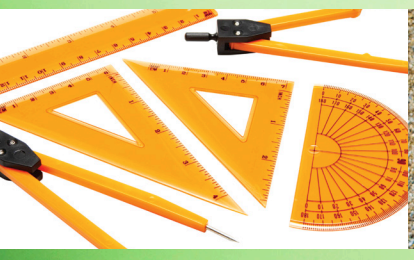
2011 NAEP-TIMSS Linking Study:

Technical Report on the Linking Methodologies and Their Evaluations




NCES 2014-461
U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

Photo Credits:

© Thomas Northcut/Jupiterimages #sb10067958gp-001; © Hans Peter Merten/Getty Images, Inc. #BB3597-004; © Chris Schmidt/Getty Images, Inc. #109725571; © Don Bayley/Getty Images, Inc. #165774250

2011 NAEP-TIMSS Linking Study:

Technical Report on the Linking Methodologies and Their Evaluations

October 2014

Gary Phillips
American Institutes for Research

Yue Jia
Xueli Xu
Educational Testing Service

Lauress L. Wise
Caroline Wiley
Tirso E. Diaz
Human Resources Research Organization

Taslima Rahman
National Center for Education Statistics

NCES 2014-461
U.S. DEPARTMENT OF EDUCATION

NAEP
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

TIMSS USA

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
Sue Betka
Acting Director

National Center for Education Statistics
Peggy G. Carr
Acting Commissioner

The National Center for Education Statistics (NCES), located within the U.S. Department of Education and the Institute of Education Sciences, is the primary federal entity for collecting and analyzing data related to education.

The National Assessment of Educational Progress (NAEP) is a congressionally authorized project sponsored by the U.S. Department of Education. The Commissioner of Education Statistics is responsible by law for carrying out the NAEP project.

The Trends in International Mathematics and Science Study (TIMSS) is an international comparative study of student achievement developed and implemented by the International Association for the Evaluation of Educational Achievement (IEA).

NCES, IES, U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

October 2014

This report was prepared with support from American Institutes for Research (AIR), CRP Incorporated, Educational Testing Service (ETS), and the Human Resources Research Organization (HumRRO) for the contract No. ED-07-CO-0107.

Full results can be found at: http://nces.ed.gov/nationsreportcard/studies/naep_timss/.

Suggested Citation
Jia, Y., Phillips, G., Wise, L.L., Rahman, T., Xu, X., Wiley, C., Diaz, T.E. (2014). *2011 NAEP-TIMSS Linking Study: Technical Report on the Linking Methodologies and Their Evaluations* (NCES 2014-461). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Content Contact
Taslima Rahman
(202) 502-7316
taslima.rahman@ed.gov

Contents

List of Tables

Table **Page**

Table **Page**

Table **Page**

Table **Page**

Table **Page**

Table **Page**

Table **Page**

List of Figures

Figure **Page**

Figure **Page**

Executive Summary

This technical report describes several methods used to establish statistical links between the 2011 National Assessment of Educational Progress (NAEP) and the 2011 Trends in International Mathematics and Science Study (TIMSS) in mathematics and science at grade 8. The goal of the 2011 NAEP-TIMSS linking study, supported by the National Center for Education Statistics (NCES), was to obtain comparable TIMSS results for U.S. states that participated in NAEP but did not participate in TIMSS.

Based on the results from the 2011 linking study, it was found that NAEP performance data can be expressed in the metric of TIMSS. By expressing both assessments in the same metric, the TIMSS mean and TIMSS benchmark percentages that each state might have obtained (had that state actually taken TIMSS) can be reported and compared to international TIMSS results.

The 2011 linking study was designed to allow NCES to perform multiple linking methods: calibration, statistical projection, and statistical moderation. Multiple contractors were involved in conducting the study. Calibration and statistical projection were performed by Educational Testing Service (ETS), while statistical moderation was performed by the American Institutes for Research (AIR). Each linking method is described in detail and descriptions of methodologies and results are presented by each respective author in this technical report. In addition, the results obtained by each method were evaluated by the Human Resources Research Organization (HumRRO), and the evaluation procedures, results, and recommendations are presented in the penultimate chapter.

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 only participated in the 2011 NAEP grade 8 mathematics and science assessments. The predicted results were validated using 2011 TIMSS results for the nine U.S. states (Alabama, California, Colorado, Connecticut, Florida, Indiana, Massachusetts, Minnesota, and North Carolina) that participated in TIMSS 2011 at the state level. The decision to use statistical moderation was based on the consideration that while all three methods of linking yielded essentially the same predicted TIMSS results, the statistical moderation technique is the simplest method among the three 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 additional samples tested with special booklets that included items from both assessments. Selecting this relatively simple and efficient methodology allows NCES to conduct additional linking studies in the future without the additional resources needed for the braided-booklet samples.

Chapter 1: Introduction

The 2011 NAEP-TIMSS linking study conducted by the National Center for Education Statistics (NCES) was designed to predict Trends in International Mathematics and Science Study (TIMSS) scores for the U.S. states that participated in 2011 National Assessment of Educational Progress (NAEP) mathematics and science assessment of eighth-grade students. 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 and 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 and 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 et al. 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.

Since the 2011 linking study required a large amount of data from both NAEP and TIMSS, a variety of samples, and multiple types of analyses were used. In addition, multiple NCES contractors were involved in the conduct of the 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. 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.

Chapter 2 of the Technical Report describes the design of the study, including information on the samples, instruments, states that participated in TIMSS at the state level, and the similarities and differences between NAEP and TIMSS. In each of the next three chapters, the linking methods and results are described in detail by each of the respective authors. Chapter 3 explains the calibration methodology approach conducted by ETS, and in Chapter 4, the statistical projection method (also conducted by ETS) is described and the findings presented. Chapter 5 discusses the statistical moderation method used by AIR and summarizes their results. Chapter 6 presents an evaluation of the results obtained by each linking method and a final set of recommendations made by HumRRO. Chapter 7, the final chapter, includes a summary, conclusions, and recommendations from the linking studies that were conducted. Tables and references relevant to each method are located at the end of each respective chapter.

References

Feuer, M.J., Holland, P.W., Green, G.F., Bertenthal, M.W., and 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.

Kolen, M.J., and 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.

Mislevy, R.J. (1992). *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects*. Princeton, NJ: Policy Information Center, Educational Testing Service.

Chapter 2: Study Design

As discussed in the Introduction, the goal of the linking study was to use grade 8 NAEP mathematics and science data to predict U.S. states' average TIMSS scores and the percentage of students reaching each of the TIMSS international benchmarks. This chapter discusses the considerations that informed the final study design in linking 2011 NAEP and 2011 TIMSS.

Various designs can be used in linking two assessments, and the design selected will affect the linking methodology used. Three designs generally used are

1. single group design,
2. random groups design, and
3. Non-Equivalent Groups Anchor Test (NEAT) design.

In single-group design, a single group of students takes both assessments. In the second method, two assessments are given to randomly equivalent groups of test takers. Equivalent groups are often formed by giving both assessments at the same time, with half of the examinees randomly selected to take one test and the remaining half taking the other (Feuer et al. 1999). In the third method, the NEAT design, assessment forms that contain common items are created for the two tests. For example, if group 1 is administered item sets A and B while group 2 is administered item sets B and C, then the items in set B are the common items or anchor items. (Refer to Kolen and Brennan 2004, for a detailed discussion of the above designs.)

The NAEP and TIMSS frameworks for mathematics and science describe the types of items that should be included in the assessments and how they should be scored.[1] There are no common items between the NAEP and TIMSS assessments because they were developed separately according to their own frameworks and test specifications. In terms of designing a study to link NAEP and TIMSS, a random groups design can be used by assigning NAEP and TIMSS assessments to two randomly equivalent groups. It is also possible to apply a single group design by assessing the same students with both the NAEP and TIMSS instruments. In addition, a NEAT design can be considered by creating new assessment forms with both NAEP and TIMSS items included in each of the forms.

There have been three previous efforts to link NAEP and TIMSS in order to predict TIMSS scores for states that participated in NAEP but not in TIMSS. All three of the studies used a random groups design, meaning that the student samples that took NAEP and TIMSS were different but were assumed to be randomly equivalent. The first study used both mathematics and science

[1] A comprehensive comparison of the NAEP and TIMSS assessment frameworks and a comparison of TIMSS assessment items against NAEP frameworks show that NAEP and TIMSS frameworks are similar but not identical in what is assessed. Results of these comparisons are available at http://nces.ed.gov/nationsreportcard/studies/naep_timss/.

results from the 1996 NAEP at grades 4 and 8 and the 1995 TIMSS at grades 4 and 8 (Johnson 1998). The second study used results from the 2000 NAEP grade 8 and the 1999 TIMSS grade 8 administrations (Johnson et al. 2003). Both studies attempted to link NAEP and TIMSS assessments—administered one year apart—to nonoverlapping samples of U.S. students. The third and most recent attempt used results from the NAEP and the TIMSS grade 8 mathematics assessments that were administered in 2007 (Phillips 2009). The two assessments were given in the same year but no individual student took both.

In addition to the random groups design, the study by Johnson et al. (2003) assessed a linking sample, where a subsample of NAEP respondents in the United States took the 1999 TIMSS instrument a few months after the 2000 NAEP administration. However, there were problems encountered in that study based on reported evidence that performance on TIMSS differed between the linking sample and the U.S. national sample that took the 1999 TIMSS assessment. The study authors further identified two contextual differences that might contribute to this discrepant performance. First, TIMSS was administered in the linking sample several months after NAEP, and the lack of consequence for NAEP may have lessened the students' motivation when taking TIMSS. The second factor was the issue of intact classrooms vs. within-school sampling. TIMSS draws one or more intact classes from the sampled schools, while the linking sample followed the NAEP approach of randomly sampling students within the sampled schools. The authors speculated that non-intact-classroom testing might also result in lower performance. Since TIMSS functioned differently in the linking sample than in the U.S. national sample, it was decided not to use the linking sample to predict the state TIMSS scores.

It is worth noting that in the three linking studies mentioned above, the relationship between scores on the NAEP and TIMSS assessments was either not available because students did not take both assessments, or could not be reliably estimated because of the potential context effect introduced by assessing the same students on NAEP and TIMSS assessments at different points in time (i.e., one administration following the other a few months later). As a result, the validity of the linkage depended upon the untestable assumption that a given level of performance on one assessment implied a certain performance on the other (Johnson et al. 2003).

The 2011 NAEP-TIMSS linking study design was intended to improve upon these previous efforts in two important ways:

1. The 2011 study was designed so that some students were part of braided-booklet samples that took items from both the 2011 NAEP and 2011 TIMSS assessments at the same time and under the same testing conditions. The design made it feasible to estimate the correlation between NAEP and TIMSS, and allowed more than one linking method to be used. The comparisons among the results from different linking methods

provided empirical evidence on the robustness of the predicted states' TIMSS scores to the linking approaches and shed light on future linking study designs.

2. Nine states participated in grade 8 TIMSS at the state level in 2011. The actual TIMSS scores from the nine states were used to validate the predicted results based on the linking study. Note that there were 3 validation states in the 1995-1996 NAEP-TIMSS grade 8 linking study, 12 validation states in the 1999-2000 grade 8 linking study, and 2 validation states in the 2007 grade 8 linking study.

The following sections in this chapter describe the design of the 2011 NAEP-TIMSS linking study and provide more details about the elements that were intended to improve upon previous NAEP-TIMSS linking efforts. Also included in the chapter is a summary of the key similarities and differences between the NAEP and TIMSS assessments.

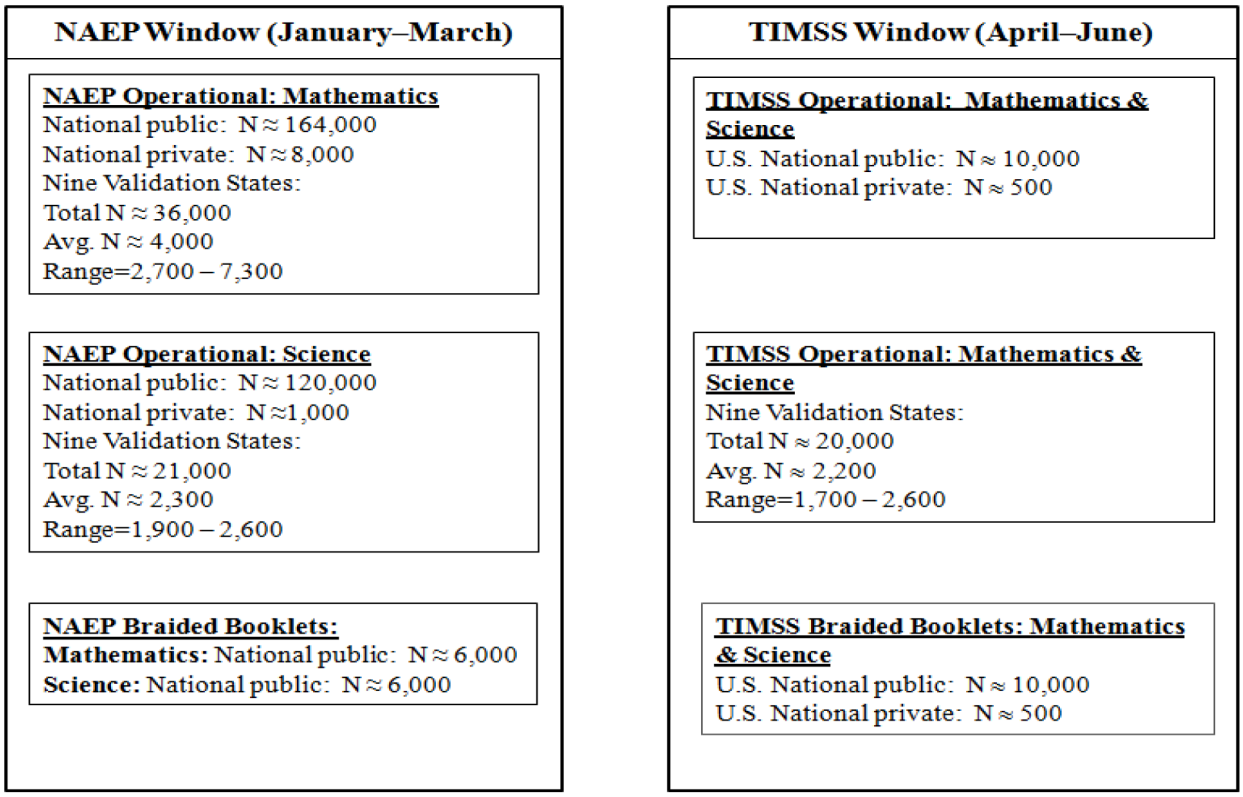
Samples and Instruments

The linking study design entailed use of data from 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 (following NAEP administration procedures) with braided booklets containing one block of NAEP items and one block of TIMSS items; and
4. Students assessed during the 2011 TIMSS testing window (following TIMSS administration procedures) with braided booklets containing one block of NAEP items and three blocks of TIMSS items.

Thus, there were samples of students that took booklets containing both NAEP and TIMSS items: two were assessed during the usual NAEP administration window (separate mathematics and science braided-booklet samples) and another during the U.S. TIMSS administration window. Figure 2.1 depicts the overall design of the study with the four instruments used in the study.

Figure 2.1. Study design and sample sizes assessed for the 2011 NAEP-TIMSS linking study



1. 2011 NAEP Operational/National Mathematics and Science Samples

NAEP mathematics and science assessments were administered at the state and national levels in winter 2011, the regular NAEP assessment window. The NAEP mathematics assessment had already been scheduled for 2011, and the National Assessment Governing Board, which sets policy for NAEP, added eighth-grade science to the assessment schedule for 2011 so that the linking study could be carried out for both mathematics and science.

Using a matrix-sampling approach and Balanced Incomplete Block (BIB) design (Allen, Donoghue, and Schoeps 2001), the 2011 NAEP mathematics assessment involved assembling a total of 155 items into a series of 10 mutually exclusive sets, or blocks, of items. The item blocks were then assembled into 50 booklets, each one including 2 blocks of items. Similarly, the NAEP 2011 science assessment involved assembling a total of 149 items into 9 blocks that were paired to form 36 booklets, each one including 2 blocks of items. The time provided for students to complete each of the mathematics or science blocks was 25 minutes. Each student was administered one subject in a single booklet.

Public school students from all 50 states and the District of Columbia participated in the state assessments. The NAEP national sample was then composed of all the state samples, a national

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

sample of private school students, as well as students from Department of Defense schools and Bureau of Indian Education schools. A total of 175,000 eighth-graders participated in the 2011 NAEP mathematics assessment, and a separate sample of approximately 122,000 eighth-graders participated in the 2011 NAEP science assessment. The number of students assessed in NAEP at grade 8 by state/jurisdiction and by subject is available at http://nationsreportcard.gov/math_2011/participation.aspx and http://nationsreportcard.gov/science_2011/participation.aspx.

2. 2011 TIMSS U.S. Operational/National Sample

The United States was one of 47 "education systems" (not counting individual participating U.S. states)[2] that participated in the 2011 TIMSS assessment at grade 8. Using a matrix-sampling approach as used for NAEP, the 2011 TIMSS assessment grouped the assessment items into blocks and then assembled the blocks into 14 booklets, with each student completing just one booklet. Each booklet in the 2011 TIMSS assessment contained four blocks—two mathematics blocks and two science blocks—but the subjects were counterbalanced across the set of booklets so that they were not always presented in the same order. Students were given 45 minutes to complete the first two blocks of items in one subject area. Then, following a short break, students were given another 45 minutes to complete two blocks of items in the other subject area. See Mullis et al. (2009) for additional details about the 2011 TIMSS assessment design.

TIMSS was administered in participating countries and subnational education systems in the northern hemisphere, including the U.S., from April through June 2011. The 2011 TIMSS U.S. national sample consisted of approximately 10,500 eighth-graders from both public and private schools and was representative at the national level and not at the state level.

3. Braided-Booklet Samples in 2011 NAEP Administration Window

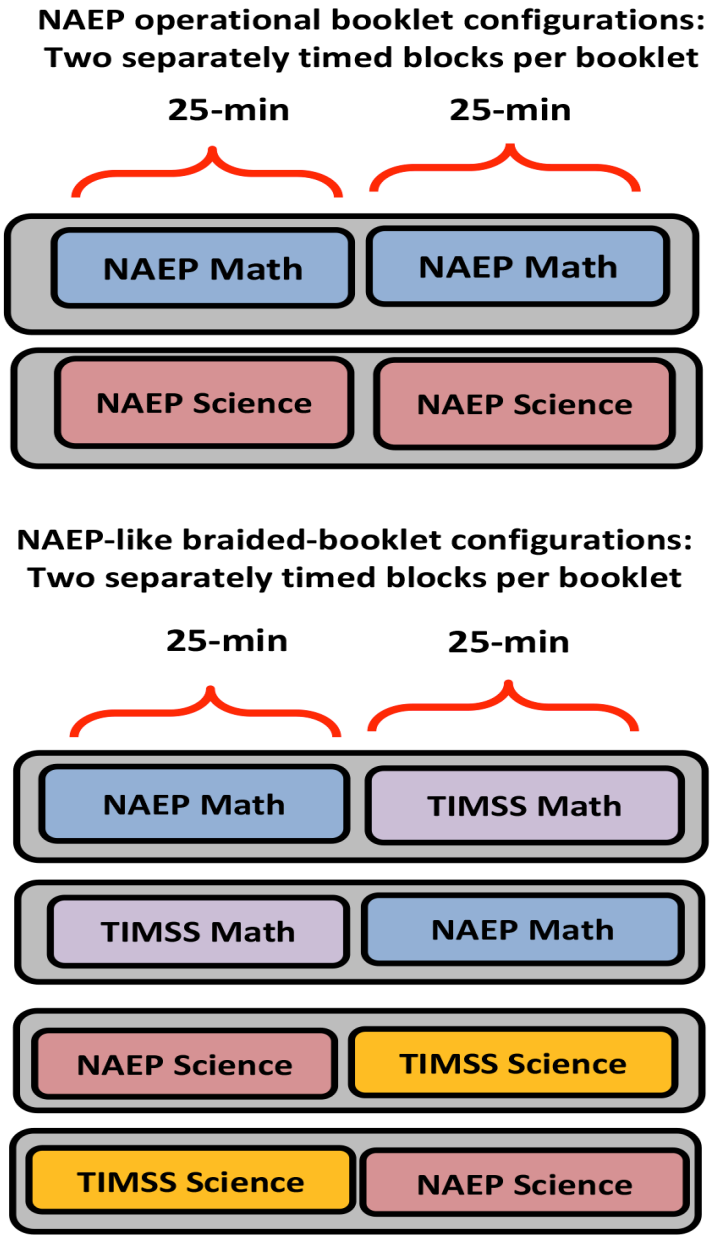
Braided mathematics booklets, a set of customized assessment booklets containing one block of NAEP mathematics items and one block of TIMSS mathematics items, were administered to a random sample of 5,700 students participating in NAEP. Similarly, braided science booklets containing one block of NAEP science items and one block of TIMSS science items were administered to a separate random sample of 6,000 students. The students in the braided-booklet samples were assessed at the same time and under the same conditions as the NAEP administration without knowing they were given an operational NAEP test booklet or a braided booklet with items from two different assessments. Only public school students were selected for the braided-booklet samples in both subjects.

All items from the 2011 NAEP mathematics and science assessments and the 2011 TIMSS assessment were administered through these braided-booklet samples. Each item block was 25 minutes long. The booklets were designed to appear as similar as possible to a regular NAEP assessment

[2] These education systems included 38 countries, such as Australia, Finland, and Japan, and nine "subnational entities," such as Alberta in Canada, Dubai in the United Arab Emirates, and England in Great Britain.

booklet, were administered under the same conditions as NAEP, and were followed by the NAEP student questionnaire. Also collected were data from the NAEP teacher and school questionnaires. Figure 2.2 shows examples of the configurations for the NAEP booklets and NAEP-like braided booklets. Note that this resulted in a slight departure from the usual operational procedures in the administration timing of TIMSS blocks; students were given 25 minutes to complete each TIMSS block in this braided-booklet sample, whereas in a regular TIMSS assessment they received 45 minutes in a timed section to complete two TIMSS blocks.

Figure 2.2. NAEP booklet and NAEP-like braided-booklet configurations



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

This approach of combining content from different assessments into booklets was similar to the braided-booklet design used by NAEP trend studies in 2009 for reading at grades 4, 8, and 12, as well as for mathematics at grade 12 (http://nces.ed.gov/nationsreportcard/reading/trend_study.asp).

4. Braided-Booklet Sample in 2011 TIMSS Administration Window

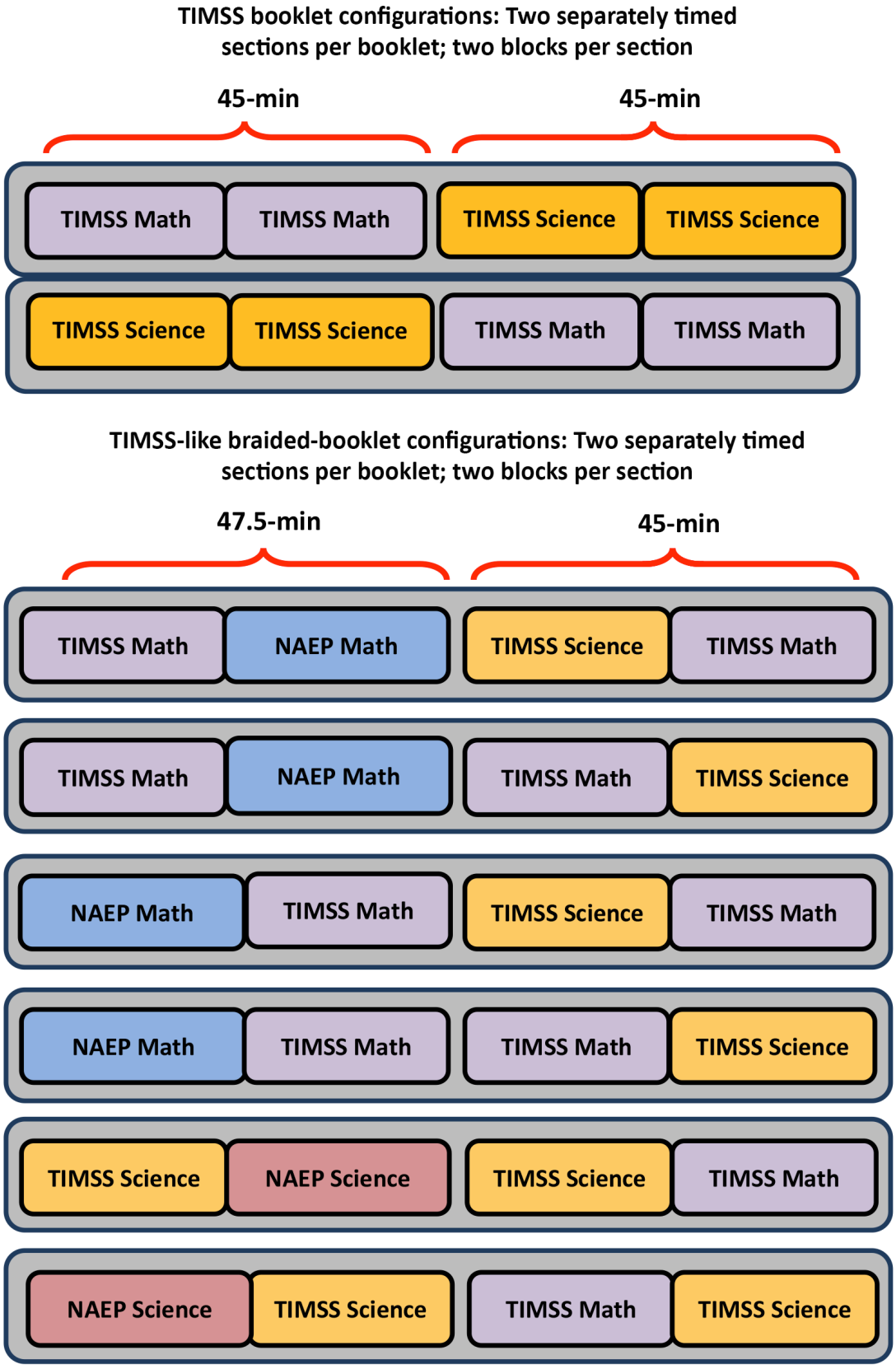
Braided assessment booklets containing a combination of TIMSS and NAEP items in mathematics and science were administered to a nationally representative sample of approximately 10,500 students. The braided-booklet sample in the 2011 TIMSS window was administered in the same schools in which TIMSS was administered, with one intact classroom randomly assigned to the U.S. TIMSS national sample and another to the braided-booklet sample. The students in the braided-booklet samples were assessed at the same time and under the same conditions as the TIMSS administration without knowing they were given an operational TIMSS test booklet or a braided booklet with items from two different assessments.

The braided booklets administered in the 2011 TIMSS administration window contained either

- one block of NAEP mathematics with two blocks of TIMSS mathematics and one block of TIMSS science, or
- one block of NAEP science with two blocks of TIMSS science and one block of TIMSS mathematics.

The booklets were designed to appear as similar as possible to a regular TIMSS assessment booklet, were administered under nearly the same conditions as TIMSS, and were followed by the TIMSS student questionnaire. Figure 2.3 depicts configurations for the TIMSS booklets and TIMSS-like braided booklets. Students were given 47.5 minutes to complete the first two blocks of items, including one TIMSS block and one NAEP block. Then, following a short break, students were given another 45 minutes to complete two blocks of TIMSS items. Note that the NAEP blocks always appeared in the first 47.5 minutes of the timed section of a braided booklet, considering that the total testing time for the NAEP assessments is 50 minutes long. Also note that the design resulted in a slight departure from standard operational TIMSS procedures for administration timing. This braided-booklet sample allowed a total of 47.5 minutes in the first timed section of the braided booklets, while the standard TIMSS practice allows 45 minutes. The 47.5 minutes timing in the first timed section was set as each NAEP block is 25-minutes long in an operational NAEP setting, while each TIMSS block is assumed to be 22.5-minutes long in a 45-minute operational TIMSS section with two TIMSS blocks. Both blocks were distinct and timed separately. This reflected the desire to allow students responding to a combined NAEP and TIMSS block section to have approximately comparable time to the operational timing in NAEP and TIMSS settings.

Figure 2.3. TIMSS booklet and TIMSS-like braided-booklet configurations



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

U.S. Validation States

In addition to the TIMSS U.S. national sample of eighth-graders, nine states participated in 2011 TIMSS directly as separate jurisdictions, with only public school students sampled from each state. These states included Alabama, California, Colorado, Connecticut, Florida, Indiana, Massachusetts, Minnesota, and North Carolina. Among those nine states, Massachusetts and Minnesota have participated in TIMSS more routinely over the years than others. The participating states did not contribute to the construction of the TIMSS scales, but did receive TIMSS results through their participation (Foy, Brossman, and Galia 2012). Thus, these states were given the opportunity to directly compare the mathematics and science achievement of their students against the TIMSS countries and subnational education systems. In the linking study, these nine states provided a "validation sample" upon which the NAEP-TIMSS link was evaluated.

The states were selected based on their state student enrollment size, their willingness to participate, their previous experience in TIMSS at the state level, and their geographic diversity. NCES also considered whether they as a group represented a substantial range of performance relative to the national average on NAEP. Public schools were selected for the TIMSS validation state samples to maximize the overlap with the TIMSS U.S. national sample and minimize the overlap with the NAEP state samples. About 1,700 to 2,600 public school students from each of the nine validation states—approximately 19,600 in total—participated in the TIMSS assessment. The actual TIMSS results from the nine states were reported along with the other international results in the TIMSS 2011 U.S. Highlights report (Provasnik et al. 2012), and the number of students assessed from the nine states is available at http://timssandpirls.bc.edu/methods/pdf/T11_Student_Sizes.pdf.

Similarities and Differences Between NAEP and TIMSS

The NAEP and TIMSS assessments in mathematics and science both measure student achievement. However, a number of key characteristics of NAEP and TIMSS have a bearing on the adequacy of any link between the two assessments. These include the following:

- NAEP and TIMSS are both designed to provide valid and reliable measurement of student group (not individual) achievement. Both assessments are administered by the National Center for Education Statistics (NCES) in the U.S.
- The frameworks that define the content measured by the NAEP and TIMSS assessments are similar but not identical (Neidorf et al. 2006; Nohara 2001; Provasnik et al. 2012).
- NAEP and TIMSS both use a matrix-sampling approach that involves assembling the entire assessment pool of items into a set of booklets, with each student completing just one booklet. This approach ensures that the range of content specified in the framework is represented by the full item pool, but that each participating student takes a portion of that overall pool to reduce student test-taking burden.

- NAEP and TIMSS both employ rigorous sampling techniques so that achievement for the overall population and for subgroups of interest may be estimated accurately by assessing a sample of students from a sample of schools.
- The state samples for all the states that participated in NAEP and the validation state samples for those that participated in TIMSS both include public school students only.
- NAEP and TIMSS both use a combination of Item Response Theory (IRT) models and population-structure latent regression models to provide estimated distributions of underlying performance for the student groups of interest.
- The two assessments share similar usage. Both assessments are used in a low-stakes fashion—that is, no serious consequences are attached to scores. NAEP can be used to compare the performance of groups of students in one state with the performance of groups of students in other states. TIMSS tells us how U.S. students as a group are doing compared with those in other countries and sub-national education systems (such as regions, districts, or provinces within a country). Neither program is designed to report how individual students are performing relative to national and international standards.

Detailed comparisons of the NAEP and TIMSS frameworks and items are available at http://nces.ed.gov/nationsreportcard/studies/naep_timss/.

While clearly similar, the NAEP and TIMSS assessments do differ in ways that will impact the link between the two; these include but are not limited to the following:

- Instrument configuration and assessment timing: In NAEP, each sampled student responded to two 25-minute blocks of items in either mathematics or science.[3] In TIMSS, the sampled students were given 45 minutes to complete two blocks of items in one subject area and another 45 minutes to complete two blocks of items in another subject area.
- Content area specifications: The NAEP and TIMSS frameworks for mathematics and science describe the content areas to be assessed. Table 2.1 compares the content area specifications between the two assessments.
- Item type specifications: Both assessments include multiple-choice and constructed-response items. There are more constructed-response items in 2011 TIMSS than in 2011 NAEP (45% vs. 26% respectively, in mathematics, and 49% vs. 34% in science). However, most of the constructed-response items in TIMSS are dichotomously-scored with just two score categories, while NAEP contains more polytomously-scored constructed-response items with three or more score levels. The emphasis and distribution of items across content areas are also different. Table 2.2 lists the distribution of items in the 2011 NAEP grade 8 assessments and the 2011 TIMSS grade 8 assessments by subject and item type.

[3] Because there were not enough students in the District of Columbia to constitute three mutually exclusive samples for NAEP assessments in reading, mathematics, and science in 2011, students took either the reading or mathematics assessment on one day, and then, on the second day, depending on which assessment they had already taken, they took either science or reading or mathematics.

- Testing aids allowed: NAEP allows regular or scientific calculators for some mathematics items. For example, ruler/protractors, geometric shapes, and other manipulatives are provided for certain mathematics items. TIMSS permits the use of regular calculators throughout the mathematics assessment. No testing aids are allowed in either NAEP or TIMSS for science items.
- Testing window: NAEP was conducted January through March 2011. TIMSS was conducted in the United States (and in most Northern Hemisphere countries) April through June 2011.
- Participation: The 2011 grade 8 NAEP assessments tested students from all 50 states, the District of Columbia, and Department of Defense schools. The 2011 grade 8 TIMSS was administered to 47 countries and subnational education systems.
- Testing population: Results reported for the states and jurisdictions by NAEP are based on students in public schools only, whereas most countries and subnational education systems in TIMSS assess students in public and private schools. In addition, NAEP and TIMSS follow different accommodation and exclusion policies. NAEP allows accommodations (e.g., extra testing time or individual rather than group administration) so that more Students with Disabilities (SD) and English Language Learners (ELL) can participate in the assessment. Unlike NAEP, TIMSS does not provide testing accommodations for SD and ELL students. As a result, the exclusion rates in TIMSS are generally higher than in NAEP. The exclusion rates and accommodation rates for the states that participated in the 2011 NAEP are listed in table 2.3. NAEP exclusion rates represent the percentage of SD and/or ELL students excluded, as a percentage of all students. The exclusion rates for the countries and subnational education system/validation states that participated in 2011 TIMSS are provided in table 2.4.
- Sample size: The NAEP national sample contains 175,000 eighth-graders for the mathematics assessment and 122,000 eighth-graders for the science assessment. In comparison, approximately 10,500 eighth-graders participated in TIMSS as the U.S. national sample.
- Within-school sampling vs. intact classrooms: NAEP assesses random samples of students within the sampled schools, while TIMSS draws one or more intact classes from the sampled schools.

Chapter 6 investigates the impact of some of the above listed differences between NAEP and TIMSS assessments on the states' predicted TIMSS results.

Tables

Table 2.1. Eighth-grade content areas specified in NAEP and TIMSS frameworks, by subject: 2011

Subject	NAEP	TIMSS
Mathematics (with framework target percentages in parentheses)	Number properties and operations (20%)	Number (30%)
	Geometry (20%)	Geometry (20%)
	Algebra (30%)	Algebra (30%)
	Data analysis, statistics, and probability (15%)	Data and chance (20%)
	Measurement (15%)	
Science (with framework target percentages in parentheses)	Physical science (30%)	Physics (25%)
	Earth and space sciences (40%)	Earth science (20%)
	Life science (30%)	Biology (35%)
		Chemistry (20%)

Table 2.2. Distribution of items in NAEP eighth-grade assessments and TIMSS eighth-grade assessment, by subject and item type: 2011

	Mathematics		Science	
Item type	2011 NAEP	2011 TIMSS	2011 NAEP	2011 TIMSS
Number of content areas	5	4	3	4
Total items	155	215	149	219
Multiple choice	115	118	98	112
Dichotomously-scored constructed response	8	82	1	90
Polytomously-scored constructed response	32	15	50	17

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 2.3. Percentage of eighth-grade public school students identified as students with disabilities and/or English language learners excluded and assessed in NAEP mathematics and science, as a percentage of all students, by state: 2011

State	Mathematics					Science				
			Assessed					Assessed		
	Identified	Excluded	Total	Without accommodations	With accommodations	Identified	Excluded	Total	Without accommodations	With accommodations
United States (public)	**18**	**3**	**15**	**5**	**10**	**18**	**2**	**16**	**5**	**11**
Alabama	12	1	11	7	4	12	1	11	7	4
Alaska	21	3	18	4	14	21	1	20	4	16
Arizona	12	1	11	2	9	12	1	11	2	9
Arkansas	16	1	14	3	12	16	1	15	3	12
California	23	1	22	15	7	23	2	22	14	8
Colorado	16	1	15	5	10	16	1	15	5	10
Connecticut	16	1	15	2	12	16	1	15	2	13
Delaware	16	3	13	2	11	16	2	14	2	12
District of Columbia	21	4	17	2	15	21	1	20	2	18
DoDEA[1]	14	3	11	3	8	14	1	13	3	10
Florida	19	2	17	1	16	19	1	17	1	16
Georgia	12	3	9	2	7	12	2	10	2	8
Hawaii	20	2	18	7	11	20	2	18	7	11
Idaho	12	1	10	3	7	12	1	10	4	7
Illinois	17	2	15	3	12	17	1	16	3	12
Indiana	17	3	14	2	12	17	1	16	3	13
Iowa	17	1	16	2	14	17	1	16	2	14
Kansas	18	1	16	7	9	18	1	16	7	9
Kentucky	13	3	10	2	8	13	3	10	2	8
Louisiana	15	1	14	1	13	15	1	14	1	13
Maine	20	2	18	4	14	20	2	18	4	14
Maryland	14	6	8	1	7	14	2	12	1	11
Massachusetts	22	4	18	3	15	22	3	19	3	16
Michigan	14	4	11	3	8	14	3	12	3	8
Minnesota	17	2	15	6	9	17	2	15	7	8
Mississippi	8	1	7	1	6	8	1	7	1	6
Missouri	14	1	12	2	10	14	1	13	3	10
Montana	13	2	12	2	9	13	2	12	3	9
Nebraska	16	4	13	4	9	16	1	15	3	12
Nevada	18	3	15	6	9	18	1	17	6	11
New Hampshire	20	2	18	4	14	20	2	18	5	13

See notes at end of table.

Table 2.3. Percentage of eighth-grade public school students identified as students with disabilities and/or English language learners excluded and assessed in NAEP mathematics and science, as a percentage of all students, by state: 2011—Continued

State	Mathematics					Science				
			Assessed					Assessed		
	Identified	Excluded	Total	Without accommodations	With accommodations	Identified	Excluded	Total	Without accommodations	With accommodations
New Jersey	19	4	15	1	14	19	1	18	1	17
New Mexico	22	2	20	10	10	22	2	20	10	10
New York	20	1	19	#	18	20	1	19	#	18
North Carolina	18	2	16	3	12	18	2	16	4	12
North Dakota	16	4	11	3	9	16	3	13	2	10
Ohio	16	5	11	1	10	16	2	14	2	12
Oklahoma	18	10	8	4	4	18	3	15	5	10
Oregon	18	1	16	6	11	18	2	16	6	10
Pennsylvania	17	2	15	2	13	17	1	16	2	15
Rhode Island	19	1	18	4	13	19	1	19	4	14
South Carolina	15	4	11	4	8	15	1	14	5	9
South Dakota	13	2	11	4	7	13	1	11	3	8
Tennessee	13	4	9	1	8	13	1	12	1	10
Texas	18	5	13	8	5	18	2	16	8	8
Utah	14	3	11	3	8	14	2	12	3	9
Vermont	20	1	18	4	15	20	1	18	4	14
Virginia	18	3	15	6	9	18	3	15	5	10
Washington	16	2	14	4	10	16	2	14	5	10
West Virginia	14	2	12	3	9	14	2	12	3	9
Wisconsin	18	2	16	2	14	18	2	16	3	14
Wyoming	14	1	13	2	11	14	1	13	2	11

Rounds to zero.

[1] Department of Defense Education Activity (overseas and domestic schools).

NOTE: Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 2.4. Exclusion rates in TIMSS assessments at grade 8, by education system/validation states: 2011

Education system	Exclusion rate
Abu Dhabi-UAE	2
Alabama-USA[1]	5
Alberta-CAN	7
Armenia	2
Australia	3
Bahrain	2
California-USA[1]	6
Chile	3
Chinese Taipei-CHN	1
Colorado-USA[1]	4
Connecticut-USA[1]	9
Dubai-UAE	4
England-GBR	2
Finland	3
Florida-USA[1]	7
Georgia	5
Ghana	1
Hong Kong SAR[2]	5
Hungary	4
Indiana-USA[1]	6
Indonesia	3
Iran, Islamic Republic of	2
Israel	23
Italy	5
Japan	3
Jordan	#
Kazakhstan	5
Korea, Republic of	2
Lebanon	1
Lithuania	5
Macedonia, Republic of	3
Malaysia	#
Massachusetts-USA[1]	8
Minnesota-USA[1]	4
Morocco	#
New Zealand	3
North Carolina-USA[1]	11
Norway	2
Oman	1
Ontario-CAN	6
Palestinian National Authority	2
Qatar	5
Quebec-CAN	5
Romania	1
Russian Federation	6
Saudi Arabia	1
Singapore	6
Slovenia	2
Sweden	5
Syrian Arab Republic	2
Thailand	2
Tunisia	#
Turkey	2
Ukraine	3
United Arab Emirates	3
United States	7

Rounds to zero.
[1] Validation state.
[2] Hong Kong SAR is a Special Administrative Region (SAR) of the People's Republic of China.
NOTE: Validation states are those U.S. states that participated in the 2011 TIMSS assessment at the state level.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

References

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

Feuer, M.J., Holland, P.W., Green, G.F., Bertenthal, M.W., and 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., and Galia, J. (2012). Scaling the TIMSS and PIRLS 2011 Achievement Data. In M.O. Martin, and 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.

Johnson, E.G. (1998). *Linking the National Assessment of Educational Progress and the Third International Mathematics and Science Study for Eighth Grade: A Research Report*. (NCES 98-499). Washington, DC: GPO.

Johnson, E.G., Cohen, J., Chen, W.-H., Jiang, T., and Zhang, Y. (2003). *2000 NAEP–1999 TIMSS Linking Report*. (NCES 2005-01). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Kolen, M. J., and Brennan, R. L. (2004). *Test Equating, Scaling, and Linking*. New York, NY: Springer.

Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y., and Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Boston College.

Neidorf, T.S., Binkley, M., Gattis, K., and 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). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

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). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Phillips, G.W. (2009). *The Second Derivative: International Benchmarks in Mathematics for American States and School Districts*. Washington, DC: American Institutes for Research.

Provasnik, S., Kastberg, D., Ferraro, D., Lemanski, N., Roey, S., and 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). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Chapter 3: Linking NAEP and TIMSS Through Calibration

Chapter 3 describes the use of calibration as a method to link the 2011 NAEP and 2011 TIMSS assessments. Chapters 4 and 5 will respectively discuss the use of statistical projection and statistical moderation as linking methods.

Calibration Linking: Use and Methods

In the literature, the term *calibration* has several different meanings and connotations. It is used here to refer to a procedure for 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, p. 430), calibration linking is a type of linking used when the two assessments are based on

- the same framework but different test specifications and different statistical characteristics, or
- 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 items 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 et al. 2006; Nohara 2001; Provasnik et al. 2012). Therefore, calibration linking is used based on the second type of linking condition listed above. Like moderation and projection linking, calibration linking is directional. That is, calibrating NAEP items onto the TIMSS scale is different from calibrating TIMSS items onto the NAEP scale.

A variety of methods exist that can be used for calibration linking. The three most commonly used are

1. concurrent calibration;
2. separate calibration with transformation; and
3. fixed parameter calibration.

In concurrent calibration, the IRT item parameters are estimated simultaneously using items and student responses from both assessments to obtain a common IRT scale. In separate calibration, item parameters are estimated separately for each assessment and then the item parameters that

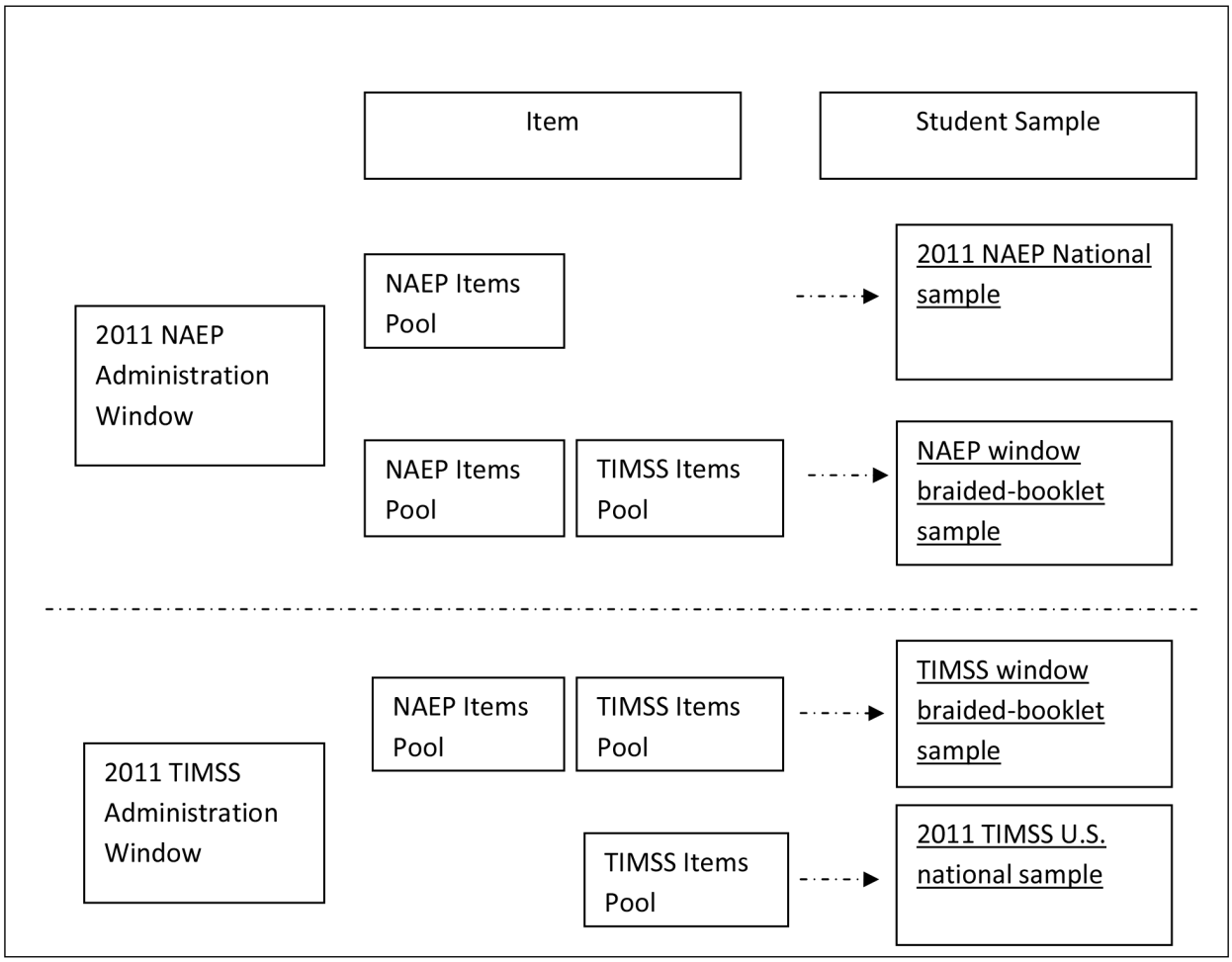
are estimated for one assessment are transformed to the scale of the other assessment using a transformation procedure such as the Stocking-Lord (Stocking and Lord 1983) or Haebara (1980) methods. To use fixed parameter calibration (Hanson and Béguin 2002; Kang and Petersen 2009; Kim 2006) to link assessment A to assessment B, the first step is to establish the scale for assessment B, just as would be done under the separate calibration method. Next, the items from assessment A are projected onto the established scale for assessment B by calibrating the items from assessments A and B together, but keeping assessment B's item parameters fixed. Compared to separate calibration, fixed parameter calibration does not require an item transformation method to place items from one assessment onto the scale of the other.

Using Calibration to Link NAEP and TIMSS

There have been no previous attempts to link NAEP and TIMSS through calibration, mainly because of the limitations of prior linking study designs, in which there were no common items shared between NAEP and TIMSS and no appropriate linking sample where students took items from both assessments. As discussed in Chapter 2, the 2011 NAEP-TIMSS linking study design was intended to improve upon previous efforts by including braided-booklet samples that contained items from both NAEP and TIMSS at the same time and under the same testing conditions. Figure 3.1 illustrates how the study design provided common items in linking NAEP and TIMSS. Both NAEP items and TIMSS items were included in the braided-booklet samples during the NAEP and the TIMSS administration windows. NAEP items were common among the 2011 NAEP sample and the two braided-booklet samples, and TIMSS items were common among the 2011 TIMSS U.S. sample and the two braided-booklet samples. The study thus supported the use of calibration linking.

Recall that the objective of the study was to use states' 2011 NAEP scores to predict their average TIMSS scores and percentages of students reaching each of the TIMSS international benchmark levels. Therefore, it was necessary for the predicted TIMSS scores to be placed on the existing TIMSS scale, which was established based on countries that participated in TIMSS. For the calibration linking analysis, the IRT item parameters for the TIMSS items were fixed at their values from the TIMSS 2011 operational analysis and then the NAEP IRT item parameters were projected onto the TIMSS scale.

Figure 3.1. Study design of the 2011 NAEP-TIMSS linking study



Three major steps were involved in calibration linking:

1. Calibrating the NAEP items onto the TIMSS IRT scale;
2. Estimating population proficiencies in TIMSS for the 2011 NAEP national sample; and
3. Transforming the proficiency distribution for the 2011 NAEP national sample to the TIMSS reporting metric.

In the following sections, each step of the calibration linking analysis is described in more detail.

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

For this first step, it was necessary to use the item parameters for the TIMSS mathematics and science items at grade 8 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. The IRT calibration procedure used in TIMSS for linking the assessments between administrations is an example of linking assessments based on the same framework. TIMSS uses

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

concurrent calibration which calibrates data from two administrations concurrently in a single IRT estimation run[4] (Foy, Brossman, and Galia 2012).

For the calibration linking done in this study, two separate fixed parameter calibrations were conducted: one for mathematics and the other for science. The item parameters for the TIMSS items were fixed at the values obtained from the TIMSS 2011 operational analysis, and the NAEP item parameters were calibrated (i.e., projected) onto the TIMSS IRT scale. The item responses from three groups of students—the 2011 NAEP national sample, 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 did not have to be included, because the NAEP item parameters needed to be estimated, and no NAEP items were administered to the 2011 TIMSS sample.

For dichotomously-scored items, two- and three-parameter logistic models (Lord and Novick 1968) were used, while for polytomously-scored items the generalized partial-credit model (Muraki 1992) was used. Item parameter estimates were obtained using the ETS proprietary version of BILOG/PARSCALE software (Muraki and Bock 1991). The student sampling weights for each of the samples were adjusted to ensure that the three samples contributed equally to the estimation of parameters for any NAEP item. The reason for balancing the sample sizes was to prevent the 2011 NAEP national sample from dominating the item parameter estimation, as the 2011 NAEP national sample was substantially larger than the two braided-booklet samples. One set of item parameters was estimated for the NAEP items.

In calibration, it was assumed that the NAEP items were functioning identically across the three groups, meaning that a single response function described the response behavior of students in all three groups who were assessed with that particular item. The fit of the IRT models was carefully checked by multiple procedures used in the operational NAEP analysis, including graphical comparisons of the empirical item response functions to the model-based (theoretical) curves, and comparisons of observed and model-predicted proportions of students obtaining a particular score on each item (Rogers et al. 2006a). (Interested readers should see Allen, Donoghue, and Schoeps (2001) for more details on the evaluation of IRT model fit in NAEP.)

Good IRT model fit is observed when the empirical results fall near the fitted curves for any given item. Figure 3.2 shows a plot of the empirical and fitted item response functions for a dichotomously-scored NAEP item from fixed parameter calibration. In the plot, the horizontal axis represents the IRT proficiency scale and the vertical axis represents the probability of having a response in a given response category. The fitted curve based on the estimated item parameters

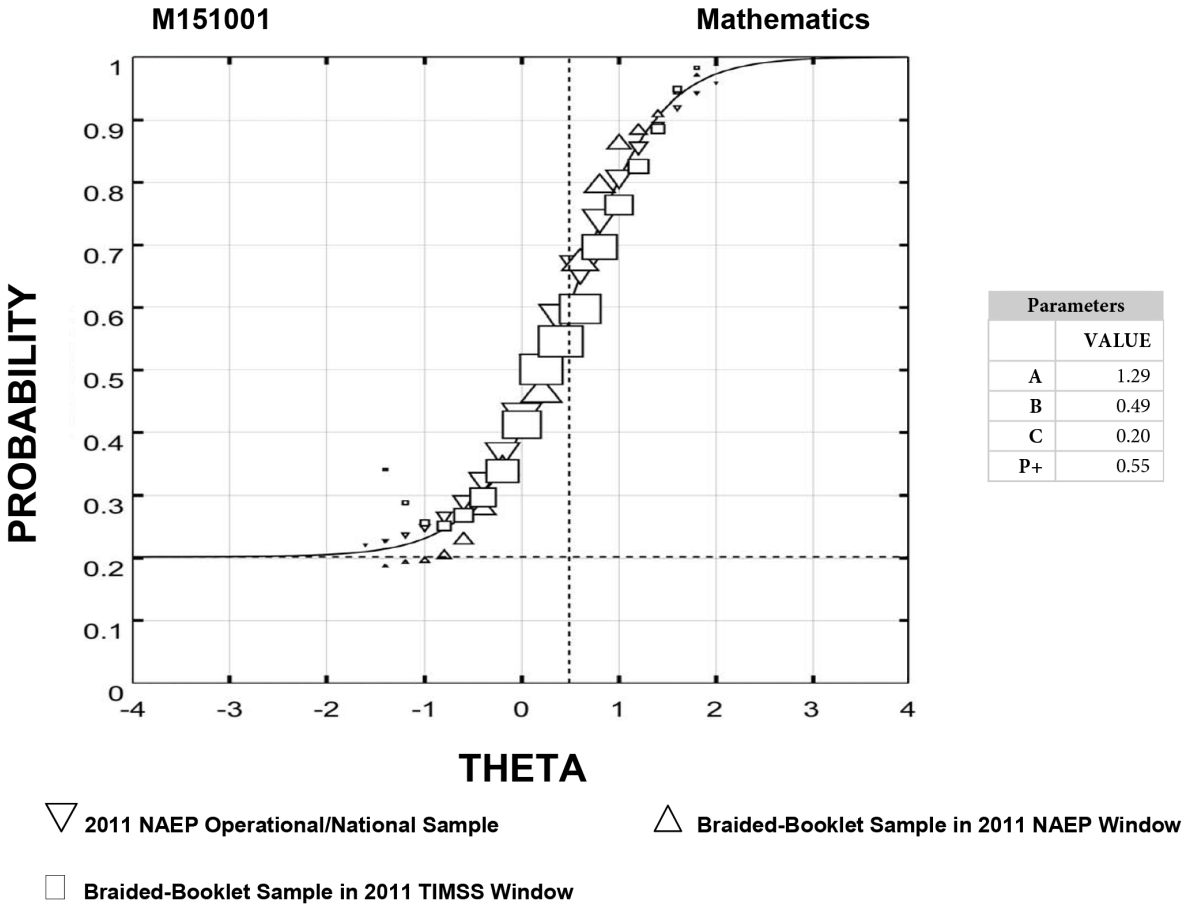
[4] In TIMSS operational analysis, common-item response functions are initially assumed for the common items, and the assumption is evaluated and modified where appropriate. In addition, separate proficiency distributions for each cohort are estimated with the item parameters. The same concurrent calibration procedure is used in linking NAEP assessments between years as well.

is shown as a solid line. Empirical results for each of the three samples (the 2011 NAEP national sample, the NAEP window braided-booklet sample, and the TIMSS window braided-booklet sample) are represented by upward triangles, downward triangles and rectangles. The center of each triangle or rectangle represents this empirical proportion of correct responses. The size of each triangle or rectangle is proportional to the number of students contributing to the estimation of its empirical proportion of correct.

Figure 3.3 contains a plot of the empirical and fitted item response functions for a polytomously-scored NAEP item. As for the dichotomous-scored item plot, the horizontal axis represents the IRT proficiency scale. But the vertical axis represents the probability of having a response in a given response category. The fitted curves that are based on the estimated item parameters are shown as solid lines. Empirical results for the three samples are represented by upward triangles, downward triangles and rectangles. The interpretation of the triangles and rectangles is the same as in figure 3.2.

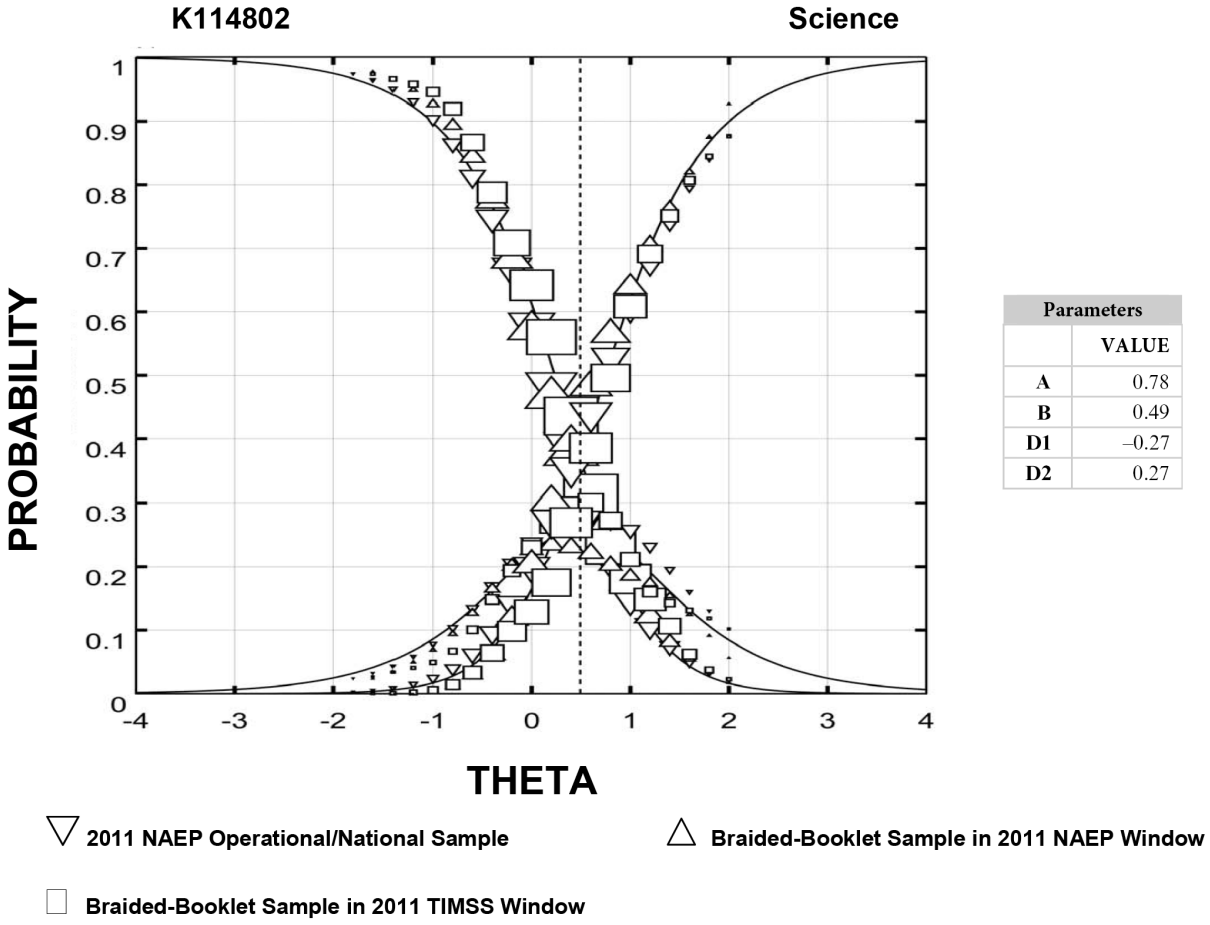
Overall, the IRT model fit to the common item response curve was acceptable for all of the NAEP items in both mathematics and science. The estimated item parameters for all 2011 NAEP mathematics and science items from fixed parameter calibration are listed in tables 3.8 and 3.9.

Figure 3.2. Example item response function for a dichotomously-scored NAEP item from fixed parameter calibration



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Figure 3.3. Example item response function for a polytomously-scored NAEP item from fixed parameter calibration



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

In the second step, the IRT item parameters for the NAEP items estimated in the first step were employed in a procedure called "conditioning" to estimate mathematics and science proficiency distributions for the 2011 NAEP national sample. The item parameters estimated in step 1 served the purpose of setting the TIMSS IRT scales on which the proficiencies were estimated.

A latent regression model was used in the conditioning analysis, given students' responses to the subset of NAEP items they received, as well as other relevant and available background information.[5] The set of background variables included in the latent regression model was identical to the set of variables used in the 2011 NAEP operational analysis. Unidimensional latent regression models were used for NAEP mathematics and NAEP science, separately. The analysis was conducted using the DGROUP set of programs (Thomas 1994; Rogers et al. 2006b).

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

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

As part of step 2, plausible values—random draws from the predictive scale score distributions for each respondent on the TIMSS IRT scale (Mislevy 1991; von Davier, Gonzalez, and 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. Twenty plausible values per respondent in the 2011 NAEP national sample were drawn.

Step 3: Transforming the proficiency distributions for the 2011 NAEP national sample to the TIMSS reporting metric

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 transformation procedure was done to transform 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 were 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 distributions. The transformation equation was as follows:

$$PV_{\text{Target}} = \hat{A} \times PV_{\text{Calibrated}} + \hat{B} \tag{3.1}$$

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 $\hat{B}$

  $\hat{A} = SD_{\text{Target}} / SD_{\text{Calibrated}}$

  $\hat{B} = M_{\text{Target}} - \hat{A} \times M_{\text{Calibrated}}$
- 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 transformation parameter estimates are listed in table 3.1.

Standard Error Estimation

Using the calibration linking procedures described above, TIMSS state-level results for all 52 states/jurisdictions that participated in the NAEP mathematics and science assessments were predicted. The error variance associated with the predicted TIMSS results from calibration linking can be expressed as

$$Var = Var_{sampling} + Var_{measurement} + Var_{transformation} \quad (3.2)$$

The sampling error accounted for the uncertainty in estimating population statistics from a sample of the population. The second variance component, measurement error, was computed from the variance between predicted TIMSS plausible values, which accounted for the uncertainty in proficiency estimation (Johnson and Rust 1992). The third variance component was associated with the transformation procedure described in step 3 of the calibration linking procedure. This source of variance was referred to as transformation error, which accounted for the uncertainty associated with estimating the transformation function parameters. A jackknife procedure was employed to estimate both sampling and transformation errors. See http://nces.ed.gov/nationsreportcard/tdw/weighting/2000_2001/2000main_varestimate_sampvar_jack.aspx for more information.

Results of Calibration Linking

This section presents the predicted state TIMSS results from calibration linking. The actual TIMSS results of the nine states that participated in the operational 2011 TIMSS assessment were used to validate the predicted results based on calibration linking. The *prediction residual error* for a state was defined as the difference between the predicted state TIMSS result ($\hat{t}_i$) and the actual TIMSS result (t_i), then the Predicted Residual Sum of Squares, or PRESS, across the nine validation states was calculated as:

$$PRESS = \sum_{i=1}^{9} \left(\hat{t}_i - t_i \right)^2 \quad (3.3)$$

And the Mean Squared Error (MSE) was used as a summary measure of the prediction results:

$$MSE = \frac{PRESS}{9} \quad (3.4)$$

Table 3.2 shows the actual TIMSS average scores for the nine validation states, and their rankings in TIMSS 2011 grade 8 mathematics. It also provides the MSE, state rankings based on their predicted state TIMSS average scores from calibration linking, *prediction residual errors*, and the rankings of the nine states based on their 2011 actual NAEP mathematics average scores. The predicted state TIMSS science results for the nine validation states are listed in table 3.3.

As shown in tables 3.2 and 3.3, the predicted TIMSS state average scores are statistically significantly different from the actual TIMSS state average scores for 4 states in mathematics, and for 3 states in science. However, as is discussed in Chapter 6, the discrepancies between predicted and actual state TIMSS results are of similar magnitude across calibration, moderation, and projection linking.

The state-level predicted TIMSS average scores from calibration linking and the predicted percentages of students reaching each of the TIMSS international benchmarks for the nine validation states, are listed in tables 3.10 and 3.11.

Further Investigation of Selection Bias and Predicted TIMSS Score Adjustments

Given the sizeable discrepancies observed between the predicted and actual state results for some of the validation states, 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 3.4, TIMSS exclusion rates are, in general, higher than in NAEP at the national level and for individual validation states. This is largely because accommodations are offered in NAEP but not in TIMSS. Such difference in the selection of assessment samples is referred to as *sample selection bias*.

An ad hoc adjustment was considered to assess and quantify the impact of selection bias due to differences in exclusion rates and accommodation policies. The state exclusion rates in NAEP were adjusted to be the same as in TIMSS. Note that the state exclusion rate for TIMSS for the validation states were known and, therefore, the following analyses were based on that subset only. With no information on which and how student groups were excluded in TIMSS but included in NAEP, this procedure presumed that those students who would most likely be excluded from TIMSS were the lowest performing accommodated (i.e., Students with Disabilities [SD] and/or English Language Learners [ELL]) students in NAEP. From each validation state sample, the exact number of accommodated students was 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.

Figure 3.4. Exclusion rate in NAEP and TIMSS assessments at grade 8, by validation state: 2011

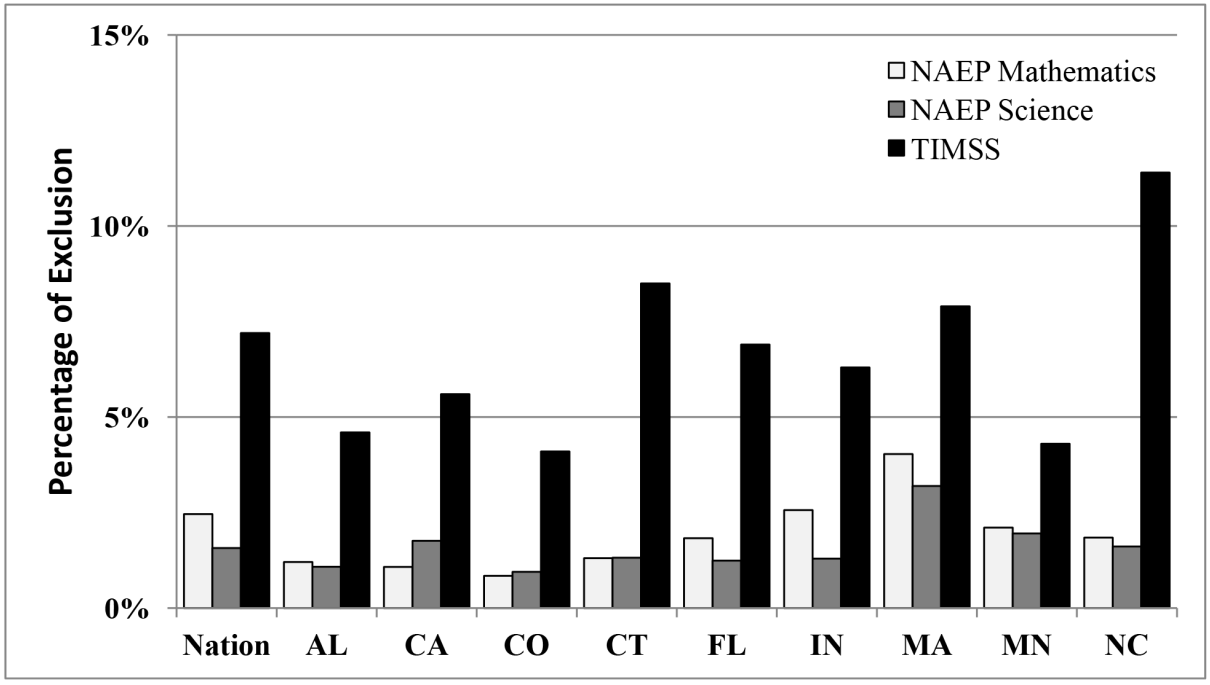


Table 3.4 provides the actual TIMSS average scores and rankings of the nine validation states in mathematics. The table also provides rankings of the validation states and the *prediction residual errors* based on

a. predicted TIMSS from calibration linking (i.e., baseline), and
b. predicted TIMSS from 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 2011 NAEP average mathematics scores are listed as well.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

MSE_prediction is defined as

$$MSE\_prediction = MSE - \frac{\sum_{i=1}^{9} Var(\hat{t}_i)}{9} - \frac{\sum_{i=1}^{9} Var(t_i)}{9} = \frac{\sum_{i=1}^{9} (\hat{t}_i - t_i)^2}{9} - \frac{\sum_{i=1}^{9} Var(\hat{t}_i)}{9} - \frac{\sum_{i=1}^{9} Var(t_i)}{9} \quad (3.5)$$

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 and MSE is defined in equation 3.4.

The PRESS, MSE and *MSE_prediction* computed from equations 3.3, 3.4 and 3.5 are presented in table 3.6. The adjustment yields smaller *prediction residual errors* for most of the validation states, and commensurate reduced PRESS and MSE values when compared to the predicted state average scores from calibration linking in the top row. Science results show similar patterns and are presented in tables 3.5 and 3.7.

Tables

Table 3.1. Coefficients of linear transformations of the univariate scale from the calibrating scale units to the units of the TIMSS reporting scale at grade 8, national assessment, by subject: 2011

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

Table 3.2. Actual TIMSS average scores, prediction residual errors, rankings, and mean squared error in eighth-grade mathematics, by validation state, calibration linking: 2011

	Actual TIMSS mathematics results			Predicted from calibration linking				
Validation state	Rank	Average score	Standard error	Rank	Average score	Standard error	Residual error	Rank in 2011 NAEP mathematics
Massachusetts	1	561	5.3	1	**540**	**3.3**	**-20**	1
Minnesota	2	545	4.6	2	**533**	**3.3**	**-12**	2
North Carolina	3	537	6.8	5	**515**	**3.5**	**-22**	5
Indiana	4	522	5.1	6	513	3.4	-9	6
Colorado	5	518	4.9	3	526	3.5	8	3
Connecticut	6	518	4.8	4	516	3.6	-1	4
Florida	7	513	6.4	7	**496**	**3.2**	**-17**	7
California	8	493	4.9	8	486	3.5	-7	8
Alabama	9	466	5.9	9	478	4.0	12	9
						Mean squared error	183	

NOTE: Residual = Predicted state TIMSS average score minus actual state TIMSS average score. Bold font indicates predicted average scores are statistically significantly different from the actual average scores. Two-tailed *t*-test, with alpha = .05, no adjustment for multiple comparisons. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 3.3. Actual TIMSS average scores, prediction residual errors, rankings, and mean squared error in eighth-grade science, by validation state, calibration linking: 2011

Validation state	Actual TIMSS science results			Predicted from calibration linking			Residual error	Rank in 2011 NAEP science
	Rank	Average score	Standard error	Rank	Average score	Standard error		
Massachusetts	1	567	5.1	1	**547**	**3.3**	**-19**	1
Minnesota	2	553	4.6	2	546	3.3	-7	2
Colorado	3	542	4.4	3	546	3.9	4	3
Indiana	4	533	4.8	5	527	3.1	-6	5
Connecticut	5	532	4.6	4	532	3.5	0	4
North Carolina	6	532	6.3	7	**515**	**3.4**	**-17**	7
Florida	7	530	7.3	6	517	3.5	-13	6
California	8	499	4.6	8	498	3.7	0	8
Alabama	9	485	6.2	9	**497**	**3.9**	**11**	9
						Mean squared error	117	

NOTE: Residual = Predicted state TIMSS average score minus actual state TIMSS average score. Bold font indicates predicted average scores are statistically significantly different from the actual average scores. Two-tailed *t*-test, with alpha = .05, no adjustment for multiple comparisons. Detail may not sum to totals because of rounding.

Table 3.4. Actual TIMSS average scores, prediction residual errors, rankings, and mean squared error in eighth-grade mathematics, by validation state, calibration linking and calibration linking with exclusion rate matching: 2011

Validation state	Actual TIMSS mathematics results			Predicted TIMSS mathematics results from calibration linking					Rank in 2011 NAEP mathematics
				(a) Predicted TIMSS mathematics			(b) Predicted w/ exclusion rate matching		
	Rank	Average score	Standard error	Rank	Residual	Standard error	Rank	Residual	
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	9	7	9
			Mean squared error		183			156	

NOTE: Residual = Predicted state TIMSS average score minus actual state TIMSS average score. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 3.5. Actual TIMSS average scores, prediction residual errors, rankings, and mean squared error in eighth-grade science, by validation state, calibration linking and calibration linking with exclusion rate matching: 2011

Validation state	Actual TIMSS science results			Predicted TIMSS science results from calibration linking					Rank in 2011 NAEP science
				(a) Predicted TIMSS science			(b) Predicted w/ exclusion rate matching		
	Rank	Average score	Standard error	Rank	Residual	Standard error	Rank	Residual	
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
Mean squared error					117			100	

NOTE: Residual = Predicted state TIMSS average score minus actual state TIMSS average score. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 3.6. PRESS and MSE values for the predicted average scores of the nine validation states in eighth-grade mathematics, calibration linking: 2011

Linking approach	PRESS	MSE	MSE prediction
(a) Calibration linking	1644	183	140
(b) Calibration linking with exclusion rate matching	1403	156	114

Table 3.7. PRESS and MSE values for the predicted average scores of the nine validation states in eighth-grade science, calibration linking: 2011

Linking approach	PRESS	MSE	MSE prediction
(a) Calibration linking	1054	117	75
(b) Calibration linking with exclusion rate matching	897	100	58

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 3.8. IRT item parameter estimates for fixed parameter calibration, NAEP grade 8 mathematics items: 2011

Block	Item	NAEP ID	a_j	b_j	c_j	d_{j1}	d_{j2}	d_{j3}	d_{j4}
MC	1	M149801	0.74	-0.96	0.18	†	†	†	†
MC	2	M149901	0.82	0.35	0.00	0.53	-0.53	†	†
MC	3	M150001	0.91	0.49	0.33	†	†	†	†
MC	4	M150101	0.69	-0.59	0.18	†	†	†	†
MC	5	M150201	1.53	0.83	0.18	†	†	†	†
MC	6	M150301	1.83	0.49	0.11	†	†	†	†
MC	7	M150401	1.05	-0.02	0.18	†	†	†	†
MC	8	M150501	1.41	0.64	0.22	†	†	†	†
MC	9	M150601	1.82	1.43	0.11	†	†	†	†
MC	10	M150701	1.35	0.41	0.15	†	†	†	†
MC	11	M150801	0.72	1.51	0.00	-1.27	1.27	†	†
MC	12	M150901	0.96	0.72	0.00	†	†	†	†
MC	13	M151001	1.29	0.49	0.20	†	†	†	†
MC	14	M151101	1.76	0.55	0.07	†	†	†	†
MC	15	M151201	0.93	0.87	0.00	†	†	†	†
MD	1	M151301	0.94	-0.18	0.17	†	†	†	†
MD	2	M151401	1.31	1.03	0.11	†	†	†	†
MD	3	M151501	0.89	0.02	0.00	0.21	-0.21	†	†
MD	4	M151601	1.09	1.32	0.00	-0.63	0.63	†	†
MD	5	M151701	1.54	0.18	0.23	†	†	†	†
MD	6	M151801	0.79	-0.50	0.20	†	†	†	†
MD	7	M151901	1.28	0.56	0.15	†	†	†	†
MD	8	M152001	1.63	0.88	0.08	†	†	†	†
MD	9	M152101	0.99	0.84	0.22	†	†	†	†
MD	10	M152201	1.53	0.32	0.20	†	†	†	†
MD	11	M152301	0.74	1.70	0.16	†	†	†	†
MD	12	M152401	1.49	0.75	0.08	†	†	†	†
MD	13	M152501	2.49	1.00	0.14	†	†	†	†
MD	14	M152602	1.01	0.53	0.00	1.33	-2.02	0.69	†
ME	1	M221201	0.87	0.15	0.21	†	†	†	†
ME	2	M221202	1.21	0.45	0.19	†	†	†	†
ME	3	M221203	0.77	0.97	0.00	1.66	-1.66	†	†
ME	4	M221204	0.73	1.08	0.00	1.17	-1.17	†	†
ME	5	M221301	1.10	0.08	0.22	†	†	†	†
ME	6	M221401	1.34	1.08	0.19	†	†	†	†
ME	7	M221501	0.94	-0.87	0.19	†	†	†	†
ME	8	M221601	1.65	0.73	0.19	†	†	†	†
ME	9	M221701	0.49	0.12	0.00	†	†	†	†
ME	10	M221801	1.20	0.48	0.16	†	†	†	†
ME	11	M221901	1.00	0.95	0.20	†	†	†	†
ME	12	M222001	1.24	1.42	0.24	†	†	†	†
ME	13	M222101	1.74	1.99	0.22	†	†	†	†
ME	14	M222201	0.90	1.13	0.12	†	†	†	†
ME	15	M222301	0.79	0.60	0.00	1.01	-1.01	†	†
MF	1	M140401	1.53	-0.29	0.12	†	†	†	†
MF	2	M140501	1.92	0.05	0.21	†	†	†	†
MF	3	M140601	0.74	-0.93	0.20	†	†	†	†
MF	4	M140701	1.01	0.14	0.16	†	†	†	†
MF	5	M140801	2.15	0.63	0.19	†	†	†	†
MF	6	M140901	1.49	0.34	0.18	†	†	†	†
MF	7	M141001	1.60	0.25	0.17	†	†	†	†
MF	8	M141101	1.12	-0.33	0.13	†	†	†	†
MF	9	M141201	1.11	0.75	0.18	†	†	†	†

See notes at end of table.

Table 3.8. IRT item parameter estimates for fixed parameter calibration, NAEP grade 8 mathematics items: 2011—Continued

Block	Item	NAEP ID	a_j	b_j	c_j	d_{j1}	d_{j2}	d_{j3}	d_{j4}
MF	10	M141301	0.72	0.84	0.00	-0.79	0.79	†	†
MF	11	M141401	0.86	0.73	0.17	†	†	†	†
MF	12	M141501	1.47	0.14	0.27	†	†	†	†
MF	13	M141601	1.40	0.70	0.00	†	†	†	†
MF	14	M141701	1.67	0.96	0.14	†	†	†	†
MF	15	M141801	1.84	0.81	0.21	†	†	†	†
MF	16	M141901	0.74	1.20	0.00	1.39	0.61	-0.43	-1.57
MG	1	M163801	0.81	-1.91	0.19	†	†	†	†
MG	2	M120701	0.81	-1.12	0.19	†	†	†	†
MG	3	M166001	0.88	-0.08	0.18	†	†	†	†
MG	4	M170101	0.30	-0.28	0.00	-3.01	3.01	†	†
MG	5	M164401	1.64	-0.30	0.17	†	†	†	†
MG	6	M169401	1.49	0.66	0.19	†	†	†	†
MG	7	M168201	1.17	0.58	0.19	†	†	†	†
MG	8	M166101	1.91	1.17	0.14	†	†	†	†
MG	9	M168701	0.73	0.21	0.00	-0.63	0.63	†	†
MG	10	M164201	1.69	1.07	0.19	†	†	†	†
MG	11	M170201	1.31	0.55	0.24	†	†	†	†
MG	12	M165301	1.49	0.68	0.00	0.10	-0.10	†	†
MG	13	M164801	1.39	0.80	0.23	†	†	†	†
MG	14	M167001	0.72	1.77	0.00	†	†	†	†
MG	15	M168401	1.30	1.61	0.11	†	†	†	†
MG	16	M1685CL	0.92	1.24	0.00	0.43	0.45	-0.61	-0.27
MH	1	M170301	1.19	-0.02	0.16	†	†	†	†
MH	2	M167801	1.14	0.40	0.19	†	†	†	†
MH	3	M163301	1.05	-0.24	0.17	†	†	†	†
MH	4	M170401	0.82	0.20	0.24	†	†	†	†
MH	5	M164501	1.33	0.26	0.16	†	†	†	†
MH	6	M164601	0.87	0.80	0.00	-2.39	2.39	†	†
MH	7	M165101	1.76	1.13	0.21	†	†	†	†
MH	8	M122501	1.52	0.92	0.18	†	†	†	†
MH	9	M166301	0.74	-0.13	0.00	-1.27	1.27	†	†
MH	10	M120901	1.44	0.80	0.08	†	†	†	†
MH	11	M170501	0.61	-0.06	0.00	-0.97	0.97	†	†
MH	12	M166601	1.30	0.58	0.20	†	†	†	†
MH	13	M164901	1.26	0.94	0.10	†	†	†	†
MH	14	M166901	2.20	1.11	0.35	†	†	†	†
MH	15	M1699CL	0.61	-0.05	0.00	0.40	1.78	-0.22	-1.96
MI	1	M119301	1.25	-0.53	0.16	†	†	†	†
MI	2	M166401	2.31	0.40	0.31	†	†	†	†
MI	3	M170601	0.74	-0.25	0.18	†	†	†	†
MI	4	M119101	1.85	0.11	0.23	†	†	†	†
MI	5	M168901	1.52	0.23	0.00	†	†	†	†
MI	6	M125301	0.85	0.11	0.18	†	†	†	†
MI	7	M166701	0.90	0.35	0.21	†	†	†	†
MI	8	M165501	0.84	-0.35	0.00	0.06	-0.06	†	†
MI	9	M166801	0.83	0.88	0.15	†	†	†	†
MI	10	M170701	2.21	0.66	0.24	†	†	†	†
MI	11	M165001	1.13	1.10	0.00	†	†	†	†
MI	12	M124901	0.80	0.25	0.19	†	†	†	†
MI	13	M170801	1.25	1.88	0.00	-1.03	1.03	†	†
MI	14	M124001	1.52	1.15	0.11	†	†	†	†
MI	15	M1657CL	0.71	0.98	0.00	-0.63	2.21	-2.51	0.93
MJ	1	M222401	0.97	-0.77	0.18	†	†	†	†
MJ	2	M222501	1.44	0.30	0.20	†	†	†	†

See notes at end of table.

Table 3.8. IRT item parameter estimates for fixed parameter calibration, NAEP grade 8 mathematics items: 2011—Continued

Block	Item	NAEP ID	a_j	b_j	c_j	d_{j1}	d_{j2}	d_{j3}	d_{j4}
MJ	3	M222601	1.54	-0.09	0.14	†	†	†	†
MJ	4	M222701	0.62	0.18	0.00	-3.16	3.16	†	†
MJ	5	M222801	2.25	0.21	0.12	†	†	†	†
MJ	6	M222901	1.53	0.04	0.12	†	†	†	†
MJ	7	M223001	1.31	0.27	0.11	†	†	†	†
MJ	8	M223101	0.84	-0.26	0.00	-1.29	1.29	†	†
MJ	9	M223201	1.57	1.30	0.29	†	†	†	†
MJ	10	M223301	0.60	0.61	0.00	-0.29	0.29	†	†
MJ	11	M223401	0.46	1.47	0.20	†	†	†	†
MJ	12	M223501	0.99	-0.04	0.25	†	†	†	†
MJ	13	M223601	1.46	0.62	0.13	†	†	†	†
MJ	14	M223701	2.41	1.19	0.16	†	†	†	†
MJ	15	M223801	0.64	0.45	0.00	-0.37	0.37	†	†
MK	1	M163101	1.00	-0.10	0.00	0.57	-0.57	†	†
MK	2	M170901	1.28	-0.05	0.17	†	†	†	†
MK	3	M122701	0.73	0.28	0.20	†	†	†	†
MK	4	M119601	0.63	0.20	0.17	†	†	†	†
MK	5	M121801	0.65	-0.36	0.23	†	†	†	†
MK	6	M171001	0.40	0.76	0.20	†	†	†	†
MK	7	M169201	1.91	0.17	0.27	†	†	†	†
MK	8	M171101	1.93	0.40	0.15	†	†	†	†
MK	9	M168301	0.94	0.74	0.00	-0.48	0.48	†	†
MK	10	M169101	0.67	1.87	0.27	†	†	†	†
MK	11	M162901	0.89	0.01	0.00	0.72	-0.72	†	†
MK	12	M164701	1.10	1.36	0.26	†	†	†	†
MK	13	M167301	1.82	0.87	0.16	†	†	†	†
MK	14	M165201	2.03	1.04	0.15	†	†	†	†
MK	15	M167901	0.72	1.17	0.00	-1.14	1.14	†	†
MK	16	M171201	1.21	0.98	0.16	†	†	†	†
MK	17	M104901	1.00	1.33	0.00	0.54	-0.54	†	†
ML	1	M152701	1.49	-0.19	0.17	†	†	†	†
ML	2	M152801	0.87	0.32	0.18	†	†	†	†
ML	3	M152901	1.61	0.55	0.15	†	†	†	†
ML	4	M153001	1.18	-0.36	0.17	†	†	†	†
ML	5	M153101	1.37	0.42	0.18	†	†	†	†
ML	6	M153201	0.47	-0.78	0.00	-1.11	1.11	†	†
ML	7	M153301	1.99	0.93	0.13	†	†	†	†
ML	8	M153401	1.03	0.90	0.16	†	†	†	†
ML	9	M153501	0.77	-0.28	0.18	†	†	†	†
ML	10	M153601	1.01	-0.54	0.00	†	†	†	†
ML	11	M153701	1.65	0.41	0.08	†	†	†	†
ML	12	M153801	2.02	0.87	0.15	†	†	†	†
ML	13	M153901	0.74	0.70	0.00	-0.13	0.13	†	†
ML	14	M154001	0.90	0.51	0.13	†	†	†	†
ML	15	M154101	1.22	0.63	0.20	†	†	†	†
ML	16	M154201	1.02	1.80	0.24	†	†	†	†
ML	17	M154301	2.32	1.11	0.17	†	†	†	†

† Not applicable.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 3.9. IRT item parameter estimates for fixed parameter calibration, NAEP grade 8 science items: 2011

Block	Item	NAEP ID	a_j	b_j	c_j	d_{j1}	d_{j2}	d_{j3}	d_{j4}
SC	1	K114201	0.59	-0.16	0.31	†	†	†	†
SC	2	K114101	1.13	1.39	0.29	†	†	†	†
SC	3	K113901	0.66	0.13	0.00	0.73	-0.73	†	†
SC	4	K114001	0.98	0.20	0.26	†	†	†	†
SC	5	K114002	0.93	0.66	0.19	†	†	†	†
SC	6	K113401	0.45	0.85	0.00	0.03	-0.25	1.87	-1.65
SC	7	K113201	0.47	-0.22	0.00	1.89	-1.30	-0.60	†
SC	8	K113102	0.80	0.88	0.29	†	†	†	†
SC	11	K113603	0.64	1.13	0.21	†	†	†	†
SC	12	K113001	1.17	-0.06	0.26	†	†	†	†
SC	13	K113801	0.87	0.21	0.27	†	†	†	†
SC	14	K113701	0.58	1.01	0.00	0.56	0.53	-0.19	-0.90
SC	15	K134201	0.56	0.21	0.00	-0.27	-0.14	0.41	†
SC	16	K113301	0.97	0.87	0.32	†	†	†	†
SC	17	K113501	0.72	0.64	0.27	†	†	†	†
SD	1	K117801	1.46	-0.41	0.28	†	†	†	†
SD	2	K114601	0.52	0.34	0.00	-0.30	0.39	-0.01	-0.08
SD	3	K122201	1.56	0.91	0.32	†	†	†	†
SD	4	K122301	0.83	0.86	0.26	†	†	†	†
SD	5	K122302	0.93	0.39	0.32	†	†	†	†
SD	6	K122303	1.01	1.44	0.23	†	†	†	†
SD	7	K122304	1.72	0.89	0.25	†	†	†	†
SD	8	K116701	1.45	0.99	0.27	†	†	†	†
SD	9	K122901	1.18	0.08	0.26	†	†	†	†
SD	10	K118901	1.20	-0.44	0.28	†	†	†	†
SD	11	K122402	0.65	1.57	0.00	0.94	-0.94	†	†
SD	12	K123001	0.84	0.46	0.24	†	†	†	†
SD	13	K122801	0.88	1.23	0.00	0.37	-0.37	†	†
SD	14	K122001	1.59	0.83	0.27	†	†	†	†
SD	15	K122602	0.58	1.21	0.00	-0.06	0.06	†	†
SD	16	K125201	0.47	0.48	0.00	0.08	0.62	-1.04	0.34
SD	17	K122501	0.54	0.63	0.26	†	†	†	†
SD	18	K121801	1.86	1.47	0.24	†	†	†	†
SE	1	K120701	0.89	0.07	0.37	†	†	†	†
SE	2	K120601	1.98	1.41	0.27	†	†	†	†
SE	3	K154501	1.09	0.31	0.27	†	†	†	†
SE	4	K121701	0.59	2.37	0.29	†	†	†	†
SE	5	K121301	0.69	1.84	0.00	-0.08	0.16	-0.08	†
SE	6	K117601	0.80	1.45	0.28	†	†	†	†
SE	7	K120801	1.12	0.98	0.16	†	†	†	†
SE	8	K120802	0.40	1.56	0.00	-0.05	0.05	†	†
SE	9	K118501	0.63	1.16	0.00	1.14	-0.12	-1.02	†
SE	10	K154601	0.54	0.51	0.00	0.63	-0.63	†	†
SE	11	K120901	1.16	0.65	0.22	†	†	†	†
SE	12	K121401	0.88	0.50	0.37	†	†	†	†
SE	13	K121402	1.01	0.24	0.26	†	†	†	†
SE	14	K121403	2.35	1.41	0.11	†	†	†	†
SE	15	K154701	0.68	1.27	0.00	2.68	-0.34	-0.54	-1.80
SE	16	K154801	1.13	-0.37	0.25	†	†	†	†
SE	17	K121201	1.27	0.04	0.24	†	†	†	†
SE	18	K120501	1.62	0.04	0.26	†	†	†	†
SF	1	K125101	0.46	-0.72	0.26	†	†	†	†
SF	2	K124501	1.86	0.82	0.26	†	†	†	†

See notes at end of table.

Table 3.9. IRT item parameter estimates for fixed parameter calibration, NAEP grade 8 science items: 2011—Continued

Block	Item	NAEP ID	a_j	b_j	c_j	d_{j1}	d_{j2}	d_{j3}	d_{j4}
SF	3	K124601	0.74	1.75	0.29	†	†	†	†
SF	4	K123801	0.73	0.71	0.00	0.58	-0.06	-0.52	†
SF	5	K123802	0.53	0.62	0.00	0.63	-0.63	†	†
SF	6	K154901	0.88	0.67	0.23	†	†	†	†
SF	7	K122101	0.54	1.26	0.00	0.38	1.08	-0.01	-1.46
SF	8	K125001	1.12	1.37	0.29	†	†	†	†
SF	9	K125002	0.99	0.80	0.25	†	†	†	†
SF	10	K125003	0.40	1.04	0.00	0.84	-0.84	†	†
SF	11	K155001	1.65	-0.02	0.31	†	†	†	†
SF	12	K155101	1.35	1.45	0.22	†	†	†	†
SF	13	K155201	1.10	0.45	0.20	†	†	†	†
SF	14	K125401	0.48	1.88	0.00	1.84	-1.84	†	†
SF	15	K125402	0.53	0.57	0.00	0.49	-0.49	†	†
SF	16	K155301	1.59	0.72	0.37	†	†	†	†
SF	17	K124401	1.52	0.43	0.29	†	†	†	†
SG	1	K111301	1.21	1.29	0.27	†	†	†	†
SG	2	K117401	1.45	0.93	0.13	†	†	†	†
SG	3	K155401	1.01	0.36	0.24	†	†	†	†
SG	4	K1107CL	0.74	-0.02	0.00	1.26	-0.04	-1.22	†
SG	5	K114501	1.59	1.50	0.19	†	†	†	†
SG	6	K110601	1.35	-0.48	0.22	†	†	†	†
SG	7	K110602	1.03	0.36	0.23	†	†	†	†
SG	8	K110603	1.58	1.41	0.24	†	†	†	†
SG	9	K123901	0.67	0.79	0.00	0.91	1.41	-0.38	-1.94
SG	10	K110401	0.55	1.11	0.00	-0.23	0.31	-0.08	†
SG	12	K110501	0.92	1.20	0.00	0.75	-0.75	†	†
SG	13	K121001	0.94	1.52	0.00	0.35	-0.35	†	†
SG	14	K1210CL	0.55	0.57	0.00	-0.84	1.66	-0.83	†
SG	15	K110801	0.92	1.67	0.23	†	†	†	†
SG	16	K110101	1.27	1.40	0.31	†	†	†	†
SH	1	K112801	1.00	-0.49	0.29	†	†	†	†
SH	2	K112501	0.50	0.76	0.27	†	†	†	†
SH	3	K111901	0.79	1.53	0.37	†	†	†	†
SH	4	K155501	1.29	0.61	0.17	†	†	†	†
SH	5	K155502	1.81	0.65	0.23	†	†	†	†
SH	6	K112401	0.70	1.34	0.00	-0.55	-2.49	3.05	†
SH	7	K112001	1.45	0.69	0.20	†	†	†	†
SH	8	K119701	1.74	0.69	0.25	†	†	†	†
SH	9	K155601	0.30	2.30	0.00	-1.56	1.54	1.17	-1.14
SH	10	K117201	0.76	0.47	0.31	†	†	†	†
SH	11	K117202	0.72	0.09	0.00	†	†	†	†
SH	12	K155701	0.59	1.93	0.00	0.46	-0.46	†	†
SH	13	K155702	0.74	2.15	0.00	1.04	-1.04	†	†
SH	14	K111601	0.90	0.49	0.22	†	†	†	†
SH	15	K112901	0.70	1.90	0.24	†	†	†	†
SH	16	K155801	1.82	0.13	0.33	†	†	†	†
SH	17	K112701	0.82	0.53	0.26	†	†	†	†
SH	18	K155901	0.65	0.25	0.28	†	†	†	†
SI	1	K115101	1.01	0.12	0.28	†	†	†	†
SI	2	K114401	1.49	0.63	0.20	†	†	†	†
SI	3	K115201	0.86	0.90	0.00	0.12	0.61	-0.73	†
SI	4	K114901	1.55	0.57	0.24	†	†	†	†
SI	5	K1560CL	0.38	2.01	0.00	1.66	-1.66	†	†

See notes at end of table.

Table 3.9. IRT item parameter estimates for fixed parameter calibration, NAEP grade 8 science items: 2011—Continued

Block	Item	NAEP ID	a_j	b_j	c_j	d_{j1}	d_{j2}	d_{j3}	d_{j4}
SI	6	K156101	0.55	1.84	0.00	-1.49	1.56	-0.06	†
SI	7	K115501	1.88	1.57	0.18	†	†	†	†
SI	8	K114802	0.78	0.49	0.00	-0.27	0.27	†	†
SI	9	K114801	0.46	1.78	0.00	-0.69	1.98	-1.28	†
SI	10	K115301	1.24	-0.12	0.25	†	†	†	†
SI	12	K124102	0.65	0.75	0.00	0.57	-0.57	†	†
SI	13	K111401	1.40	1.48	0.30	†	†	†	†
SI	14	K156201	1.46	0.77	0.37	†	†	†	†
SK	1	K119401	1.05	-0.02	0.29	†	†	†	†
SK	2	K123201	0.77	0.71	0.28	†	†	†	†
SK	3	K119201	1.10	0.85	0.33	†	†	†	†
SK	4	K156701	0.97	1.66	0.22	†	†	†	†
SK	5	K123601	0.47	-0.18	0.00	1.10	-2.30	1.27	-0.07
SK	6	K111801	1.17	0.98	0.00	0.45	-0.45	†	†
SK	7	K111802	0.95	0.98	0.00	-0.12	0.12	†	†
SK	8	K111803	1.01	1.57	0.00	0.25	-0.25	†	†
SK	9	K120101	1.88	1.21	0.15	†	†	†	†
SK	10	K117001	0.57	1.02	0.00	0.31	-0.31	†	†
SK	11	K1170CL	0.46	1.52	0.00	0.09	-0.28	0.19	†
SK	12	K117005	0.99	-0.45	0.34	†	†	†	†
SK	13	K117006	1.78	1.00	0.27	†	†	†	†
SK	14	K119902	0.49	1.67	0.00	1.53	-1.53	†	†
SK	15	K120301	0.94	0.74	0.27	†	†	†	†
SK	16	K119601	0.62	1.16	0.25	†	†	†	†
SL	1	K118001	1.06	-0.11	0.29	†	†	†	†
SL	2	K156801	1.67	0.76	0.18	†	†	†	†
SL	3	K156901	1.60	1.59	0.25	†	†	†	†
SL	4	K118601	1.46	1.15	0.14	†	†	†	†
SL	5	K124001	0.82	0.33	0.24	†	†	†	†
SL	6	K116901	0.55	-0.08	0.00	-0.37	0.37	†	†
SL	7	K123301	0.91	0.56	0.00	†	†	†	†
SL	9	K119505	1.07	0.52	0.00	†	†	†	†
SL	10	K124301	0.78	1.64	0.19	†	†	†	†
SL	11	K123401	0.71	0.59	0.00	0.14	-0.39	0.25	†
SL	12	K112601	0.79	1.96	0.00	0.56	-0.56	†	†
SL	13	K118801	0.83	1.00	0.19	†	†	†	†
SL	14	K157101	1.04	0.95	0.30	†	†	†	†
SL	15	K115601	0.97	0.34	0.26	†	†	†	†

† Not applicable.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 3.10. Predicted state TIMSS average scores, predicted benchmark results, and standard errors from calibration linking in eighth-grade mathematics, by validation state: 2011

	Predicted TIMSS mathematics results from calibration linking									
	Average score		Percentage of students reaching TIMSS low international benchmark		Percentage of students reaching TIMSS intermediate international benchmark		Percentage of students reaching TIMSS high international benchmark		Percentage of students reaching TIMSS advanced international benchmark	
Validation state	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error
Alabama	478	4	84	1.6	54	2.2	17	1.7	2	0.9
California	486	3.5	85	1.1	56	1.6	22	1.3	5	0.7
Colorado	526	3.5	95	1.1	76	1.5	39	1.9	9	1.2
Connecticut	516	3.6	94	1	71	2.2	34	1.9	7	1.2
Florida	496	3.2	90	1.2	62	1.8	24	1.7	4	0.6
Indiana	513	3.4	94	0.8	71	1.7	31	2	5	0.8
Massachusetts	540	3.3	96	0.6	82	1.5	46	2.2	11	1.2
Minnesota	533	3.3	95	0.6	80	1.4	43	2.1	10	1.5
North Carolina	515	3.5	93	1.5	70	1.9	33	1.9	7	1.3

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 3.11. Predicted state TIMSS average scores, predicted benchmark results, and standard errors from calibration linking in eighth-grade science, by validation state: 2011

	Predicted TIMSS science results from calibration linking									
	Average score		Percentage of students reaching TIMSS low international benchmark		Percentage of students reaching TIMSS intermediate international benchmark		Percentage of students reaching TIMSS high international benchmark		Percentage of students reaching TIMSS advanced international benchmark	
Validation state	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error
Alabama	497	3.9	87	1.4	64	2	27	2	4	1
California	498	3.7	86	1.2	63	2	29	1.8	6	0.8
Colorado	546	3.9	96	1.1	82	1.8	51	2.5	15	1.8
Connecticut	532	3.5	94	0.9	77	1.7	44	2.1	11	1.4
Florida	517	3.5	91	1.2	71	1.8	37	2.6	8	0.9
Indiana	527	3.1	94	1	77	1.5	42	2	8	1
Massachusetts	547	3.3	95	0.7	83	1.4	53	1.7	16	1.2
Minnesota	546	3.3	96	0.9	84	1.3	52	1.8	13	1.4
North Carolina	515	3.4	92	1.8	71	1.6	35	1.8	7	0.9

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

References

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

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.

Foy, P., Brossman, B., and Galia, J. (2012). Scaling the TIMSS and PIRLS 2011 Achievement Data. In M.O. Martin, and 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.

Haebara, T. (1980). Equating Logistic Ability Scales by a Weighted Least Squares Method. *Japanese Psychological Research*, 22, 144–149.

Hanson, B.A., and Béguin, A.A. (2002). Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate Versus Concurrent Estimation in the Common-Item Equating Design. *Applied Psychological Measurement*, *26*, 3-24.

Johnson, E.G., and Rust, K.F. (1992). Population Inferences and Variance Estimation for NAEP Data. *Journal of Educational Statistics*, *17*, 175–190.

Kang, T., and Petersen, N.S. (2009). *Linking Item Parameters to a Base Scale*. Iowa City, IA: ACT Research Report Series 20090-2.

Kim, S. (2006). A Comparative Study of IRT Fixed Parameter Calibration Methods. *Journal of Educational Measurement*, *43(4)*, 355-381.

Kolen, M.J., and Brennan, R.L. (2004). *Test Equating, Scaling, and Linking*. New York, NY: Springer.

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

Mislevy, R.J. (1991). Randomization-Based Inference About Latent Variables From Complex Samples. *Psychometrika,* 56(2), 177-196.

Mislevy, R.J., Beaton, A.E., Kaplan, B.A., and Sheehan, K.M. (1992). Estimating Population Characteristics From Sparse Matrix Samples of Item Responses. *Journal of Educational Measurement*, *29*, 133-161.

Mislevy, R., Johnson, E., and 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.

Muraki, E. and Bock, R.D. (1991). *PARSCALE: Parameter Scaling of Rating Data [Computer Program]*. Chicago, IL: Scientific Software, Inc.

Neidorf, T.S., Binkley, M., Gattis, K., and 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). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

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). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Provasnik, S., Kastberg, D., Ferraro, D., Lemanski, N., Roey, S., and 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). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Rogers, A., Gregory, K., Davis, S., and Kulick, E. (2006a). *User's Guide to NAEP Model-Based P-value Programs*. Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Rogers, A., Tang, C., Lin, M.-J., and Kandathil, M. (2006b). *DGROUP [Computer Software]*. Princeton, NJ: Educational Testing Service.

Stocking, M.L., and Lord, F.M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement*, *7*, 201–210.

Thomas, N. (1994). *CGROUP and BGROUP: Modifications of the MGROUP Program to Estimate Group Effects in Multivariate Models [Computer Programs]*. Princeton, NJ: Educational Testing Service.

von Davier, M., Gonzalez, E., and 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. Princeton, NJ: IEA-ETS Research Institute. Retrieved August 19, 2013, from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf.

Chapter 4: Linking NAEP and TIMSS Through Projection Linking

This chapter describes the linking of the 2011 NAEP to the 2011 TIMSS assessments in mathematics and science through statistical projection.

Projection Linking: Use and Methods

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. 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. Projection linking uses the linking sample to model the relationships between scores on the two assessments.

Projection linking can be implemented using various statistical methods, including regression (Pashley and Phillips 1993) and direct estimation of joint score distributions (Mislevy 1992). To implement the regression approach in linking two tests, the score from one test can be used as the explanatory variable and the score from the other test as the response variable. Note that NAEP and TIMSS both use a combination of Item Response Theory (IRT) models and latent regression population models to provide estimated distributions of underlying performance for student groups of interest and to impute student proficiency values or plausible values (e.g., von Davier et al. 2007). This regression function derived from the linking sample could then serve as the projection linking function. However, each student in the linking sample answered a small portion of the NAEP and TIMSS cognitive item pools. As a result, reliable individual proficiency estimates could not be obtained.

Furthermore, in both NAEP and TIMSS, consistent estimates of proficiency (e.g., averages) and the dispersion of proficiency (e.g., variances) in various reporting groups of interest are estimated by including the variables of interest as predictors in the population model (Mislevy et al. 1992). Statistics based on variables not included in the population model are subject to asymptotic (secondary) biases. The bias typically results in an underestimate of the effect of the variables not included (Mislevy 1984, 1985). Analogously, deriving two independent sets of plausible values for NAEP and TIMSS does not consider the relationship between NAEP and TIMSS in their population models. Consequently, estimating a regression function with NAEP scores as explanatory variables and TIMSS scores as responses could potentially underestimate the relationship between the two assessments of interest.

In this linking study, the joint NAEP-TIMSS score distribution and the relationship between the two scales was directly estimated from the braided-booklet sample. Compared to a regression type of projection using independently generated NAEP and TIMSS scores, the correlation between NAEP and TIMSS estimated from the joint distribution is expected to be more accurate. Using this approach, 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.

There have been two previous studies that used statistical projection to link NAEP and other large-scale survey assessments. A regression type of projection linking was used in both studies. Pashley and Phillips (1993) linked the 1991 International Assessment of Educational Progress (IAEP) to the 1992 NAEP mathematics assessment in order to evaluate countries' performance with respect to the NAEP benchmark levels. The linking sample consisted of a subsample of students that had been selected to participate in the 1992 NAEP who were re-tested with the 1991 IAEP assessment. The reported percentages of U.S. students at or above each NAEP achievement level were within or close to the confidence intervals of the predicted values from projection linking. However, there was no empirical evidence available to evaluate the extent to which the prediction relationship generalized to other countries that were assessed with IAEP but not with NAEP.

In the second study, Johnson et al. (2003) linked the 2000 NAEP assessment to the 1999 TIMSS assessment to project U.S. states' average TIMSS mathematics and science scores from their NAEP scores in the same subjects. The linking sample comprised a group of students who had been assessed by NAEP in 2000 and were re-administered the 1999 TIMSS instrument a few months after the 2000 NAEP administration. When 12 states also participated in the 1999 grade 8 TIMSS assessments at the state level, the projection linkage consistently under-predicted their actual TIMSS scores. Johnson et al. (2003) attributed (hypothesized) that the under-prediction was due, in part, to the differences in administration conditions experienced by the students in the linking sample and those in the 1999 TIMSS assessment.

One important difference between the current NAEP-TIMSS linking study design and the designs used in the previous studies mentioned above is that the current study administered the braided-booklet samples (both NAEP and TIMSS test items) at the same time and under the same conditions. Those braided-booklet samples were used to develop the projection linking functions. As discussed in Chapter 2, the current study design included braided-booklet samples from both NAEP and TIMSS administration windows. In addition to responding to cognitive items, the braided-booklet samples of students assessed during the NAEP administration window were given the NAEP survey questionnaires. Likewise, the braided-booklet sample of students under the TIMSS administration

window took the TIMSS survey questionnaires. Therefore, the current study took a different approach than previous studies: the joint NAEP-TIMSS population-structure model was directly estimated by using survey questionnaires and students' responses to the cognitive items. In addition, it took into account the relationship between the two assessments.

Given the availability of the braided-booklet samples under both NAEP and TIMSS administration windows, it was possible to derive two projection functions for each subject domain and to 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 mathematics or science, and those in the TIMSS window took items from both subjects. In addition, as will be discussed later in this chapter, there is empirical evidence of discrepancies between the predicted results using projection functions from the NAEP and TIMSS window braided-booklet samples. Therefore, the braided-booklet samples across assessment windows were not combined in deriving projection functions.

Using Projection to Link NAEP and TIMSS

In this section, the step-by-step procedures that were used to carry out projection linking are described. This section also discusses the relevant similarities and differences associated with implementing the linking procedures among the braided-booklet samples.

Step 1: Applying the NAEP and TIMSS latent proficiency scale parameters to the braided-booklet sample item responses

The NAEP and TIMSS latent proficiency scales were both estimated based on a combination of IRT models (Allen, Donoghue, and Schoeps 2001; Foy, Brossman, and Galia 2012). For dichotomously-scored items, two-parameter and three-parameter logistic models (Lord and Novick 1968) were used, while for polytomously-scored items the generalized partial-credit model (Muraki 1992) was used.

As described in Chapter 2, the braided instrument that was administered to the braided-booklet samples included the complete pool of items administered in the 2011 NAEP and TIMSS assessments. The operational 2011 NAEP item parameter estimates[6] were used to calculate NAEP proficiency estimates for the braided-booklet samples. Likewise, the operational 2011 TIMSS item parameter estimates from the overall TIMSS mathematics and science scales were applied in the

[6] 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 linking 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 linking samples.

calculation of TIMSS proficiency estimates. The fit of the IRT models was carefully checked by multiple procedures, including graphical comparisons of the empirical item response functions to the model-based (theoretical) curves and comparisons of observed and model-predicted proportions of students obtaining a particular score on each item (Rogers et al. 2006a). The evaluation of the IRT models was done using the ETS proprietary version of BILOG/PARSCALE software (Muraki and Bock 1991). The IRT model fit for the NAEP items in the braided-booklet samples was reasonable and comparable to the model fit in the 2011 NAEP national samples, taking sample size into account. The IRT model fit for the TIMSS items in the braided-booklet samples was also comparable to the model fit observed in the 2011 TIMSS U.S. national sample.

Step 2: Estimating the projection function for the braided-booklet sample

In the second step, a procedure called "conditioning"[7] was employed to estimate the joint NAEP and TIMSS proficiency distribution through a latent regression model. The latent regression model was based on the IRT parameters from step 1 and the student responses to the subset of items they received, as well as other relevant and available background information. For the mathematics braided-booklet sample in the NAEP administration window, a bivariate latent regression model was used to estimate this joint distribution of NAEP and TIMSS mathematics scores. The analysis was conducted using the DGROUP set of programs (Thomas 1994; Rogers et al. 2006b). The DGROUP program uses the EM algorithm to estimate all population parameters simultaneously. This program represents the latent proficiencies in the model through "plausible values"—random draws from the predictive scale score distributions for each respondent on the IRT scale (Mislevy 1991; von Davier, Gonzalez, and Mislevy 2009). These plausible values can subsequently be used to represent probabilities in the joint NAEP-TIMSS proficiency distribution and allow unbiased group-level estimates. In this study, 20 plausible values were drawn per respondent on NAEP mathematics and TIMSS mathematics, respectively.

The same conditioning procedures were used to estimate the joint distribution of NAEP and TIMSS science proficiencies from the science braided-booklet sample in the NAEP administration window. Students in the TIMSS window braided-booklet sample were administered items from both subjects and assessments; therefore, 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: Transforming the proficiency distribution for the braided-booklet sample from the IRT metrics to the reporting metric

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 was to place the proficiency distributions on the NAEP and TIMSS reporting metrics.

[7] Full descriptions of the conditioning procedure can be found in Beaton, 1987; Mislevy, Johnson, and Muraki, 1992; and Mislevy et al. 1992. The description of the procedure is also available at http://nces.ed.gov/nationsreportcard/tdw/analysis/est.aspx.

Both NAEP and TIMSS apply linear transformations to transform results from IRT metrics to the appropriate reporting metrics (Allen, Donoghue, and Schoeps 2001; Foy, Brossman, and Galia 2012). Essentially, 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. Student plausible values are used in computing the means and standard deviations of the score distribution. For TIMSS, as there were five plausible values per student, a total of five sets of transformation parameters were available. Those transformation parameter estimates, $\hat{A}_i$ and $\hat{B}_i$, were then used in a linear transformation equation as follows:

$$PV_{i,\text{Target}} = \hat{A}_i \times PV_{i,\text{ Calibrated}} + \hat{B}_i \quad (4.1)$$

Where

- i = 1,2,3,4,5;
- $PV_{i,\text{Target}}$ was the plausible value i on the transformed TIMSS reporting scale;
- $PV_{i,\text{Calibrated}}$ 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 one set of transformation parameters $\hat{A}$ and $\hat{B}$, that is computed by first averaging the means and standard deviations of the score distribution obtained from both metrics.

For the braided-booklet samples, given that the original 2011 NAEP item parameter estimates were used in estimating the plausible values on the calibration scale, the transformation parameter estimates $\hat{A}$ and $\hat{B}$ from the operational 2011 NAEP analysis[8] were applied 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 TIMSS 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. The transformation parameter estimates from 2011 NAEP and TIMSS are listed in tables 4.1 and 4.2.

Step 4: Smoothing the projection functions from the braided-booklet sample

Taking the NAEP and TIMSS plausible values obtained in step 3, the discrete joint NAEP-TIMSS

[8] For 2011 NAEP science, an overall univariate scale was established in the operational analysis. Therefore the transformation parameter estimates $\hat{A}$ and $\hat{B}$ 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 parameter estimates $\hat{A}$ and $\hat{B}$ obtained from the overall NAEP mathematics scale were applied to the linking samples.

proficient score distribution for each subject was smoothed using a bivariate continuous exponential family of distributions (Haberman 2011). For the braided-booklet samples, it was specified that the first four moments of the bivariate continuous exponential family distribution should match the first four moments of the joint NAEP-TIMSS score distribution. 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: Predicting TIMSS scores for all the states

The projection functions derived in step 4 were used to predict TIMSS 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 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 state-level average scores and percentages of students reaching each of the TIMSS international benchmark levels.

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 were estimates of how students in the 2011 NAEP sample would have performed if they had taken TIMSS during the NAEP window under NAEP conditions, to the extent that differences in testing conditions were accounted for in the projection functions. However, it is of more interest to determine how they would have performed if they had taken TIMSS during the TIMSS window and under TIMSS conditions, as that would facilitate comparisons to other countries and subnational education systems that participated in TIMSS. Therefore, the distributions of predicted TIMSS plausible values from the 2011 NAEP national sample were then aligned (through a mean-sigma transformation adjustment) to the distribution of TIMSS plausible values from the 2011 TIMSS U.S. national sample. The adjustment was conducted separately for mathematics and science.

$$PV_{\text{Target\_with adjustment}} = \hat{A} \times PV_{\text{Target}} + \hat{B} \tag{4.2}$$

Where

- PV_{Target} was the plausible value on the TIMSS reporting scale from step 5 of projection linking;
- $PV_{\text{Target\_with adjustment}}$ 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 adjustment function parameter estimates.

Table 4.3 contains the linear adjustment function parameter estimates for both mathematics and science, and for the different projection functions obtained from the NAEP and TIMSS window braided-booklet samples.

Standard Error Estimation

TIMSS eighth-grade mathematics and science achievement results for the participating countries, subnational education systems, and the nine validation states were released in December 2012. In addition to reporting average scores, TIMSS reports on the performance of students at four international benchmarks for each subject and grade: Advanced, High, Intermediate, and Low. The standard errors of the actual TIMSS average scores and the percentages of students reaching each TIMSS international benchmark (hereafter referred to as "benchmark percentages") include sampling and measurement components:

$$Var = Var_{sampling} + Var_{measurement} \tag{4.3}$$

From projection linking, TIMSS state results were predicted for all 52 states/jurisdictions that participated in the NAEP mathematics and science assessments. The error variance associated with the predicted TIMSS results can be expressed as the following:

$$Var = Var_{sampling} + Var_{measurement} + Var_{adjustment} \tag{4.4}$$

The sampling error accounted for the uncertainty in estimating population statistics from a sample of the population. The second variance component, measurement error, was computed from the variance between predicted TIMSS plausible values, which accounted for the uncertainty in proficiency estimation (Johnson and Rust 1992). The third variance component was associated with the adjustment described in step 6 of the projection linking procedure, in which the predicted TIMSS plausible values for the 2011 NAEP U.S. national sample were adjusted to have the same mean and standard deviation as the 2011 TIMSS U.S. national sample. A jackknife procedure was employed to estimate the sampling and adjustment errors.

The linking study can be thought of as an estimation and prediction question in which the state-level TIMSS results can be predicted from the linking function. The variance estimated in equation 4.4 captures the uncertainty of the NAEP and TIMSS results, the uncertainty of the projection function, and the adjustment function. However, there is also uncertainty associated with predicting a new score point, i.e., the state TIMSS estimate, based on the linking function, which is referred to as *prediction error variance*.

How to estimate prediction error variance from the linking results is a challenging question. In the current study, where the actual state TIMSS results were available for the nine validation states, the *prediction residual error* for a state was defined as the difference between the predicted state

TIMSS result ($\hat{t}_i$) and the actual TIMSS result (t_i), then the Predicted Residual Sum of Squares, or PRESS, across the nine validation states is calculated as:

$$PRESS = \sum_{i=1}^{9} \left(\hat{t}_i - t_i \right)^2 \tag{4.5}$$

and the Mean Squared Error (MSE) is used as a summary measure of the prediction results:

$$MSE = \frac{\sum_{i=1}^{9} \left(\hat{t}_i - t_i \right)^2}{9} \tag{4.6}$$

The MSE measure reflects bias as well as variability. In particular, the bias portion of the predicted results can be expected to be considerable due to the many differences in NAEP and TIMSS administration policies and procedures. Consequently, using the MSE measure as an estimate of the prediction error variance in score comparisons (e.g., *t*-tests or Z-tests) would 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. Thus, the projection linking results were reported with the three-part error variance as shown in equation 4.4, without taking into account the prediction error component.

Results of Projection Linking

Recall that nine states participated in the 2011 TIMSS assessment at the state level, meaning that they have actual operational TIMSS assessment results. For the linking study, those states served as validation states, wherein their actual TIMSS scores were used to evaluate the accuracy of their predicted scores. Chapter 6 provides a detailed evaluation of the results from all three linking approaches employed in the study—calibration, projection, and moderation linking. In this section, results are provided for the predicted state TIMSS average scores from projection linking, before and after the linear adjustment described in step 6 of the projection linking procedure.

Table 4.4 shows the actual TIMSS state average scores and ranking of the nine validation states in mathematics. Also provided are the state rankings based on the predicted state TIMSS average scores from projection linking, the *prediction residual errors*, and MSE values, before and after the linear adjustment (step 6 of the projection linking procedure). For reference, the rankings of the nine states based on their actual 2011 NAEP mathematics scores are listed as well. It can be seen that the *prediction residual errors* changed in value when applying the linear adjustment to the predicted TIMSS scores; the MSE changed from 237 before the adjustment to 204 after the adjustment.

The predicted state TIMSS average mathematics scores that were obtained using the projection function from the TIMSS window braided-booklet sample are shown in table 4.5. In this case, the

linear adjustment also noticeably improved prediction precision, with the MSE changed in value from 263 to 195. A comparison between tables 4.4 and 4.5 indicates that, in general, the adjustment made a larger impact on the projected scores using the projection function derived from the TIMSS window braided-booklet sample than the one derived from the NAEP window braided-booklet sample.

The predicted state TIMSS science results are listed separately by administration window in tables 4.6 and 4.7. Similar to what was observed in the predicted TIMSS mathematics scores, the linear adjustment made less difference for the predicted TIMSS science results when using the projection function from the NAEP window braided-booklet sample. The MSE moderately changed in value after the adjustment (141 before the adjustment and 133 after the adjustment). On the other hand, the adjustment made an appreciable difference in improving prediction precision when the projection function was derived from the TIMSS window braided-booklet sample. The MSE changed in value from 786 before the adjustment to 124 after the adjustment.

When comparing the predicted state TIMSS average scores to their actual values for the nine validation states, it was observed from the *prediction residual error* ("Residual") column in tables 4.4 to 4.7 that there were discrepancies between the predicted and actual state results, regardless of which braided-booklet sample was used to derive the projection function.

For the nine validation states, the state-level predicted TIMSS average scores from projection linking and the predicted percentages of students reaching each of the TIMSS international benchmarks are listed in table 4.10 for mathematics and table 4.11 for science. These tables contain the predicted results that were based on the projection function derived from the NAEP window braided-booklet sample and reflect the linear adjustment.

The relationship between the NAEP and TIMSS mathematics latent scales was estimated using the mathematics braided-booklet sample from the NAEP administration window. Similarly, the NAEP window science braided-booklet sample was used to estimate the relationship between the NAEP and TIMSS science latent scales. The NAEP and TIMSS scales of the same content domain were fairly highly correlated. As shown in table 4.8, the estimated Pearson correlations were .92 for NAEP and TIMSS mathematics, and .90 for NAEP and TIMSS science.

Further Investigation of Multigroup Projection Linking

As discussed earlier in this chapter, the current study design offers the advantage of having students who receive the braided-booklet sample take items from both NAEP and TIMSS at the same time and under the same administration conditions. Because students did not take NAEP and TIMSS in separate test administrations, the design precludes the possible impact of a prior low-stakes

assessment on students' motivation on a second low-stakes assessment given later in time. Also, a linear adjustment was applied to help account for the difference in time of the year for NAEP and TIMSS testing. However, for the nine validation states, there were still sizeable discrepancies found when comparing the adjusted projection-based state TIMSS average scores to their actual TIMSS average scores. As shown in Chapter 6, the magnitudes of the discrepancies were comparable to the values observed from moderation linking (without the two-stage adjustment) and from calibration linking.

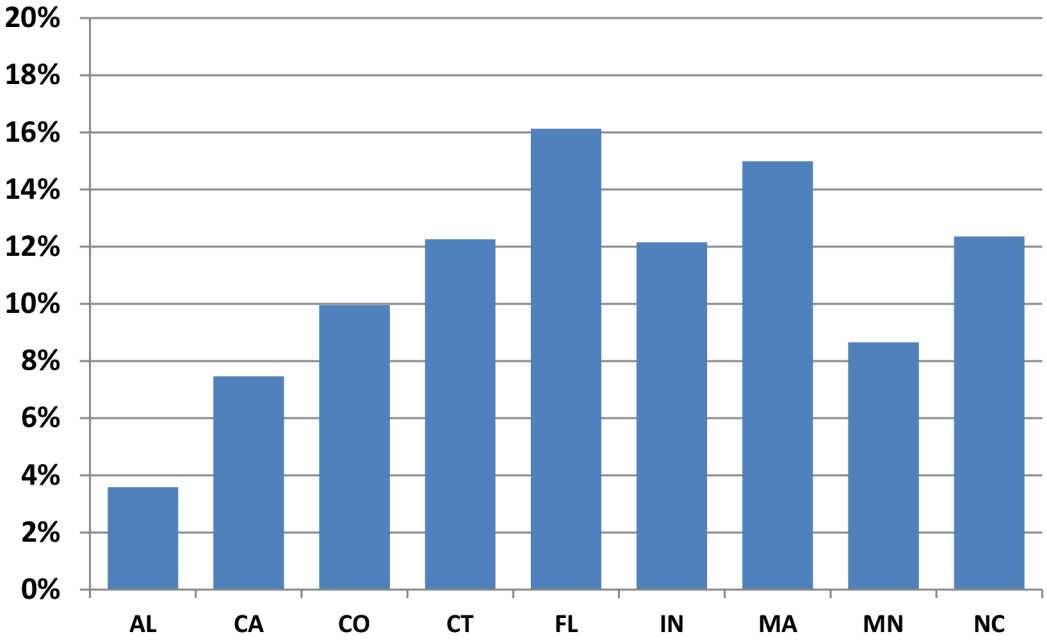
Consequently, additional steps were taken to further improve the prediction accuracy of the projection-based linkage. As discussed in Mislevy (1992), for projection linking the relationship between assessments can differ systematically for students with different backgrounds and characteristics. Thus, to properly support intended inferences, it is desirable to build a projection function that models relationships among not only assessments, but also "other student variables that will be involved in the inference" (Mislevy 1992, p. 54).

In this study, the overall joint NAEP-TIMSS distributions that were estimated for the braided-booklet samples took into consideration a large number of student demographic and background variables that NAEP and TIMSS routinely collect. In addition, the overall joint distributions were used to derive the projection function. In this section, the possibility of deriving student group-specific projection functions from the overall joint NAEP-TIMSS distribution is further explored. The variables of inference in this linking study were individual states. Thus, ideally, one would want to estimate and apply state-specific projection functions. However, the braided-booklet sample was a nationally-representative sample that did not contain representative state samples to support the estimation of state-specific projection functions. Alternatively, grouping variables were sought that (1) showed variability across states and (2) could potentially explain the differences between predicted and actual state TIMSS scores, then derive projection functions for each group separately.

The mathematics braided-booklet sample from the NAEP administration window was used as an example to demonstrate the procedures in deriving group specific projection functions. For the nine validation states, when using the *prediction residual error* computed in the results section as a measure of discrepancy between predicted and actual state TIMSS average scores, two student-level grouping variables were identified, where the group membership correlated fairly highly with the state-level *prediction residual error*. One grouping variable was *student accommodation status*, a binary variable with yes/no responses. For the validation states, the percentage of accommodated students per state was negatively correlated with the *prediction residual error* for state TIMSS mathematics average scores, with a Pearson correlation of -.76. Figure 4.1 shows a plot of the percentage of accommodated students by state for the 2011 NAEP mathematics assessment.

The second grouping variable was *student mathematics course-taking*, a categorical variable that measures students' opportunity to learn. The original categories of the second variable were collapsed into three groups so that the sample size within each group is reasonably large—basic or general eighth-grade mathematics, introduction to algebra or pre-algebra, and algebra and above (including geometry, algebra II, algebra I (1-year course), 1st year of 2-year algebra I, 2nd year of 2-year algebra I, integrated or sequential mathematics, other mathematics class). The percentages of student mathematics course taking per state were also correlated with state-level *prediction residual error*. For example, the Pearson correlation between the percentages of students taking basic or general eighth-grade mathematics per state and the *prediction residual error* was -.52.

Figure 4.1. Percentage of eighth-grade public school students identified as students with disabilities and/or English language learners assessed in NAEP mathematics with accommodations, as a percentage of all students, by validation state: 2011



To obtain projected TIMSS scores by student accommodation status, the NAEP and TIMSS plausible values from the NAEP window braided-booklet sample were used to conduct the following two-step procedure:

1. Obtain two separately-smoothed joint NAEP-TIMSS distributions from the NAEP window mathematics braided-booklet sample: one for the students tested with accommodations and one for the students tested without accommodations.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

2. Use the conditional distributions derived from the two joint distributions to predict distributions of TIMSS scores for students in the 2011 NAEP sample separately by their accommodation status.

The predicted TIMSS plausible values were then used to estimate the state-level TIMSS results. The same procedure was used to model and apply separate projection functions for the mathematics course-taking groupings. Table 4.9 contains the actual state TIMSS average mathematics scores, the *prediction residual errors* from table 4.4, and the *prediction residual errors* from the group-specific projection functions. For comparison purposes, all the *prediction residual errors* were based on results from projection linking without the linear adjustment. Rankings of states according to their actual and predicted TIMSS scores and their NAEP scores are also provided.

For the two grouping variables examined in this investigation, it can be seen clearly that the group-specific projection functions resulted in predicted state TIMSS results that were similar to those obtained from the projection functions derived from the overall NAEP-TIMSS joint distribution.

Tables

Table 4.1. NAEP coefficients of linear transformations of the univariate scale from the calibrating scale units to the units of the reporting scale at grade 8, by subject: 2011

Subject	$\hat{A}$	$\hat{B}$
Mathematics	36.737	283.284
Science	36.882	150.018

Table 4.2. TIMSS coefficients of linear transformations of the univariate scale from the calibrating scale units to the units of the reporting scale at grade 8, by subject: 2011

Overall mathematics	$\hat{A}$	$\hat{B}$
PV1	111.734	477.077
PV2	112.921	477.205
PV3	113.235	477.166
PV4	113.357	476.782
PV5	113.052	477.443
Overall science	$\hat{A}$	$\hat{B}$
PV1	109.112	486.672
PV2	108.793	486.531
PV3	107.813	487.546
PV4	109.266	486.444
PV5	108.546	487.171

Table 4.3. Projection linking linear adjustment parameter estimates at grade 8, by subject: 2011

Projection with NAEP window braided-booklet sample	$\hat{A}$	$\hat{B}$
Mathematics	.937	34.336
Science	.984	9.298
Projection with TIMSS window braided-booklet sample	$\hat{A}$	$\hat{B}$
Mathematics	.906	51.929
Science	.917	62.789

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 4.4. Actual TIMSS average scores, prediction residual errors, rankings, and mean squared errors in eighth-grade mathematics, by validation state, projection function derived from the NAEP window braided-booklet sample: 2011

Validation state	Actual TIMSS mathematics results			Predicted TIMSS mathematics results based on NAEP window braided-booklet sample								Rank in 2011 NAEP mathematics
				Projection before adjustment				Projection after adjustment				
	Rank	Average score	Standard error	Rank	Average score	Standard error	Residual	Rank	Average score	Standard error	Residual	
Massachusetts	1	561	5.3	1	**537**	**2.1**	**-23**	1	**538**	**3.3**	**-22**	1
Minnesota	2	545	4.6	2	**531**	**2.2**	**-14**	2	**532**	**3.4**	**-13**	2
North Carolina	3	537	6.8	5	**512**	**2.3**	**-25**	5	**514**	**3.4**	**-23**	5
Indiana	4	522	5.1	6	**510**	**1.9**	**-12**	6	512	3.2	-9	6
Colorado	5	518	4.9	3	524	2.4	6	3	525	3.5	8	3
Connecticut	6	518	4.8	4	514	2.7	-4	4	516	3.7	-2	4
Florida	7	513	6.4	7	**493**	**2.1**	**-20**	7	**497**	**3.2**	**-17**	7
California	8	493	4.9	8	483	2.6	-9	8	487	3.4	-5	8
Alabama	9	466	5.9	9	475	2.9	9	9	480	3.7	14	9
						Mean squared error	237				204	

NOTE: Residual = Predicted state TIMSS mean minus actual state TIMSS mean. Bold font indicates predicted average scores are statistically significantly different from the actual average scores. Two-tailed *t*-test, with alpha = .05, no adjustment for multiple comparisons. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 4.5. Actual TIMSS average scores, prediction residual errors, rankings, and mean squared errors in eighth-grade mathematics, by validation state, projection function derived from the TIMSS window braided-booklet sample: 2011

Validation state	Actual TIMSS mathematics results			Predicted TIMSS mathematics results based on TIMSS window braided-booklet sample								Rank in 2011 NAEP mathematics
				Projection before adjustment				Projection after adjustment				
	Rank	Average score	Standard error	Rank	Average score	Standard error	Residual	Rank	Average score	Standard error	Residual	
Massachusetts	1	561	5.3	1	**537**	**2.2**	**-24**	1	**538**	**3.3**	**-22**	1
Minnesota	2	545	4.6	2	**530**	**2.3**	**-15**	2	**532**	**3.4**	**-13**	2
North Carolina	3	537	6.8	5	**511**	**2.5**	**-26**	5	**515**	**3.5**	**-22**	5
Indiana	4	522	5.1	6	**509**	**2.3**	**-13**	6	513	3.4	-9	6
Colorado	5	518	4.9	3	523	2.4	5	3	525	3.5	8	3
Connecticut	6	518	4.8	4	512	2.6	-5	4	516	3.6	-2	4
Florida	7	513	6.4	7	**491**	**2.2**	**-22**	7	**497**	**3.2**	**-17**	7
California	8	493	4.9	8	**480**	**2.8**	**-13**	8	486	3.5	-6	8
Alabama	9	466	5.9	9	471	3.2	5	9	479	3.8	13	9
						Mean squared error	263				195	

NOTE: Residual = Predicted state TIMSS mean minus actual state TIMSS mean. Bold font indicates predicted average scores are statistically significantly different from the actual average scores. Two-tailed *t*-test, with alpha = .05, no adjustment for multiple comparisons. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 4.6. Actual TIMSS average scores, prediction residual errors, rankings, and mean squared errors in eighth-grade science, by validation state, projection function derived from the NAEP window braided-booklet sample: 2011

Validation state	Actual TIMSS science results			Predicted TIMSS science results based on NAEP window braided-booklet sample								Rank in 2011 NAEP science
				Projection before adjustment				Projection after adjustment				
	Rank	Average score	Standard error	Rank	Average score	Standard error	Residual	Rank	Average score	Standard error	Residual	
Massachusetts	1	567	5.1	1	**544**	**2.6**	**-22**	1	**545**	**3.4**	**-22**	1
Minnesota	2	553	4.6	2	543	2.6	-10	2	544	3.4	-9	2
Colorado	3	542	4.4	3	543	2.9	1	3	544	3.7	2	3
Indiana	4	533	4.8	5	526	2.2	-7	5	527	3.2	-6	5
Connecticut	5	532	4.6	4	530	2.6	-1	4	531	3.5	0	4
North Carolina	6	532	6.3	7	**515**	**2.4**	**-16**	7	**516**	**3.4**	**-15**	7
Florida	7	530	7.3	6	517	2.7	-13	6	518	3.5	-12	6
California	8	499	4.6	9	499	3.1	0	9	500	3.7	1	8
Alabama	9	485	6.2	8	499	3.0	13	8	**500**	**3.8**	**15**	9
						Mean squared error	141				133	

NOTE: Residual = Predicted state TIMSS mean minus actual state TIMSS mean. Bold font indicates predicted average scores are statistically significantly different from the actual average scores. Two-tailed *t*-test, with alpha = .05, no adjustment for multiple comparisons. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 4.7. Actual TIMSS average scores, prediction residual errors, rankings, and mean squared errors in eighth-grade science, by validation state, projection function derived from the TIMSS window braided-booklet sample: 2011

Validation state	Actual TIMSS science results			Predicted TIMSS science results based on TIMSS window braided-booklet sample								Rank in 2011 NAEP science
				Projection before adjustment				Projection after adjustment				
	Rank	Average score	Standard error	Rank	Average score	Standard error	Residual	Rank	Average score	Standard error	Residual	
Massachusetts	1	567	5.1	1	**526**	**2.8**	**-40**	1	**546**	**3.5**	**-21**	1
Minnesota	2	553	4.6	2	**526**	**2.5**	**-28**	2	545	3.2	-8	2
Colorado	3	542	4.4	3	**525**	**3.1**	**-17**	3	544	3.7	2	3
Indiana	4	533	4.8	5	**507**	**2.4**	**-26**	5	528	3.2	-5	5
Connecticut	5	532	4.6	4	**511**	**2.9**	**-21**	4	531	3.5	0	4
North Carolina	6	532	6.3	7	**494**	**2.9**	**-38**	7	**516**	**3.6**	**-16**	7
Florida	7	530	7.3	6	**495**	**2.9**	**-34**	6	517	3.5	-13	6
California	8	499	4.6	8	**475**	**3.5**	**-23**	8	499	3.8	0	8
Alabama	9	485	6.2	9	474	3.6	-11	9	498	4.0	13	9
						Mean squared error	786				124	

NOTE: Residual = Predicted state TIMSS mean minus actual state TIMSS mean. Bold font indicates predicted average scores are statistically significantly different from the actual average scores. Two-tailed *t*-test, with alpha = .05, no adjustment for multiple comparisons. Detail may not sum to totals because of rounding.

Table 4.8. Marginal correlation between NAEP and TIMSS reporting scales at grade 8 for braided-booklet samples in 2011 NAEP administration window, by subject area scale: 2011

Assessment	NAEP mathematics
NAEP mathematics	1.00
TIMSS mathematics	.92
Assessment	**NAEP science**
NAEP science	1.00
TIMSS science	.90

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 4.9. Actual TIMSS average scores, prediction residual errors, rankings, and mean squared errors in eighth-grade mathematics, by validation state, projection function derived from the overall NAEP window braided-booklet sample, and by subgroup: 2011

Validation state	Actual TIMSS mathematics results			Predicted TIMSS mathematics results based on NAEP window braided-booklet sample						Rank in 2011 NAEP mathematics
				Projection before adjustment		Projection by accommodation status		Projection by mathematics course taking		
	Rank	Average score	Standard error	Rank	Residual	Rank	Residual	Rank	Residual	
Massachusetts	1	561	5.3	1	-23	1	-24	1	-23	1
Minnesota	2	545	4.6	2	-14	2	-14	2	-13	2
North Carolina	3	537	6.8	5	-25	5	-25	5	-25	5
Indiana	4	522	5.1	6	-12	6	-12	6	-12	6
Colorado	5	518	4.9	3	6	3	6	3	6	3
Connecticut	6	518	4.8	4	-4	4	-4	4	-3	4
Florida	7	513	6.4	7	-20	7	-20	7	-19	7
California	8	493	4.9	8	-9	8	-8	8	-7	8
Alabama	9	466	5.9	9	9	9	10	9	10	9
Mean squared error					237		238		225	

NOTE: Residual = Predicted state TIMSS mean minus actual state TIMSS mean. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 4.10. Predicted state TIMSS average scores, predicted benchmark results, and standard errors from projection linking in eighth-grade mathematics, by validation state: 2011

	Predicted TIMSS mathematics results from projection linking									
	Average score		Percentage of students reaching TIMSS low international benchmark		Percentage of students reaching TIMSS intermediate international benchmark		Percentage of students reaching TIMSS high international benchmark		Percentage of students reaching TIMSS advanced international benchmark	
Validation state	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error
Alabama	480	3.7	85	1.4	54	2.1	18	1.5	2	0.5
California	487	3.4	85	1.2	57	1.9	23	2	4	0.7
Colorado	525	3.5	95	0.7	76	1.5	39	2.4	8	1.5
Connecticut	516	3.7	93	0.8	71	1.8	34	1.7	7	1.1
Florida	497	3.2	89	1.1	62	1.7	25	1.4	4	0.6
Indiana	512	3.2	94	0.9	71	1.7	31	1.6	5	1
Massachusetts	538	3.3	96	0.7	81	1.7	46	2	11	1.4
Minnesota	532	3.4	95	0.7	79	1.6	42	1.7	9	1.3
North Carolina	514	3.4	93	1.1	70	2	33	1.8	7	1

NOTE: The predicted TIMSS state mathematics score was calculated from the projection linking using the NAEP window braided-booklet sample.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 4.11. Predicted state TIMSS average scores, predicted benchmark results, and standard errors from projection linking in eighth-grade science, by validation state: 2011

Validation state	Predicted TIMSS science results from projection linking									
	Average score		Percentage of students reaching TIMSS low international benchmark		Percentage of students reaching TIMSS intermediate international benchmark		Percentage of students reaching TIMSS high international benchmark		Percentage of students reaching TIMSS advanced international benchmark	
	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error
Alabama	500	3.8	88	1.3	65	2	29	1.9	5	0.9
California	500	3.7	87	1.5	64	2	30	1.7	7	1
Colorado	544	3.7	96	0.7	82	1.4	49	1.8	14	1.8
Connecticut	531	3.5	94	0.8	77	1.7	43	2.1	11	1.2
Florida	518	3.5	92	1.2	72	1.8	37	2.2	8	0.8
Indiana	527	3.2	94	1.1	76	1.8	41	1.7	9	0.9
Massachusetts	545	3.4	95	0.8	82	1.3	51	1.9	15	1.6
Minnesota	544	3.4	96	0.7	83	1.2	50	2	13	1.3
North Carolina	516	3.4	92	1.1	72	2.2	35	1.8	8	1

NOTE: The predicted TIMSS state science score was calculated from the projection linking using the NAEP window braided-booklet sample.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

References

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

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., and 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., and Galia, J. (2012). Scaling the TIMSS and PIRLS 2011 Achievement Data. In M.O. Martin, and 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.

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). New York, NY: Springer, LLC.

Johnson, E.G., and Rust, K.F. (1992). Population Inferences and Variance Estimation for NAEP Data. *Journal of Educational Statistics*, *17*, 175–190.

Johnson, E.G., Cohen, J., Chen, W.-H., Jiang, T., and Zhang, Y. (2003). *2000 NAEP–1999 TIMSS Linking Report*. (NCES 2005-01). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

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

Mislevy, R.J. (1984). Estimating Latent Distributions. *Psychometrika*, *49*, 359-381.

Mislevy, R.J. (1985). Estimation of Latent Group Effects. *Journal of the American Statistical Association*, *80*, 993-997.

Mislevy, R.J. (1991). Randomization-Based Inference About Latent Variables From Complex Samples. *Psychometrika*, *56(2)*, 177-196.

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., and Sheehan, K.M. (1992). Estimating Population Characteristics from Sparse Matrix Samples of Item Responses. *Journal of Educational Measurement*, *29*, 133-161.

Mislevy, R.J., Johnson, E., and 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.

Muraki, E. and Bock, R.D. (1991). *PARSCALE*: *Parameter Scaling of Rating Data [computer program]*. Chicago, IL: Scientific Software, Inc.

Pashley, P.J. and Phillips, G.W. (1993). *Toward World-Class Standards: A Research Study Linking International and National Assessments*. Princeton, NJ: Educational Testing Service.

Rogers, A., Gregory, K., Davis, S., and Kulick, E. (2006a). *User's Guide to NAEP Model-Based P-value Programs*. Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Rogers, A., Tang, C., Lin, M.-J., and Kandathil, M. (2006b). *DGROUP [Computer Software]*. Princeton, NJ: Educational Testing Service.

Thomas, N. (1994). *CGROUP and BGROUP: Modifications of the MGROUP Program to Estimate Group Effects in Multivariate Models [Computer Programs]*. Princeton, NJ: Educational Testing Service.

von Davier, M., Gonzalez, E., and 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. Princeton, NJ: IEA-ETS Research Institute. Retrieved August 19, 2013, from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf.

von Davier, M., Sinharay, S., Oranje, A., and Beaton, A. (2007). *Statistical Procedures Used in the National Assessment of Educational Progress (NAEP): Recent Developments and Future Directions*. In C. R. Rao and S. Sinharay (Eds.), *Handbook of Statistics: Vol. 26*.

Chapter 5: Linking 2011 NAEP to 2011 TIMSS Using Statistical Moderation

Overview

This chapter describes the statistical moderation linking analysis conducted by AIR to link the 2011 NAEP to the 2011 TIMSS. In statistical moderation, a technique also applied by Johnson et al. (2003) in an earlier attempt to link NAEP with TIMSS, the estimated scores were 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 linking was conducted using the grade 8 U.S. national NAEP and TIMSS samples and validated with samples of nine states—Alabama, California, Colorado, Connecticut, Florida, Indiana, Massachusetts, Minnesota, and North Carolina—that participated in both 2011 NAEP and TIMSS. After the statistical link was established between NAEP and TIMSS, the link was applied to the remaining states in the study in order to estimate TIMSS performance.

Method

The 2011 NAEP-TIMSS link using statistical moderation was accomplished in five steps. It should be noted that, steps 1 and 2 correspond to the first stage of 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. The TIMSS-equivalent, $\overline{z}_{1j}$ associated with a NAEP state mean $\overline{x}_j$ is

$$\overline{z}_{1j} = \left(\overline{y} - \frac{\hat{\sigma}_y}{\hat{\sigma}_x}\overline{x} \right) + \left(\frac{\hat{\sigma}_y}{\hat{\sigma}_x} \right) \overline{x}_j \tag{5.1}$$

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

$$\hat{B} = \frac{\hat{\sigma}_y}{\hat{\sigma}_x} \tag{5.2}$$

In equations 5.1 and 5.2,

- $\hat{A}$ is an estimate of the intercept of a straight line, and $\hat{B}$ is an estimate of the slope,
- $\overline{x}$ and $\overline{y}$ are the national public school means of the U.S. NAEP and U.S. TIMSS results,
- $\hat{\sigma}_x$ and $\hat{\sigma}_{\overline{y}}$ are the public school standard deviations for NAEP and TIMSS respectively,
- $\overline{z}_{1j}$ is the TIMSS-equivalent of the NAEP mean $\overline{x}_j$ in state *j*.

The error variances in the mean TIMSS-equivalents are

$$\hat{\sigma}^2_{\overline{z}_{1j}} = \hat{B}^2\hat{\sigma}^2_{\overline{x}_j} + \hat{\sigma}^2_A + 2\left(\overline{x}_j\right)\hat{\sigma}_{AB} + \left(\overline{x}_j\right)^2\hat{\sigma}^2_B \tag{5.3}$$

The square root of equation 5.3 is the standard error of linking. According to Johnson et al. (2003), 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).

$$\hat{\sigma}^2_A = \hat{B}^2\hat{\sigma}^2_{\overline{x}} + \hat{\sigma}^2_{\overline{y}} + \overline{x}^2\hat{B}^2\left[\frac{\hat{\sigma}^2_{\sigma_y}}{\hat{\sigma}^2_y} + \frac{\hat{\sigma}^2_{\sigma_x}}{\hat{\sigma}^2_x}\right]$$

$$\hat{\sigma}_{AB} = -\overline{x}\hat{B}^2\left[\frac{\hat{\sigma}^2_{\sigma_y}}{\hat{\sigma}^2_y} + \frac{\hat{\sigma}^2_{\sigma_x}}{\hat{\sigma}^2_x}\right] \tag{5.4}$$

$$\hat{\sigma}^2_B = \hat{B}^2\left[\frac{\hat{\sigma}^2_{\sigma_y}}{\hat{\sigma}^2_y} + \frac{\hat{\sigma}^2_{\sigma_x}}{\hat{\sigma}^2_x}\right]$$

Estimates of the Means and Standard Deviations

The process began 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 drawn from 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 $\overline{p} = \sum_{n=1}^{N}\frac{p_n}{N}$. Table 5.1 shows the calculations for the parameter estimates of the means and standard deviations.

Error variance (sampling) of the mean and standard deviation

The error variances for the parameter estimates in table 5.1 each have two components: error variance due to sampling (S) and error variance due to measurement (M). The sampling error in the estimates of the means and standard deviations were obtained using a jackknife error variance approach for complex samples. More information on the jackknife procedure can be found at http://nces.ed.gov/nationsreportcard/tdw/weighting/2000_2001/2000main_varestimate_sampvar_jack.aspx. 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 as follows:

$$S = \sum_{n=1}^{N}\frac{S_n}{N}\ .$$

This process results in the variance estimates reported in table 5.2, which are estimates of error variance due to sampling for the mean and standard deviations.

Error variance (measurement) of the mean and standard deviation

The error variance due to measurement is estimated by the variance between plausible values. This is estimated by $M = \frac{1+(1/N)}{N-1}\sum_{n=1}^{N}(p_n - \bar{p})^2$. The error variance due to measurement is shown in table 5.3.

Error variance (total) of the mean and standard deviation

The total error variance is $T = S + M$ and is shown in table 5.4.

Estimates of the linking parameters A and B

The linking parameters were calculated for each plausible value using equation 5.2. The linking parameter estimates were then averaged over the five plausible values as reported in table 5.5.

Error variance (sampling) of the linking parameters A and B are outlined in table 5.6.

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 5.7.

Error variance (total) of the linking parameters A and B is shown in table 5.8.

The TIMSS-equivalents of the nine validation state NAEP means are contained in tables 5.9 and 5.10.

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

An investigation of the relationships between state-level accommodation rates and mean scores was conducted, and it was recommended that the state TIMSS-equivalent means be adjusted to account for differences in the accommodation rates among states which predict differences between NAEP and TIMSS exclusion rates. The derivations of specific adjustments are described in Chapter 6 of this report, *Adjustments to Predicted State Mean Estimates*. The following adjustments were used following the HumRRO recommendations.

- For mathematics: $\hat{T}_{adj}(j) = \hat{T}(j) + (2.65(\%\ Acc_j - 9.7))$
  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: $\hat{T}_{adj}(j) = \hat{T}(j) + (2.21(\%\ Acc_j - 10.6))$
  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 state accommodation rates are estimated in tables 5.11 and 5.12. The TIMSS-equivalents of the nine validation state NAEP means with adjustments for accommodations are contained in tables 5.13 and 5.14.

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 a determination of 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 was to estimate state performance on TIMSS based on NAEP performance in the 43 states/jurisdictions in which TIMSS was not administered. This can be addressed by taking advantage of the correlation between NAEP and TIMSS (or equivalently the correlation between the TIMSS-equivalents and the actual TIMSS) estimated from the nine validation states. The prediction of state TIMSS from state TIMSS-equivalents can then be accomplished through statistical prediction.

$$\overline{z}_{2j} = \left(\overline{\overline{y}} - \hat{\rho} \frac{\hat{\sigma}_{\bar{y}}}{\hat{\sigma}_{\overline{z}_1}} \overline{\overline{z}}_1 \right) + \left(\hat{\rho} \frac{\hat{\sigma}_{\bar{y}}}{\hat{\sigma}_{\overline{z}_1}} \right) \overline{z}_{1j} \tag{5.5}$$

With intercept and slope regression parameters

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

$$\hat{\beta} = \hat{\rho} \frac{\hat{\sigma}_{\bar{y}}}{\hat{\sigma}_{\overline{z}_1}} \tag{5.6}$$

The quantities in equation 5.5 are defined as follows:

- $\overline{z}_{2j}$ is the *projected* state TIMSS mean for a given TIMSS-equivalent $\overline{z}_{1j}$,
- $\overline{\overline{z}}_1$ 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,

- $\hat{\sigma}_{\bar{z}_1}$ is the weighted standard deviation of the adjusted state means of TIMSS-equivalents in the nine validation states,
- $\bar{\bar{y}}$ is the weighted mean of the actual TIMSS means among the nine validation states,
- $\hat{\sigma}_{\bar{y}}$ 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 $\bar{z}_{1j}$ and actual TIMSS state means $\bar{y}_j$ in the nine validation states.

The error variance in the projection is found by

$$\hat{\sigma}^2_{\bar{z}_{2j}} = \hat{\beta}^2 \hat{\sigma}^2_{\bar{z}_{1j}} + \hat{\sigma}^2_{\hat{\alpha}} + 2\left(\bar{z}_{1j}\right)\hat{\sigma}_{\hat{\alpha},\hat{\beta}} + \left(\bar{z}_{1j}\right)^2 \hat{\sigma}^2_{\hat{\beta}}. \tag{5.7}$$

In equation 5.7 the projection error variance components are as follows:

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

The variances and co-variances of α and β in equation 5.7 are

$$\begin{aligned}\hat{\sigma}^2_{\hat{\alpha}} \approx\ & \hat{\beta}^2 Var(\bar{\bar{z}}_1) + \hat{\beta}^2 \frac{\bar{\bar{z}}_1^2}{\hat{\sigma}^2_{\bar{z}_1}} Var(\hat{\sigma}_{\bar{z}_1}) + Var(\bar{\bar{y}}) + \hat{\beta}^2 \frac{\bar{\bar{z}}_1^2}{\hat{\sigma}^2_{\bar{y}}} Var(\hat{\sigma}_{\bar{y}}) + \hat{\beta}^2 \frac{\bar{\bar{z}}_1^2}{\hat{\rho}^2} Var(\hat{\rho}) \\ & -2\hat{\beta} Cov(\bar{\bar{z}}_1, \bar{\bar{y}}) - 2\hat{\beta}^2 \frac{\bar{\bar{z}}_1^2}{\hat{\sigma}_{\bar{z}_1}\hat{\sigma}_{\bar{y}}} Cov(\hat{\sigma}_{\bar{z}_1}, \hat{\sigma}_{\bar{y}}) - 2\hat{\beta}^2 \frac{\bar{\bar{z}}_1^2}{\hat{\rho}\hat{\sigma}_{\bar{z}_1}} Cov(\hat{\sigma}_{\bar{z}_1}, \hat{\rho}) \\ & +2\hat{\beta}^2 \frac{\bar{\bar{z}}_1^2}{\hat{\rho}\hat{\sigma}_{\bar{y}}} Cov(\hat{\sigma}_{\bar{y}}, \hat{\rho})\end{aligned} \tag{5.8}$$

$$\begin{aligned}\hat{\sigma}^2_{\hat{\beta}} \approx\ & \hat{\beta}^2 \frac{1}{\hat{\sigma}^2_{\bar{z}_1}} Var(\hat{\sigma}_{\bar{z}_1}) + \hat{\beta}^2 \frac{1}{\hat{\sigma}^2_{\bar{y}}} Var(\hat{\sigma}_{\bar{y}}) + \hat{\beta}^2 \frac{1}{\hat{\rho}^2} Var(\hat{\rho}) \\ & -2\hat{\beta}^2 \frac{1}{\hat{\sigma}_{\bar{z}_1}\hat{\sigma}_{\bar{y}}} Cov(\hat{\sigma}_{\bar{z}_1}, \hat{\sigma}_{\bar{y}}) - 2\hat{\beta}^2 \frac{1}{\hat{\rho}\hat{\sigma}_{\bar{z}_1}} Cov(\hat{\sigma}_{\bar{z}_1}, \hat{\rho}) \\ & +2\hat{\beta}^2 \frac{1}{\hat{\rho}\hat{\sigma}_{\bar{y}}} Cov(\hat{\sigma}_{\bar{y}}, \hat{\rho})\end{aligned} \tag{5.9}$$

$$\hat{\sigma}_{\hat{\alpha},\hat{\beta}} \approx -\hat{\beta}^2 \frac{\overline{\overline{z}}_1}{\hat{\sigma}^2_{\bar{z}_1}} Var(\hat{\sigma}_{\bar{z}_1}) - \hat{\beta}^2 \frac{\overline{\overline{z}}_1}{\hat{\sigma}^2_{\bar{y}}} Var(\hat{\sigma}_{\bar{y}}) - \hat{\beta}^2 \frac{\overline{\overline{z}}_1}{\hat{\rho}^2} Var(\hat{\rho})$$
$$+2\hat{\beta}^2 \frac{\overline{\overline{z}}_1}{\hat{\sigma}_{\bar{z}_1}\hat{\sigma}_{\bar{y}}} Cov(\hat{\sigma}_{\bar{z}_1}, \hat{\sigma}_{\bar{y}}) + 2\hat{\beta}^2 \frac{\overline{\overline{z}}_1}{\hat{\rho}\hat{\sigma}_{\bar{z}_1}} Cov(\hat{\sigma}_{\bar{z}_1}, \hat{\rho}) \quad (5.10)$$
$$-2\hat{\beta}^2 \frac{\overline{\overline{z}}_1}{\hat{\rho}\hat{\sigma}_{\bar{y}}} Cov(\hat{\sigma}_{\bar{y}}, \hat{\rho})$$

The components of equations 5.8 to 5.10 can be estimated as follows:

$$Var(\overline{\overline{z}}_1) = \frac{\hat{\sigma}^2_{\bar{z}_1}}{n} \quad (5.11)$$

$$Var(\overline{\overline{y}}) = \frac{\hat{\sigma}^2_{\bar{y}}}{n} \quad (5.12)$$

$$Var(\hat{\sigma}_{\bar{z}_1}) = \frac{\hat{\sigma}^2_{\bar{z}_1}}{2(n-1)} \quad (5.13)$$

$$Var(\hat{\sigma}_{\bar{y}}) = \frac{\hat{\sigma}^2_{\bar{y}}}{2(n-1)} \quad (5.14)$$

$$Var\left(\hat{\sigma}^2_{\bar{z}_1}\right) \approx 4\hat{\sigma}^2_{\bar{z}_1} Var\left(\hat{\sigma}_{\bar{z}_1}\right) \quad (5.15)$$

$$Var\left(\hat{\sigma}^2_{\bar{y}}\right) \approx 4\hat{\sigma}^2_{\bar{y}} Var\left(\hat{\sigma}_{\bar{y}}\right) \quad (5.16)$$

$$Var\left(\hat{\rho}\right) \approx \left(1-\hat{\rho}^2\right)^2 \left\{ \frac{1}{n-1} + \frac{11\hat{\rho}^2}{2(n-1)^2} + \frac{-24\hat{\rho}^2 + 75\hat{\rho}^4}{16(n-1)^3} \right\} \quad (5.17)$$

$$Var\left(\hat{\rho}^2\right) = 4\hat{\rho}^2 Var\left(\hat{\rho}\right) \quad (5.18)$$

The quantities in equations 5.11 to 5.18 are defined as follows:

- $Var(\bar{\bar{z}}_1)$ in equation 5.11 is error variance in the weighted mean of the adjusted TIMSS-equivalent means among the nine validation states.
- $Var(\bar{\bar{y}})$ in equation 5.12 is the error variance in the weighted mean of the actual TIMSS means among the nine validation states.
- $Var(\hat{\sigma}_{\bar{z}_1})$ in equation 5.13 is the error variance in the weighted standard deviation of the adjusted state means of TIMSS-equivalents in the nine validation states.
- $Var(\hat{\sigma}_{\bar{y}})$ in equation 5.14 is the error variance in the weighted standard deviation of the state means of actual TIMSS among the nine validation states.
- $Var\left(\hat{\sigma}^2_{\bar{z}_1}\right)$ in equation 5.15 is the error variance in the weighted variance of the adjusted state means of TIMSS-equivalents in the nine validation states.
- $Var\left(\hat{\sigma}^2_{\bar{y}}\right)$ in equation 5.16 is the error variance in the weighted variance of the state means of actual TIMSS among the nine validation states.
- $Var\left(\hat{\rho}\right)$ in equation 5.17 is the error variance in the weighted correlation between the state's mean TIMSS-equivalents $\bar{z}_{1j}$ and actual TIMSS state means $\bar{y}_j$ in the nine validation states.
- $Var\left(\hat{\rho}^2\right)$ in equation 5.18 is the error variance in the weighted square of the correlation between the state's mean TIMSS-equivalents $\bar{z}_{1j}$ and actual TIMSS state means $\bar{y}_j$ in the nine validation states.

$$Cov(\bar{\bar{z}}_1, \bar{\bar{y}}) = \hat{\rho}\sqrt{Var(\bar{\bar{z}}_1)Var(\bar{\bar{y}})} \tag{5.19}$$

$$Cov\left(\hat{\sigma}_{\bar{z}_1}, \hat{\sigma}_{\bar{y}}\right) \approx \hat{\rho}^2\sqrt{Var\left(\hat{\sigma}_{\bar{z}_1}\right)Var\left(\hat{\sigma}_{\bar{y}}\right)} \tag{5.20}$$

$$Cov\left(\hat{\sigma}^2_{\bar{z}_1}, \hat{\sigma}^2_{\bar{y}}\right) \approx \hat{\rho}^2\sqrt{Var\left(\hat{\sigma}^2_{\bar{z}_1}\right)Var\left(\hat{\sigma}^2_{\bar{y}}\right)} \tag{5.21}$$

$$Cov\left(\hat{\rho}, \hat{\sigma}_{\bar{z}_1}\right) \approx \hat{\rho}_{\hat{\rho}^2, \hat{\sigma}^2_{\bar{z}_1}}\sqrt{Var\left(\hat{\rho}\right)Var\left(\hat{\sigma}_{\bar{z}_1}\right)} \tag{5.22}$$

$$Cov\left(\hat{\rho}^2, \hat{\sigma}^2_{\bar{z}_1}\right) \approx \hat{\rho}_{\hat{\rho}^2, \hat{\sigma}^2_{\bar{z}_1}}\sqrt{Var\left(\hat{\rho}^2\right)Var\left(\hat{\sigma}^2_{\bar{z}_1}\right)} \tag{5.23}$$

$$\hat{\rho}_{\hat{\rho}^2, \hat{\sigma}^2_{\bar{z}_1}} \approx \frac{\left(\hat{\rho}^2\hat{\sigma}^2_{\bar{z}_1} + \frac{\left(1-\hat{\rho}^2\right)}{n-1}\hat{\sigma}^2_{\bar{z}_1} - \left(Var\left(\hat{\rho}\right) + \hat{\rho}^2\right)\left(Var\left(\hat{\sigma}_{\bar{z}_1}\right) + \hat{\sigma}^2_{\bar{z}_1}\right)\right)}{\sqrt{Var\left(\hat{\sigma}^2_{\bar{z}_1}\right)Var\left(\hat{\rho}^2\right)}} \tag{5.24}$$

$$Cov\left(\hat{\rho}, \hat{\sigma}_{\bar{y}}\right) \approx \hat{\rho}_{\hat{\rho}^2, \hat{\sigma}^2_{\bar{y}}}\sqrt{Var\left(\hat{\rho}\right)Var\left(\hat{\sigma}_{\bar{y}}\right)} \tag{5.25}$$

$$Cov\left(\hat{\rho}, \hat{\sigma}^2_{\bar{y}}\right) \approx \hat{\rho}_{\hat{\rho}^2, \hat{\sigma}^2_{\bar{y}}}\sqrt{Var\left(\hat{\rho}\right)Var\left(\hat{\sigma}^2_{\bar{y}}\right)} \tag{5.26}$$

$$\hat{\rho}_{\hat{\rho}^2, \hat{\sigma}^2_{\bar{y}}} \approx \frac{\left(\hat{\rho}^2\hat{\sigma}^2_{\bar{y}} + \frac{\left(1-\hat{\rho}^2\right)}{n-1}\hat{\sigma}^2_{\bar{y}} - \left(Var\left(\hat{\rho}\right) + \hat{\rho}^2\right)\left(Var\left(\hat{\sigma}_{\bar{y}}\right) + \hat{\sigma}^2_{\bar{y}}\right)\right)}{\sqrt{Var\left(\hat{\sigma}^2_{\bar{y}}\right)Var\left(\hat{\rho}^2\right)}} \tag{5.27}$$

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 correlations between TIMSS-equivalent means and actual TIMSS means were improved by the adjustment for accommodations. Since the accommodation adjustments in both mathematics and science should improve the projections, it was decided to use the accommodation adjustments as part of the projections. The prediction is conducted among the nine validation states with the accommodation adjustments in tables 5.15 – 5.18.

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 5.1 by substituting z_{1j} for $\overline{z}_{1j}$ and x_j for $\overline{x}_j$. Once the distribution of z_{1j} is determined, an estimate of the proportion above various cut-scores on z_{1j} can be done. 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 estimated to be above the international benchmarks on TIMSS-equivalents in each state. The quantity $1-\hat{p}_{1j}$ can be estimated via a normal approximation.

$$\begin{aligned} 1-\hat{p}_{1j} &= \Pr\left(z_{1j} \geq z_{benchmark}\right) \\ &= \int_{-\infty}^{\infty} \Pr\left(z_{1j} \geq z_{benchmark} \mid x_j\right) f\left(x_j \mid \overline{x}_j, \hat{\sigma}^2_{X_j}\right) dx_j \\ &= \int_{z_{benchmark}}^{\infty} f\left(x_j \mid A + B\overline{x}_j, B^2\hat{\sigma}^2_{X_j}\right) dx_j \end{aligned} \tag{5.28}$$

This value can be defined as $h(z_{benchmark}, \overline{z}_{1j}, \hat{\sigma}_{Z_{1j}}) = \int_{z_{benchmark}}^{\infty} f\left(x_j \mid A + B\overline{x}_j, B^2\hat{\sigma}^2_{X_j}\right) dx_j$. 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

$$\begin{aligned} \sigma^2_{L(1-p_{1j})} &= Var(h(z_{benchmark}, \overline{z}_{1j}, \hat{\sigma}_{z_{1j}})) \\ &\approx \left(\frac{\exp\left(-(z_{benchmark}-\overline{z}_{1j})^2/(2\hat{\sigma}^2_{z_{1j}})\right)}{\sqrt{2\pi}\hat{\sigma}_{Z_{1j}}}\right)^2 Var\left(z_{1j}\right) \\ &+ \left(\frac{\exp\left(-(z_{benchmark}-\overline{z}_{1j})^2/(2\hat{\sigma}^2_{z_{1j}})\right)}{\sqrt{2\pi}\hat{\sigma}_{z_{1j}}}\right)^2 Var\left(\overline{z}_{1j}\right) \\ &+ \left(\left(\frac{z_{benchmark}-\overline{z}_{1j}}{\hat{\sigma}_{z_{1j}}}\right)\frac{\exp\left(-(z_{benchmark}-\overline{z}_{1j})^2/(2\hat{\sigma}^2_{z_{1j}})\right)}{\sqrt{2\pi}\hat{\sigma}_{z_{1j}}}\right)^2 Var\left(\hat{\sigma}_{z_{1j}}\right) \end{aligned} \tag{5.29}$$

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
- $Var\left(\hat{\sigma}_{Z_1 j}\right)$ is the error variance in the standard deviation of z_{1j}.

Step 5: Predicting the Percentages at and Above International Benchmarks

To predict the percentages at and above international benchmarks in the projected distribution $1 - p_{2j}$ equations 5.5, 5.6, and 5.7 were used 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 $\overline{\overline{y}}$
- the mean of $1 - p_{1j}$ is substituted for $\overline{\overline{z}_1}$

The parameter estimates that were needed to conduct the projections for both the means (step 3) and the international benchmarks (step 5) are contained in tables 5.19 – 5.34.

Tables

Table 5.1. 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

Table 5.2. 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

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.3. 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

Table 5.4. 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

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.5. 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 $(\overline{p})$
$\hat{A}$ (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	171.192	170.125	170.776
$\hat{B}$ (science)	2.351	2.325	2.313	2.325	2.342	2.331

Table 5.6. Sampling error variance in *A* and *B* linking parameters (S_A, S_B, S_{AB})

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

Table 5.7. Measurement error variance in *A* and *B* linking parameters (M_A, M_B, M_{AB})

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

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.8. 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
Co-variance 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
Co-variance 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

Table 5.9. TIMSS-equivalents of state means in mathematics

State	Equivalent state mean without accommodation adjustment	Error linking	TIMSS state mean	Error state TIMSS	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: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.10. TIMSS-equivalents of state means in science

State	Equivalent state mean without accommodation adjustment	Error linking	TIMSS state mean	Error state TIMSS	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: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.11. Accommodation rates in mathematics

State	Accommodation rate
Alabama	4
Alaska	14
Arizona	9
Arkansas	12
California	7
Colorado	10
Connecticut	12
Delaware	11
District of Columbia	15
DoDEA	8
Florida	16
Georgia	7
Hawaii	11
Idaho	7
Illinois	12
Indiana	12
Iowa	14
Kansas	9
Kentucky	8
Louisiana	13
Maine	14
Maryland	7
Massachusetts	15
Michigan	8
Minnesota	9
Mississippi	6
Missouri	10
Montana	9
U.S. National	10
National Private	5
National Public	10
Nebraska	9
Nevada	9
New Hampshire	14
New Jersey	14
New Mexico	10
New York	18
North Carolina	12
North Dakota	9
Ohio	10
Oklahoma	4
Oregon	11
Pennsylvania	13
Rhode Island	13
South Carolina	8
South Dakota	7
Tennessee	8
Texas	5
Utah	8
Vermont	15
Virginia	9
Washington	10
West Virginia	9
Wisconsin	14
Wyoming	11

Table 5.12. Accommodation rates in science

State	Accommodation rate
Alabama	4
Alaska	16
Arizona	9
Arkansas	12
California	8
Colorado	10
Connecticut	13
Delaware	12
District of Columbia	18
DoDEA	10
Florida	16
Georgia	8
Hawaii	11
Idaho	7
Illinois	12
Indiana	13
Iowa	14
Kansas	9
Kentucky	8
Louisiana	13
Maine	14
Maryland	11
Massachusetts	16
Michigan	8
Minnesota	8
Mississippi	6
Missouri	10
Montana	9
U.S. National	11
National Private	5
National Public	11
Nebraska	12
Nevada	11
New Hampshire	13
New Jersey	17
New Mexico	10
New York	18
North Carolina	12
North Dakota	10
Ohio	12
Oklahoma	10
Oregon	10
Pennsylvania	15
Rhode Island	14
South Carolina	9
South Dakota	8
Tennessee	10
Texas	8
Utah	9
Vermont	14
Virginia	10
Washington	10
West Virginia	9
Wisconsin	14
Wyoming	11

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.13. TIMSS-equivalents of state means with adjustments for accommodations in grade 8 mathematics

State	TIMSS-equivalent state mean without accommodation adjustment	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: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

Table 5.14. TIMSS-equivalents of state means with adjustments for accommodations in grade 8 science

State	TIMSS-equivalent state mean without accommodation adjustment	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: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.15. Projection parameters for mathematics means

	Parameter estimates	
Correlation	α	β
.94	32.1584	0.9457
Variance-covariance		
	α	β
α	15.1720	-0.0294
β	-0.0294	0.0001

Table 5.16. Projection for mathematics with accommodation adjustments

State	Projection	Standard error linking	Standard error prediction	Standard error projection	Actual TIMSS	Standard error TIMSS	Overall standard error	Z-Test	Significant difference
Alabama	469	4.0	0.4	3.8	466	5.9	7.0	0.46	NS
California	486	3.7	0.3	3.5	493	4.9	6.0	-1.06	NS
Colorado	530	3.5	0.2	3.4	518	4.9	5.9	2.07	Significant
Connecticut	526	3.5	0.2	3.3	518	4.8	5.9	1.51	NS
Florida	518	3.2	0.2	3.0	513	6.4	7.1	0.66	NS
Indiana	522	3.4	0.2	3.2	522	5.1	6.0	0.12	NS
Massachusetts	556	3.2	0.4	3.1	561	5.3	6.1	-0.72	NS
Minnesota	533	3.4	0.2	3.2	545	4.6	5.6	-2.05	Significant
North Carolina	525	3.4	0.2	3.2	537	6.8	7.6	-1.53	NS

NOTE: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.17. Projection parameters for science means

	Parameter estimates	
Correlation	α	β
.97	20.3460	0.9680
Variance-covariance		
	α	β
α	7.9064	-0.0150
β	-0.0150	0.0000

Table 5.18. Projection for science with accommodation adjustments

State	Projection	Standard error linking	Standard error prediction	Standard error projection	Actual TIMSS	Standard error TIMSS	Overall standard error	Z-Test	Significant difference
Alabama	488	4.2	0.3	4.1	485	6.2	7.4	0.31	NS
California	496	4.0	0.2	3.9	499	4.6	6.0	-0.34	NS
Colorado	547	4.0	0.2	3.8	542	4.4	5.8	0.94	NS
Connecticut	539	3.7	0.1	3.6	532	4.6	5.8	1.26	NS
Florida	533	3.7	0.1	3.6	530	7.3	8.1	0.34	NS
Indiana	536	3.3	0.1	3.2	533	4.8	5.7	0.52	NS
Massachusetts	561	3.7	0.2	3.6	567	5.1	6.2	-0.93	NS
Minnesota	544	3.5	0.2	3.4	553	4.6	5.8	-1.57	NS
North Carolina	522	3.6	0.1	3.5	532	6.3	7.2	-1.29	NS

NOTE: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.19. Projection parameters for low international benchmark in mathematics

	Parameter estimates	
Correlation	α	β
.90	17.4697	0.8063
Variance-covariance		
	α	β
α	0.6641	-0.0071
β	-0.0071	0.0001

Table 5.20. Predicted TIMSS-equivalents for low benchmark with adjustments for accommodations in grade 8 mathematics

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: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.21. Projection parameters for intermediate international benchmark in mathematics

	Parameter estimates	
Correlation	α	β
.92	10.7261	0.8567
Variance-covariance		
	α	β
α	0.3554	-0.0049
β	-0.0049	0.0001

Table 5.22. Predicted TIMSS-equivalents for intermediate benchmark with adjustments for accommodations in grade 8 mathematics

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: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.23. Projection parameters for high international benchmark in mathematics

	Parameter estimates	
Correlation	α	β
.94	2.1544	1.0356
Variance-covariance		
	α	β
α	0.0888	-0.0024
β	-0.0024	0.0001

Table 5.24. Predicted TIMSS-equivalents for high benchmark with adjustments for accommodations in grade 8 mathematics

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: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.25. Projection parameters for advanced international benchmark in mathematics

	Parameter estimates	
Correlation	α	β
.93	0.4132	1.1453
Variance-covariance		
	α	β
α	0.0087	-0.0009
β	-0.0009	0.0001

Table 5.26. Predicted TIMSS-equivalents for advanced benchmark with adjustments for accommodations in grade 8 mathematics

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: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.27. Projection parameters for low international benchmark in science

	Parameter estimates	
Correlation	α	β
.92	18.2179	0.7977
Variance-covariance		
	α	β
α	0.4500	-0.0048
β	-0.0048	0.0001

Table 5.28. Predicted TIMSS-equivalents 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: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.29. Projection parameters for intermediate international benchmark in science

	Parameter estimates	
Correlation	α	β
.95	12.1405	0.8437
Variance-covariance		
	α	β
α	0.2030	-0.0027
β	-0.0027	0.0000

Table 5.30. Predicted TIMSS-equivalents 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: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.31. Projection parameters for high international benchmark in science

	Parameter estimates	
Correlation	α	β
.97	1.4500	1.0586
Variance-covariance		
	α	β
α	0.0550	-0.0013
β	-0.0013	0.0000

Table 5.32. Predicted TIMSS-equivalents 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: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 5.33. Projection parameters for advanced international benchmark in science

Correlation	Parameter estimates	
	α	β
.96	-0.7354	1.1930
Variance-covariance		
	α	β
α	0.0088	-0.0007
β	-0.0007	0.0001

Table 5.34. Predicted TIMSS-equivalents 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: Two-tailed Z-test, with alpha = .05, no adjustment for multiple comparisons.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

References

Johnson, E.G., Cohen, J., Chen, W.-H., Jiang, T., and Zhang, Y. (2003). *2000 NAEP–1999 TIMSS linking report*. (NCES 2005-01). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

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

Chapter 6: Evaluation of the Quality of NAEP-TIMSS Linkages

Overview

HumRRO, serving as the NAEP Quality Assurance (QA) contractor for NCES, evaluated the results of the NAEP-TIMSS linking study and recommended how linkage results should be used in reporting estimated TIMSS distributions for each state participating in the 2011 grade 8 NAEP assessment. The purpose of this chapter is to convey key findings from their evaluation and summarize the evidence underlying these findings.

Overall Study Design

As described previously in Chapter 2, the design for the NAEP-TIMSS linking study included data from the 2011 operational NAEP assessment and the 2011 operational U.S. TIMSS assessment. In addition, "braided" booklets containing blocks of NAEP and TIMSS items were administered to special samples during the NAEP testing window using NAEP administration procedures and during the TIMSS testing window using TIMSS administration procedures. Figure 2.1 shows the samples and sample sizes used in developing and evaluating the NAEP-TIMSS linkages. Three key differences between NAEP and TIMSS administrations are illustrated in figure 2.1. The most obvious is the difference in the testing window: in the northern hemisphere, TIMSS is administered at the end of the school year; approximately three months after the NAEP tests are administered. A second difference is that the NAEP mathematics and science assessments are separate, each administered to a separate sample of students, while TIMSS combines the two assessments, administering both mathematics and science blocks to the same students. A final difference is that the NAEP state samples are included in the overall national sample. For the purposes of this study, TIMSS was administered to samples in nine states that were not included in the overall national TIMSS sample.

More detailed information on the braided-booklet samples and their use in developing the calibration (CAL), statistical projection (PRO), and statistical moderation (MOD) linkages are provided in Chapters 3, 4, and 5 of this report.

Evaluation Design

Three stages were included in the plan for evaluating the results of the linking study. The first stage of the evaluation involved identifying key differences between the two assessments that might affect the linkages or threaten the validity of the interpretation of predicted state level results. The second 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. The third stage included further examination and follow-up analyses of the findings from the primary evaluation in Stage 2.

Stage 1: Threats to Validity

HumRRO maintains the Quality Assurance Technical Panels (QATP),[9] a cadre of consultants with expertise in various aspects of assessments. Prior to analyzing any data, HumRRO conducted discussions with key QATP members to identify differences between the two assessments that might plausibly affect the scale score linkages. Figure 6.1 (shown below) lists key differences in six assessment components: (1) content, (2) sampling, (3) administration, (4) inclusion and accommodations, (5) analysis and scaling, and (6) reporting. Table 6.1 also displays the proposed plans for how both ETS and AIR addressed each difference in their analyses.

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 content or testing windows, could not be investigated directly from the available data and further investigations were beyond the scope of this evaluation. Note that the braided-booklet samples did provide estimates from each of the two assessments during each testing window, but the braided-booklet samples were too small to support separate analyses by state. Thus, HumRRO researchers were not able to investigate state differences in the additional learning students appeared to have obtained between the testing windows. Additionally, testing window differences were confounded with other differences in test administration procedures (e.g., testing length or testing time). Refer to *U.S. States in a Global Context* (NCES 2013-460) for specific differences between NAEP and TIMSS mathematics and science assessments.

Figure 6.1. Key differences between the NAEP and TIMSS assessments

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 accommodations	▪ Accommodation policy ▪ Exclusion policy
Analysis and scaling	▪ Conditioning model ▪ Treatment of not-reached items ▪ Establishing trend
Reporting	▪ Benchmarks ▪ Scale (score range, mean, SD)

[9] 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.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Stage 2: Primary Evaluation

Stage 2 of the evaluation involved analyzing the NAEP-TIMSS linking study data and applying the three linkages to the NAEP state samples for the nine validation states. Prior to analyzing the data, NAEP and TIMSS reports were reviewed and the statistics most likely to be used in reporting results from the linkages were identified as scale score means and the percentage of students at or above each of the TIMSS benchmark levels. Differences in the estimated TIMSS scale score standard deviations for each validation state were also examined, 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, differences in linkage estimates were also examined for subgroups defined by gender and, where sample size permitted, race/ethnicity.

Means

Tables 6.2 and 6.3 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. 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 using the MOD method include error variance associated with error in estimating the linkage functions. Figures 6.2 and 6.3 show confidence bounds estimated for each of the empirical and linkage-based estimates of state means.

Figure 6.2. Confidence bounds for state mean estimates for overall sample – mathematics

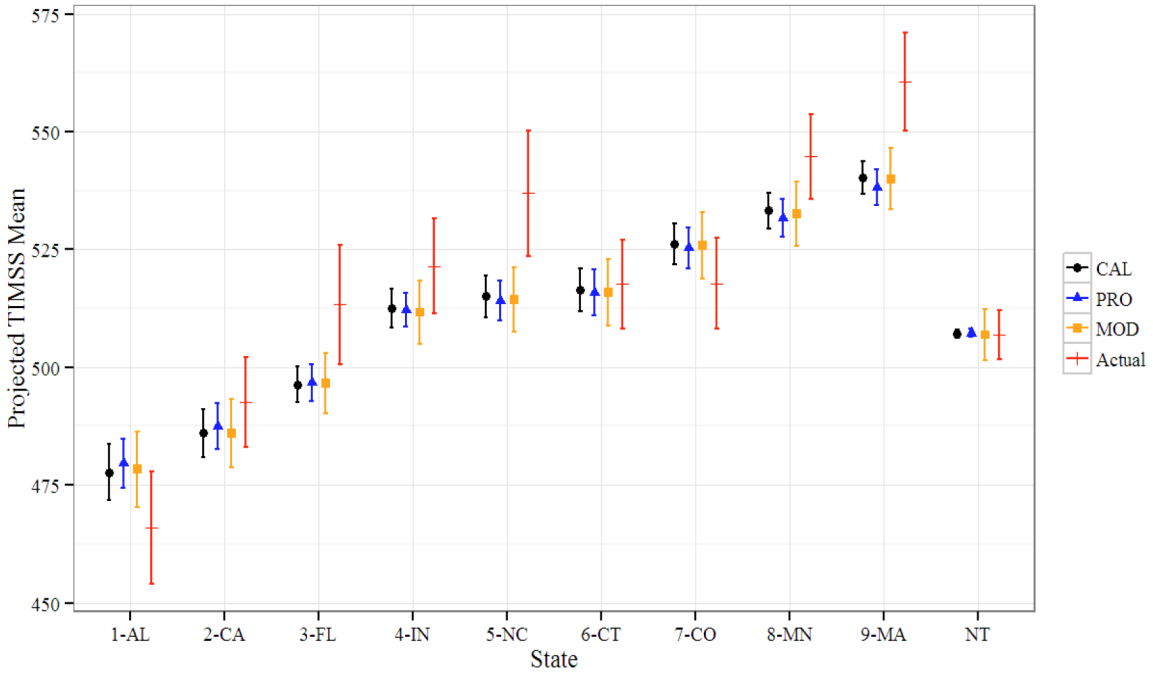
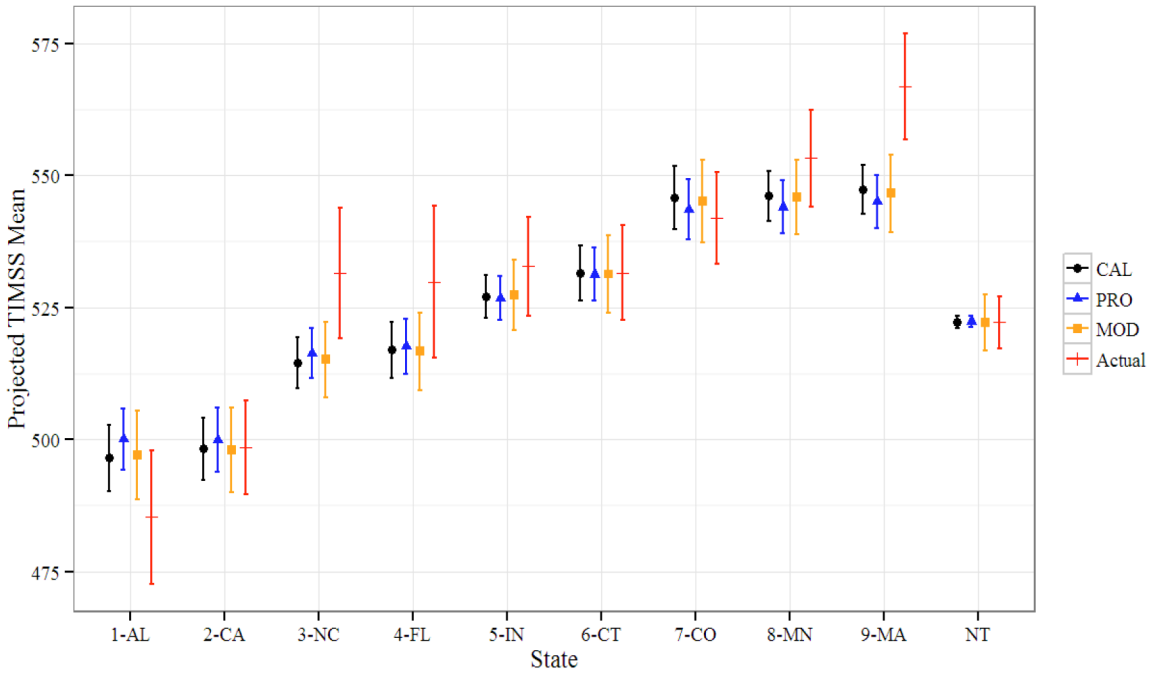


Figure 6.3. Confidence bounds for state mean estimates for overall sample – science



A key finding illustrated in figures 6.2 and 6.3 is that all three linkage methods yielded very similar projections of the state means. However, the confidence bounds for the empirical and each of the linkage-based estimates of state means did not overlap for several validation states. Note that the confidence bounds for the empirical TIMSS means are larger than for the linkage-based

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

projections because the TIMSS state samples are considerably smaller than the NAEP state samples used in generating the linkage-based projections.

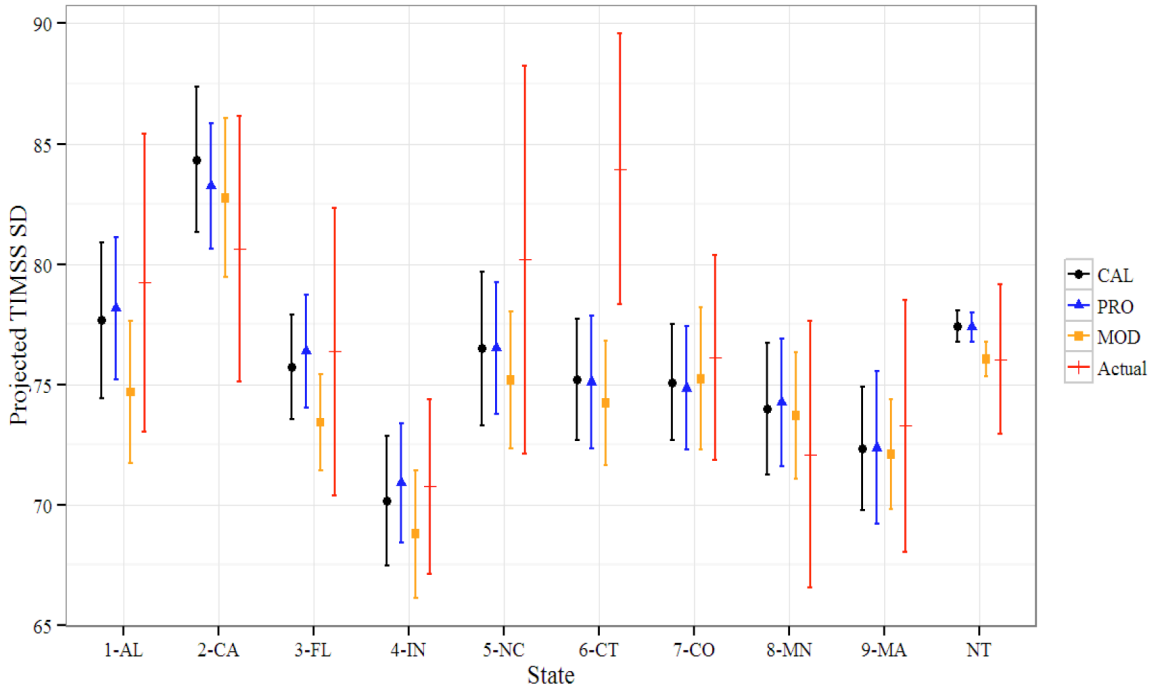
Tables 6.4 and 6.5 show results from tests of the statistical significance of the differences between the empirical and linkage-based estimates for mathematics and science respectively. As shown, the differences were statistically significant for nearly half of the validation states in mathematics and for at least two states in science, based on a two-tailed Z-test with significance level 0.05.

Standard Deviations

Tables 6.6 and 6.7 show differences between estimates of the standard deviation (SD) of TIMSS scale scores from the operational TIMSS validation state samples and from the NAEP state samples using each of the three linkage methods. 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 6.4 and 6.5 show confidence bounds estimated for each of the empirical and linkage-based estimates of state SDs.

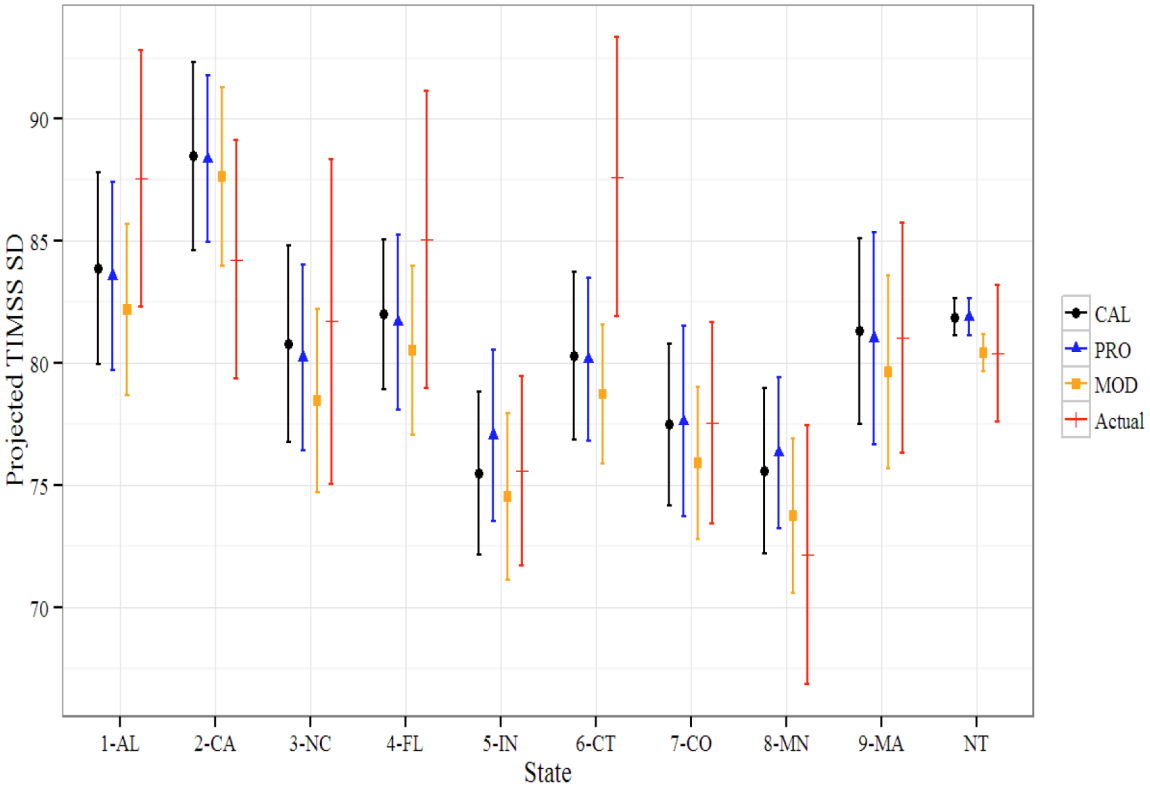
As shown in figures 6.4 and 6.5 below, the confidence bounds for the empirical and linkage-based estimates of state SDs overlapped for most, but not all validation states. Note that similar to the results for the state means, the confidence bounds for the empirical TIMSS SDs are larger than for the linkage-based projections because the TIMSS state samples are considerably smaller than the NAEP state samples used in generating the linkage-based projections.

Figure 6.4. Confidence bounds for state SD estimates for overall sample – mathematics



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Figure 6.5. Confidence bounds for state SD estimates for overall sample – science



Tables 6.8 and 6.9 show results from tests of the statistical significance of the differences between the empirical and linkage-based estimates for mathematics and science respectively. As shown, the differences were statistically significant for one of the nine validation states, suggesting that the SD projected estimates were more similar to the actual SD estimates than were the mean estimates.

Score Distributions

As part of the initial analyses, the extent to which the distributions of scores for each linkage method were similar to the actual TIMSS distributions for each validation state was also examined. As shown in figures 6.6 and 6.7, the projected mean score distributions in each validation state were similar to their respective actual TIMSS score distributions.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Figure 6.6. Score distribution for overall sample – mathematics

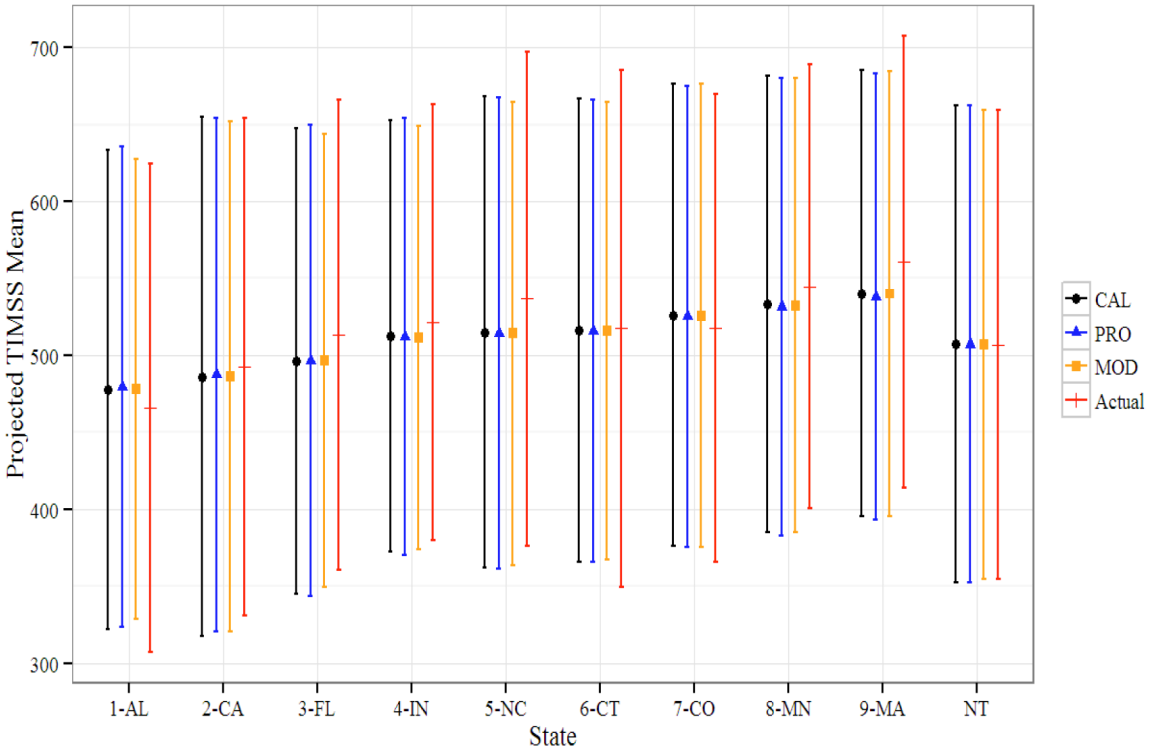
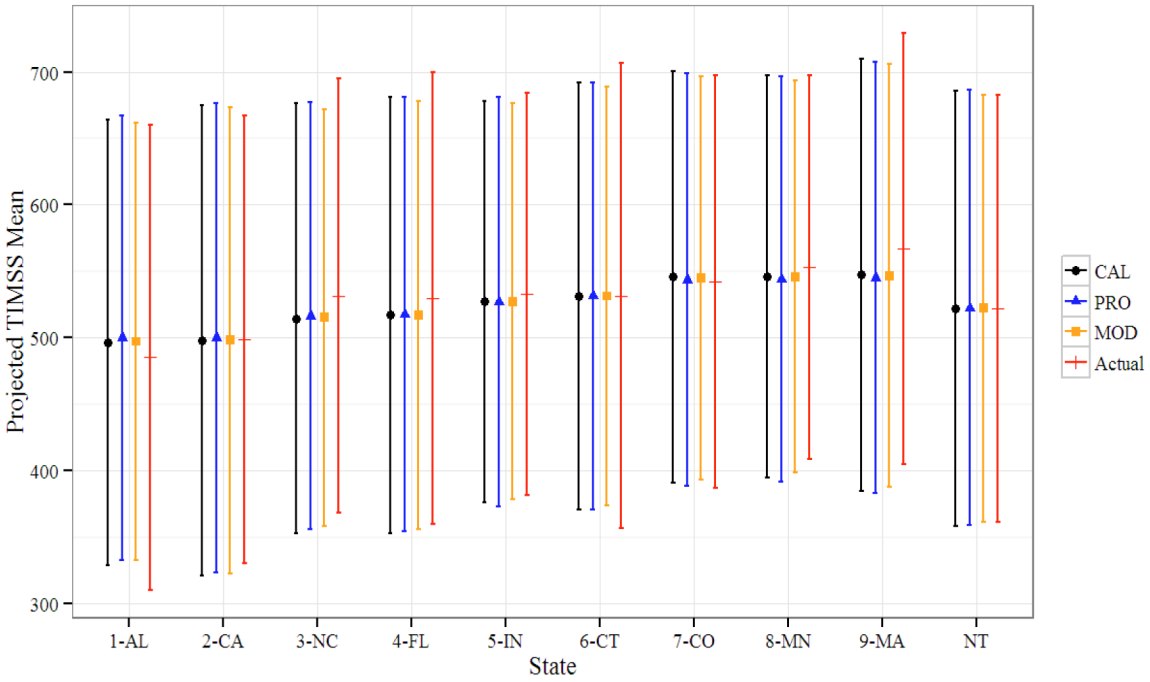


Figure 6.7. Score distribution for overall sample – science

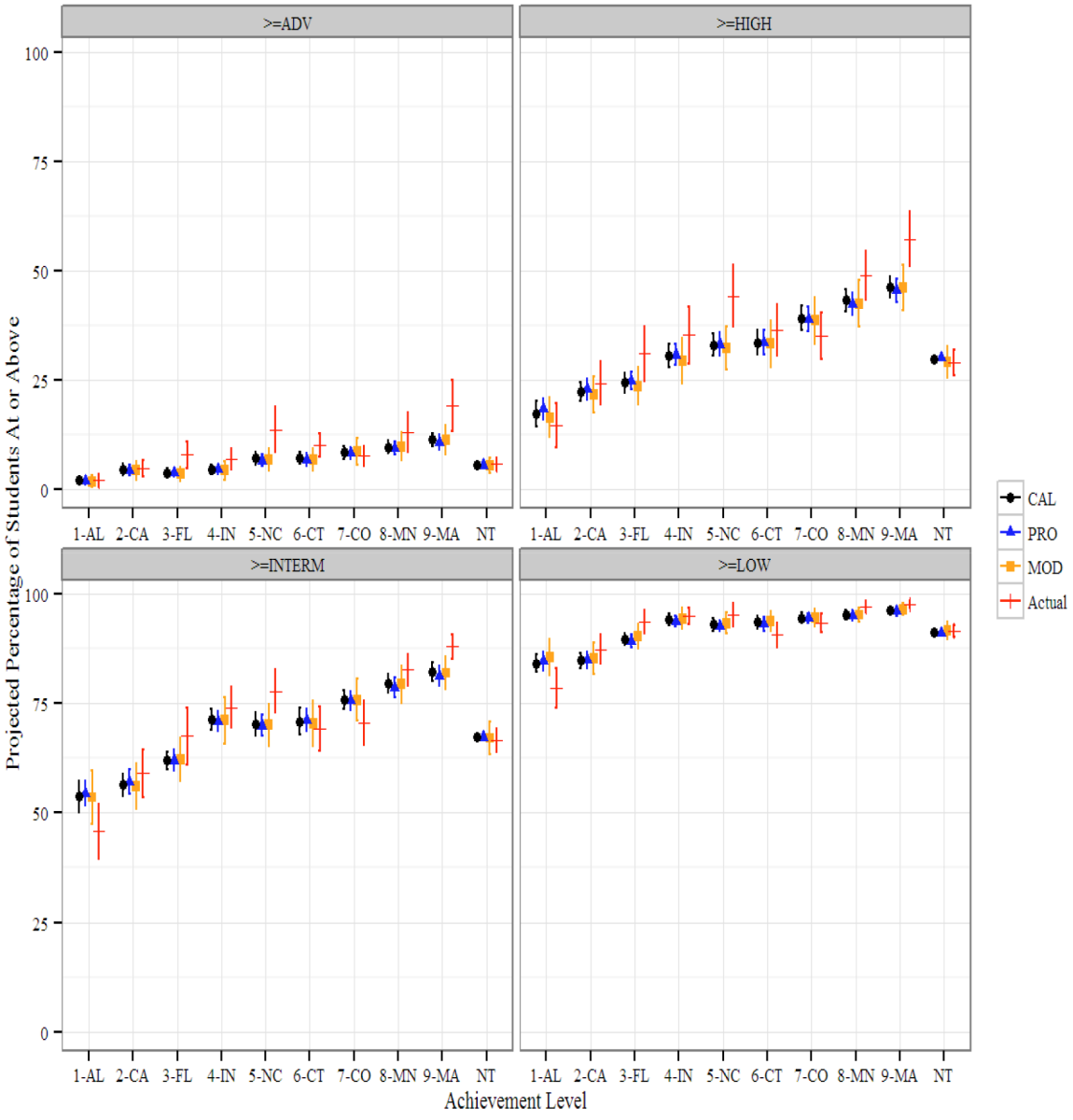


SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Benchmark Levels

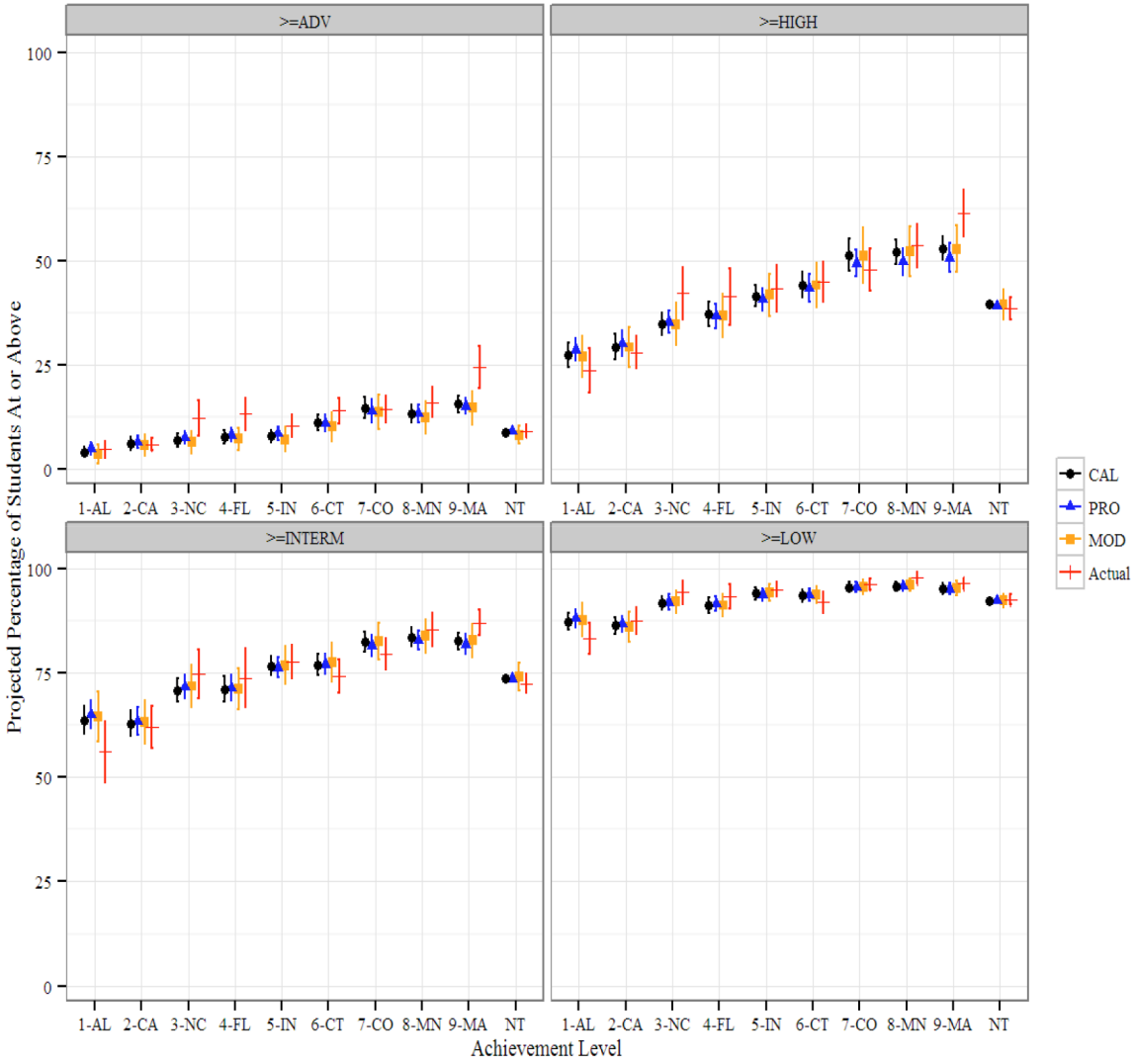
Tables 6.10 through 6.13 show differences in estimates of the percent above each of the TIMSS benchmark level cut points along with statistical tests of these differences. As with the state means estimates, differences between empirical and linkage-based estimates were larger than would be expected based on estimates provided by AIR and ETS of the standard error of each estimate (see also figures 6.8 and 6.9).

Figure 6.8. Confidence bounds for benchmark levels for overall sample – mathematics



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Figure 6.9. Confidence bounds for benchmark levels for overall sample – science



Gender

HumRRO researchers examined the differences in projected TIMSS means and projected percentage of students at or above each benchmark level between males and females. Tables 6.14 and 6.15 show differences for each gender between estimates of mean TIMSS scale scores from the operational TIMSS and 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 errors greater than half a scale score point 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 6.10 and 6.11 display the confidence bounds for the empirical linkage-based estimates of mean scores for both males and females within each validation state.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Figure 6.10. Confidence bounds for state mean estimates for male and female students – mathematics

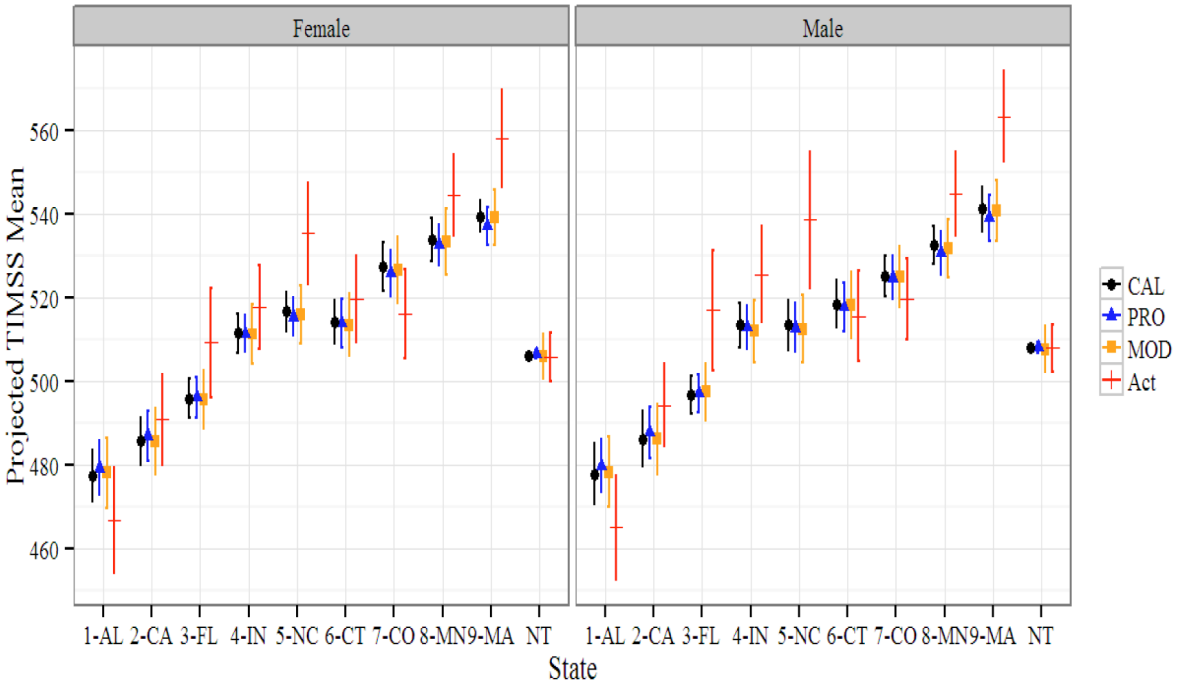
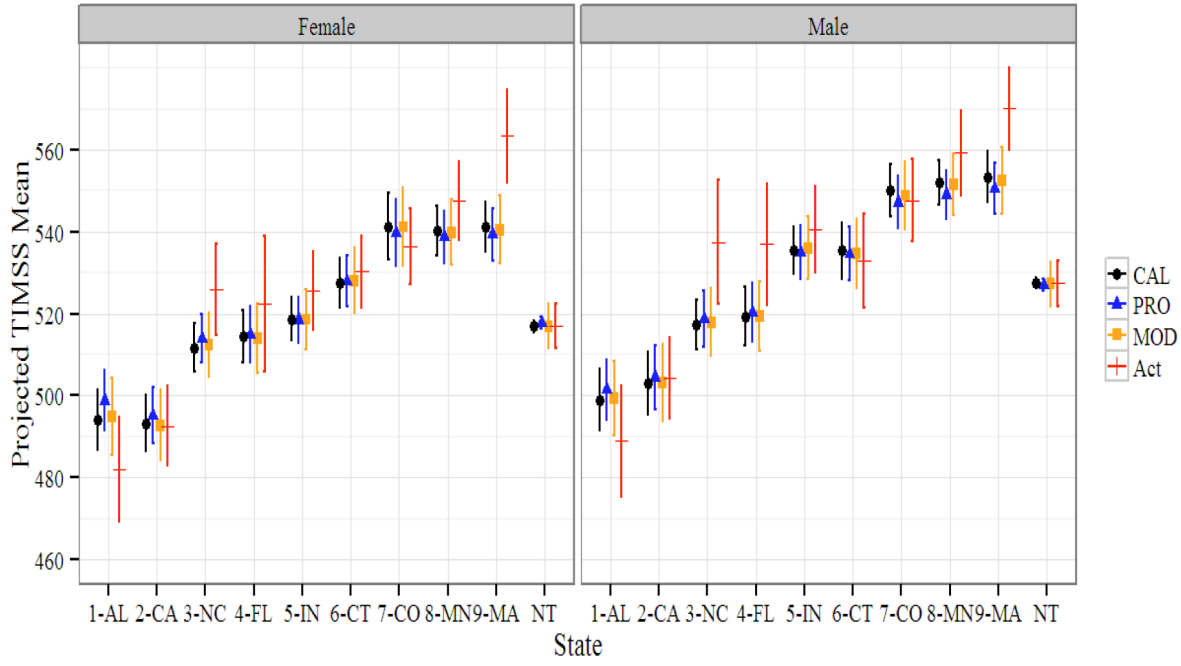


Figure 6.11. Confidence bounds for state mean estimates for male and female students – science



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Race/Ethnicity

The differences in projected TIMSS means for different racial/ethnic groups including White, Black, Hispanic, and Asian students were also examined. Tables 6.16 through 6.23 show differences between estimates of mean TIMSS scale scores from the operational TIMSS and each of the three linkage methods for each of these racial/ethnic groups. At the national level, some groups yielded not statistically significant estimation errors that were greater than several scale score points, compared to estimation errors by gender which were 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 (PRO), which accounted for some demographic information, yielded far smaller differences by race/ethnicity compared to the other two methods.

Stage 3: Preliminary Findings

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

Finding 1: The three different linkage methods yielded 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, implying that other differences between the two assessments must be contributing significant amounts of variation in the state level estimates.

Stage 4: Differences in Populations and Item Properties

Based on preliminary results, the factors that would most reasonably be associated with the larger-than-preferred linking error were explored. Specifically, the impact of two key differences between the NAEP and TIMSS assessments were investigated:

1. differences in exclusion and accommodation policies and
2. differences in the distribution of test item difficulty and item formats.

Additionally, whether the differences in racial/ethnic group size by state could be accounted for and thus reduce the linkage error was investigated. Specifically, how each adjustment might impact the difference between the empirical and linkage-based scale score estimates was examined. Other differences, such as the difference in testing window (and associated differences in exposure to instruction), could not be investigated within the scope of the current study.

Differences in Accommodation and Exclusion Rates

Tables 6.24 and 6.25 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.[10] It is important to note that the NAEP 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, NAEP accommodation rates have grown while exclusion rates have 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. The unavailability of specific accommodations likely results in students being excluded from TIMSS who would be included in NAEP. As shown, TIMSS exclusion rates are considerably higher than NAEP exclusion rates. The difference between the percentage of students excluded in the NAEP and TIMSS assessments also varies considerably from state to state.

Table 6.26 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 (from .72 to .81) 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. For the nation as a whole, 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 included and excluded students with disabilities and English language learners that provide information about specific student disabilities and characteristics. TIMSS does not collect comparable background information on the students tested, and no information is available about excluded students.

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

The first investigation involved excluding certain populations from the linking that are related to the exclusion and accommodation differences between NAEP and TIMSS. More specifically, the groups excluded from the linking were (1) all accommodated students and (2) all students with disabilities (SD) and English Language Learners (ELL) to see to what extent, if any, the linkage-based estimates would become more accurate.

Exclude All SD/ELL Students ("NoSDE")

First, all SD/ELL students were excluded, whether they received accommodations or not. Tables 6.27 and 6.28 display the differences between the empirical and linkage-based estimates for mathematics and science when all SD and ELL students were excluded. This approach led to large errors at the national level (more than 10 points) and did not substantially reduce the error in state-level predictions as indicated by RMSEs (compared to tables 6.2 and 6.3).

Exclude All Accommodated Students ("NoACC")

Tables 6.29 and 6.30 display the differences between the empirical and linkage-based estimates for mathematics and science when only accommodated students were excluded. Although the RMSEs were lower than they were for the overall sample (see tables 6.2 and 6.3), excluding all accommodated students still yielded large errors at the national level with a modest improvement in the accuracy of state level estimates. Given the large errors at the national level, it was concluded that neither the NoSDE nor the NoACC approaches were warranted.

Reweight Accommodated Students ("AccRW")

Several methods for adjusting the NAEP samples to reduce the impact of differences in exclusion and accommodation rates were then investigated. The first adjustment treated all accommodated students as a homogenous group and assumed that some students who would be accommodated on NAEP would be excluded on TIMSS. The Accommodations Reweighted approach (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. This approach rendered the impact of excluded students relatively equal for NAEP and TIMSS by reducing the contribution of accommodated students in NAEP. 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. As seen in tables 6.31 and 6.32, compared to the NoACC approach, the AccRW adjustment led to smaller errors at the national level and decreased the RMSEs for the state-level estimates for both mathematics and science. However, differences between the empirical and the linkage-based estimates remained large.

Differentially Reweight Accommodated Students ("AccDRW")

This exploration was then refined with a finer treatment of accommodated students. Options for reweighting accommodated students differentially based on type of accommodation were examined. The AccDRW adjustment assumed some accommodated students are more likely to be

excluded from TIMSS than are others. The lowest scoring groups based on type of accommodation (or number of accommodations) were identified. Assuming that the low-scoring groups were more likely to be excluded from TIMSS, this would lead to the largest differences in population estimates. To determine what the lowest scoring groups were, the NAEP means for students with different types of accommodations were examined and six accommodation groups for mathematics and science were defined separately. Cases (set weights to zero) for the lowest performing groups were then eliminated until the overall TIMSS exclusion rate was reached. As can be seen in tables 6.33 and 6.34, the differential reweighting had little to no impact on overall NAEP means and thus, little to no impact on TIMSS estimates would be expected. Thus, the AccDRW adjustment provided essentially the same results as the proportional reweighting.

Adjust for Accommodation Differences ("AccADJ")

The Accommodations Adjustment (AccADJ) involved an empirically derived adjustment based on the percentage of students accommodated in NAEP. When state results for all states were estimated, the actual TIMSS exclusion rates for states not participating in TIMSS were not available. However, the percent accommodated in NAEP was the best available predictor of the difference in NAEP and TIMSS exclusion rates. The correlations shown in table 6.26 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. More precisely, adjustment coefficients were estimated for each linkage method by regressing the difference between the empirical and linkage-based estimates on the difference between the national and state accommodation rates, suppressing the intercept:

$$\left(\overline{\boldsymbol{T}}(\boldsymbol{i}) - \widehat{\boldsymbol{T}}(\boldsymbol{i})\right) = \boldsymbol{\beta_{adj}} \times (\%\, \boldsymbol{Acc}(\boldsymbol{i}) - \%\, \boldsymbol{AccNT}) + \boldsymbol{Error} \tag{6.1}$$

where $\overline{\boldsymbol{T}}(\boldsymbol{i})$ is the empirical estimate, $\widehat{\boldsymbol{T}}(\boldsymbol{i})$ is the linkage-based estimate, $\boldsymbol{\%\ Acc(i)}$ is the NAEP accommodation rate for state *i* and $\%\ \boldsymbol{AccNT}$ is the national accommodation rate. The adjustment coefficient in this model is estimated by

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{adj}} = \frac{\sum_{\boldsymbol{i}}(\%\, \boldsymbol{Acc}(\boldsymbol{i}) - \%\, \boldsymbol{AccNT}) \times \left(\overline{\boldsymbol{T}}(\boldsymbol{i}) - \widehat{\boldsymbol{T}}(\boldsymbol{i})\right)}{\sum_{\boldsymbol{i}}(\%\, \boldsymbol{Acc}(\boldsymbol{i}) - \%\, \boldsymbol{AccNT})^{\boldsymbol{2}}} \tag{6.2}$$

See table 6.35 for the AccADJ coefficients to use in computing the AccADJ mean estimates. Equations 6.3 and 6.4 were used to apply the adjustments to the estimated state TIMSS mean, $\widehat{\boldsymbol{T}}(\boldsymbol{i})$:

- For mathematics:

$$\widehat{\boldsymbol{T}}_{\boldsymbol{adj}}(\boldsymbol{i}) = \widehat{\boldsymbol{T}}(\boldsymbol{i}) + (\widehat{\boldsymbol{\beta}}_{\boldsymbol{adj}} \times (\%\, \boldsymbol{Acc}(\boldsymbol{i}) - \boldsymbol{9.68})) \tag{6.3}$$

  where, 9.68 is the national accommodation rate for mathematics, and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{adj}}$ is the corresponding mathematics coefficient from table 6.35.

- For science:

$$\widehat{T}_{adj}(i) = \widehat{T}(i) + (\widehat{\beta}_{adj} \times (\%\, Acc(i) - 10.59)) \quad (6.4)$$

where, 10.59 is the national accommodation rate for science, and $\widehat{\beta}_{adj}$ is the corresponding science coefficient from table 6.35.

A comparison of tables 6.36 and 6.37 to tables 6.2 and 6.3 reveals that the AccADJ introduced no additional error at the national level and resulted in substantially lower RMSEs for the state-level estimates compared to the unadjusted linkage-based estimates.

Racial/Ethnic Differences

In reviewing initial results with a 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 rates among racial/ethnic groups or might have resulted from differences in school and class participation rates by race/ethnicity that were not fully accounted for in nonresponse adjustments. Race/ethnicity differences may also have resulted from sampling error, particularly in states with relatively small frequencies for some groups.

Adjust for Differences in Racial Distribution ("RaceADJ")

The Race Adjustment (RaceADJ) involved reweighting the TIMSS samples for each state to yield the racial/ethnic distribution of the NAEP state sample. This resulted in adjustment to the empirical TIMSS estimate to reflect the NAEP racial distribution, which was generally more stable because of the large sample sizes. As seen in tables 6.38 and 6.39, the RaceADJ minimally reduced the RMSEs relative to tables 6.2 and 6.3. In any event, this was not a practical adjustment due to the fact that this adjustment required states to have an empirical TIMSS score.

Adjust for Differences in Accommodations and Racial Distribution ("RaceAccADJ")

An Accommodations and Racial Adjustment (RaceAccADJ) that combined the race/ethnicity adjustment and the adjustment based on accommodation rates was also examined. The prediction error is equivalent to the difference between the accommodations-adjusted (AccADJ) mean estimates (see tables 6.36 and 6.37) and the race-adjusted (RaceADJ) empirical TIMSS means (see tables 6.38 and 6.39). As seen in tables 6.40 and 6.41, the RaceAccADJ resulted in substantially smaller RMSEs compared to the unadjusted linkage-based estimates in tables 6.2 and 6.3. Figures 6.12 and 6.13 display the adjusted means using the RaceAccADJ. While the RaceAccADJ did improve prediction, it is not feasible to use this approach for states not participating in TIMSS, since TIMSS race/ethnicity distributions would not be available.

Figure 6.12. Adjusted projected TIMSS means using the race and accommodation adjustment (RaceAccADJ) with confidence bounds – mathematics

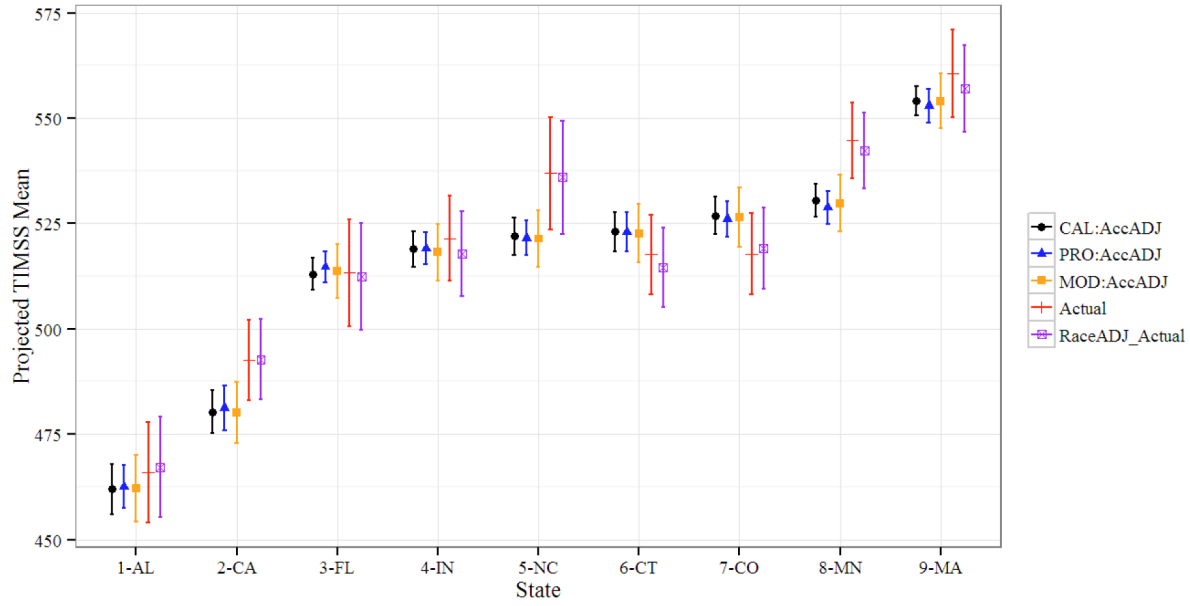
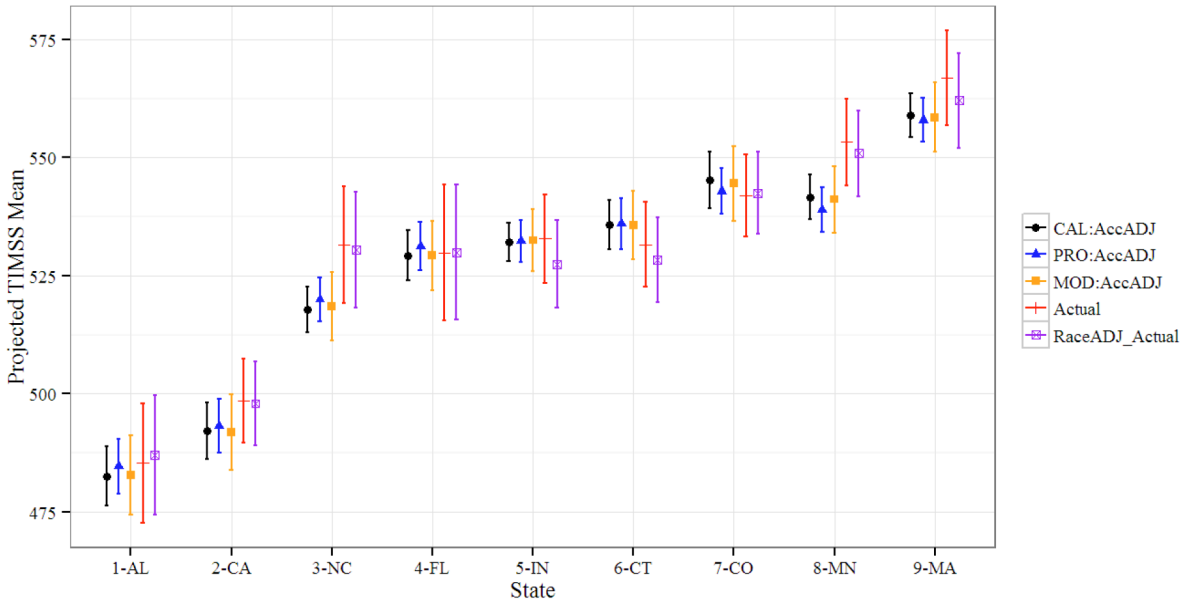


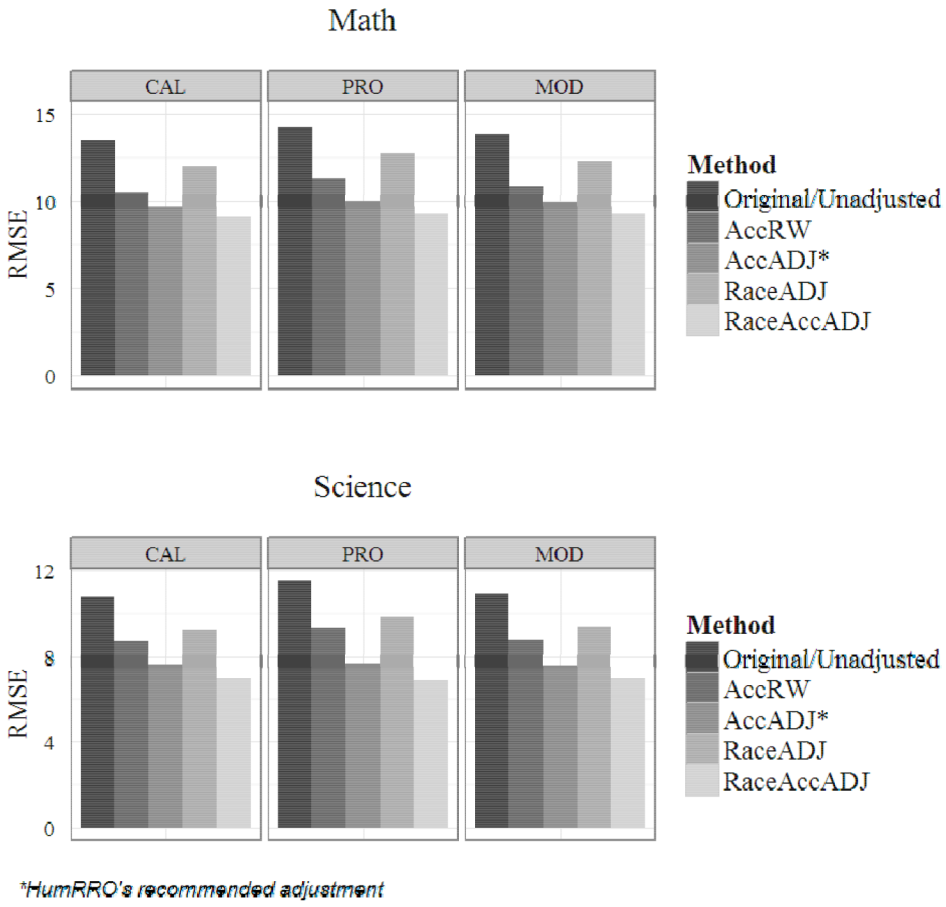
Figure 6.13. Adjusted projected TIMSS means using the race and accommodation adjustment (RaceAccADJ) with confidence bounds – science



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Figure 6.14 shows the root mean square error for estimates of state means using each of the three linkage methods and each of the four adjustments that were pursued in detail. The NoSDE and NoACC approaches were clearly inferior (as described above) and AccRW and AccDRW yielded the same predictions as AccRW, so these methods are not included here. Inspection of figure 6.14 reveals that the race/ethnicity adjustment, by itself, yielded 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 6.14. Comparison of error rates resulting from each of the four adjustments for exclusion and accommodation differences



Finding 3. 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. Item difficulties were found to be similar for the

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

two assessments, although there were differences in the number of short and extended constructed response item types. A check was made to see if students in some states performed better on these item types compared to students in other states. As shown in table 6.42, while the constructed response items were more difficult overall, there were no significant interactions between state and item type. Based on these findings, it was concluded that there was little or no possibility of creating a useful adjustment to state mean estimates based on item difficulty or item type differences between NAEP and TIMSS.

Recommended Linkage Method and Final Adjustment

Based on the various accommodation adjustments that were examined, it was found that the empirically derived accommodation adjustment (AccADJ) would result in the state mean predictions with the lowest RMSE.[11] Additionally, because each of the three linkage methods resulted in similar differences between the empirical and linkage-based estimates, additional analyses using the statistical moderation method were pursued. This method is the least complex and easiest to implement relative to the statistical projection and statistical calibration methods.

Examining Sources of Error

Adjustments to Predicted State Mean Estimates

After adjustment for state differences in accommodation rates (AccADJ) (equations 6.3 and 6.4), 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. See tables 6.43 and 6.44 for unadjusted and adjusted mean estimates for statistical moderation. In order to account for the residual prediction error after the accommodation adjustment, ways to estimate additional variance in the linkage-based estimates were examined.

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 estimate requires examination of the variance of differences between the empirical and linkage-based estimates of validation state means and subtraction of known estimates of variance due to NAEP and TIMSS sampling error, measurement error, and linkage error. 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, the projected TIMSS mean estimates that included the accommodation adjustment described above were used.

[11] Although the RMSE was lowest for the RaceAccADJ, because the contractor would not be able to apply the adjustment to all 52 states/jurisdictions, they did not view this as a viable adjustment.

The original standard error estimate included measurement (*m*, based on plausible value variance), sampling (*s*, using the jackknife weights), and linkage (*l*) error components. The standard error, $\boldsymbol{SE}_{\widehat{\boldsymbol{T}}_{\boldsymbol{adj}}(\boldsymbol{i})}$, is the square root of the sum of these error variance components:

$$\boldsymbol{SE}_{\widehat{\boldsymbol{T}}_{\boldsymbol{adj}}(\boldsymbol{i})} = \sqrt{\boldsymbol{s}^{\boldsymbol{2}}_{\widehat{\boldsymbol{T}}(\boldsymbol{i}),\boldsymbol{s}} + \boldsymbol{s}^{\boldsymbol{2}}_{\widehat{\boldsymbol{T}}(\boldsymbol{i}),\boldsymbol{m}} + \boldsymbol{s}^{\boldsymbol{2}}_{\widehat{\boldsymbol{T}}(\boldsymbol{i}),\boldsymbol{l}}} \tag{6.5}$$

The estimated and expanded standard error estimate included the measurement, sampling, and linkage error components, and a "model or prediction error" component. Empirical estimates of model error *variance* from the nine validation states were developed.

$$\boldsymbol{SE}'_{\widehat{\boldsymbol{T}}_{\boldsymbol{adj}}(\boldsymbol{i})} = \sqrt{\boldsymbol{s}^{\boldsymbol{2}}_{\widehat{\boldsymbol{T}}(\boldsymbol{i}),\boldsymbol{s}} + \boldsymbol{s}^{\boldsymbol{2}}_{\widehat{\boldsymbol{T}}(\boldsymbol{i}),\boldsymbol{m}} + \boldsymbol{s}^{\boldsymbol{2}}_{\widehat{\boldsymbol{T}}(\boldsymbol{i}),\boldsymbol{l}} + \boldsymbol{s}^{\boldsymbol{2}}_{\widehat{\boldsymbol{T}}(\mathbf{model\ error})}} \tag{6.6}$$

Tables 6.45 and 6.46 show estimates of the different NAEP and TIMSS variance components for each validation state and the squared difference between the linkage-based and empirical TIMSS mean estimates for the state. The variance component estimates across the nine validation states were averaged, and then these variance components were subtracted 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 the specific amounts of variance cannot be attributed to specific differences. This residual variation was labeled as "model error" to indicate that the variance results from differences in the two assessment models. The estimates were 55.78 for mathematics and 13.15 for science.[12] This adjustment accounted for NAEP and TIMSS differences due to a variety of sources, other than accommodation rates, that introduce state-level variation.

Tables 6.47 and 6.48 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 6.15 and 6.16 display the accommodation adjustment means (AccADJ) for mathematics and science with model error.

[12] The model or prediction error estimates takes the residual error from predicting the TIMSS mean from the Projected NAEP mean and the percent accommodated and adjusts (enlarges) this error estimate to account for the use of one degree of freedom. Estimates of NAEP and TIMSS sampling and measurement error and linkage error variance were subtracted from the residual error variance. What is left is the "model" error.

Figure 6.15. Adjusted projected TIMSS means using the accommodation adjustment (AccADJ) and incorporating model error in the confidence bands – mathematics

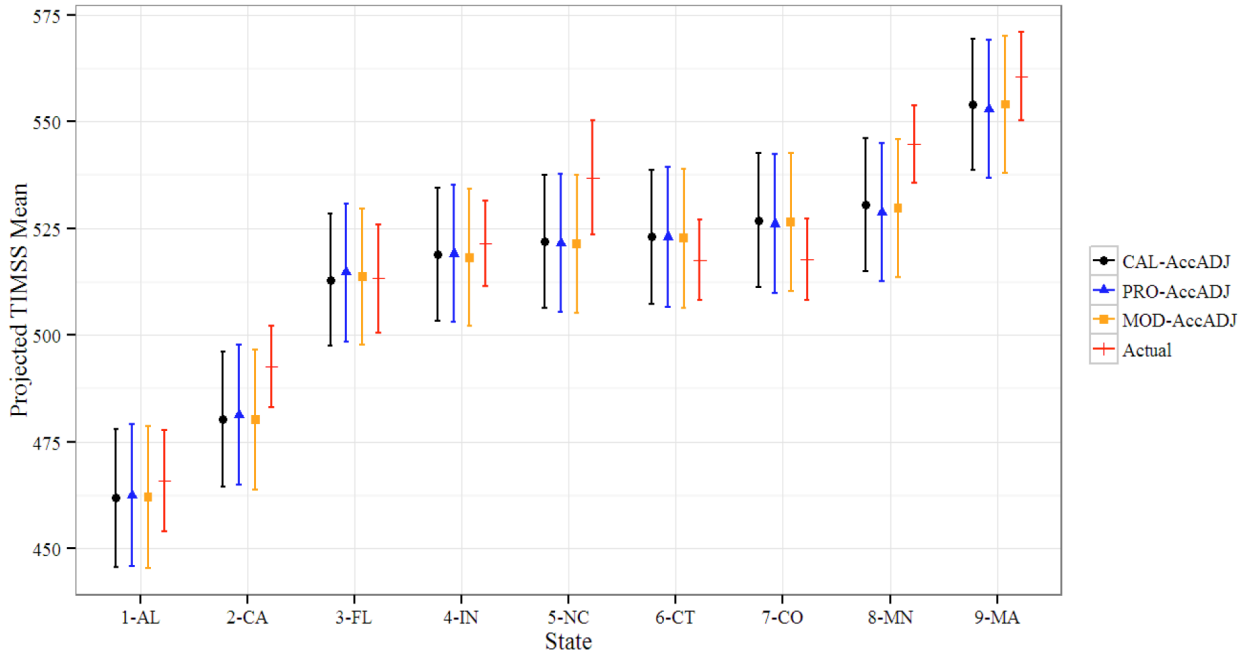
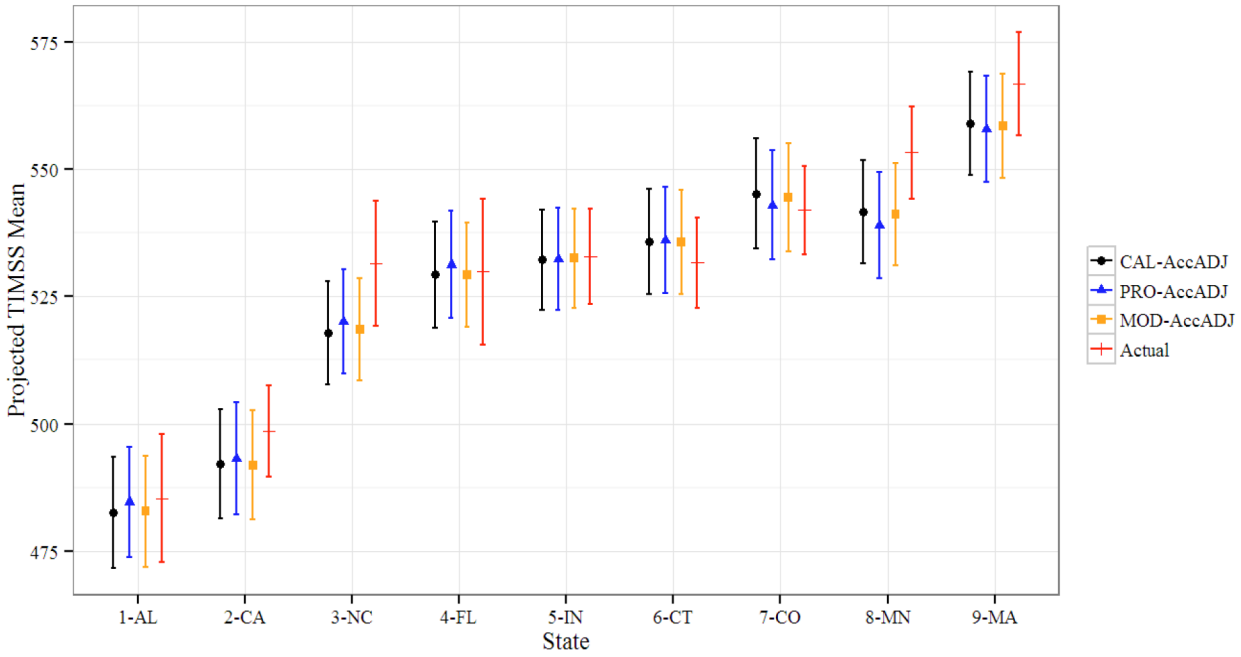


Figure 6.16. Adjusted projected TIMSS means using the accommodation adjustment (AccADJ) and incorporating model error in the confidence bands – science



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Adjustments to the Estimates of Percent Above Benchmark Cut Points

The contractor investigated two approaches for adjusting percent above (benchmark level) cut estimates using empirical adjustment based on the percentage of students accommodated in NAEP: normal approximation and direct adjustment.

Normal Approximation

The first approach is based on a normal approximation to the projected TIMSS score distribution (AccADJ_Normal). In this approach, the original percent above cut estimate was first converted into the TIMSS scale score metric using the inverse normal cumulative distribution with mean equal to the unadjusted TIMSS mean estimate. This calculation yields a "normalized" cut score which may differ from the original cut score depending on how the projected 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. In sum, the adjustment comprised three steps:

1. Apply an inverse cumulative normal transformation to the original percentage estimates using the unadjusted mean to put them into the TIMSS scale score metric. This gives a "normalized" cut score which may differ slightly from the original cut score if the score distribution is not originally normal. In equation 6.7 below, $\widehat{\boldsymbol{C}}_{\boldsymbol{b}}(\boldsymbol{i})$ is equal to the normalized cut score for each benchmark *b* in state *i*, $\widehat{\boldsymbol{P}}_{\boldsymbol{b}}(\boldsymbol{i})$ is equal to the unadjusted percent above cut estimates in state *i*, and $\boldsymbol{F^{-1}}$ is equal to the inverse normal distribution function.

$$\widehat{\boldsymbol{C}}_{\boldsymbol{b}}(\boldsymbol{i}) = \boldsymbol{F}^{-1}(\mathbf{1} - \widehat{\boldsymbol{P}}_{\boldsymbol{b}}(\boldsymbol{i})) \tag{6.7}$$

2. Apply the accommodation adjustment given above (equations 6.3 and 6.4) for estimated state means based on the percentage of students accommodated in each state to obtain $\widehat{\boldsymbol{T}}_{\boldsymbol{adj}}(\boldsymbol{i})$.

3. Use the cumulative normal distribution with mean equal to the adjusted mean (from Step 2) to convert the normalized cut score (from Step 1) back to the percentile metric, that is to estimate the percent above the cut given the shift in the mean. In equation 6.8 below, $\widehat{\boldsymbol{P}}_{\boldsymbol{adj,b}}(\boldsymbol{i})$ is equal to the adjusted percent-above-cut estimate for each benchmark in state *i*, and $\boldsymbol{SD}_{\widehat{\boldsymbol{T}}}(\boldsymbol{i})$ is the standard deviation estimated TIMSS score distribution in state *i*.

$$\widehat{\boldsymbol{P}}_{\boldsymbol{adj,b}}(\boldsymbol{i}) = \mathbf{1} - \boldsymbol{F}^{-1}\left(\frac{\widehat{\boldsymbol{C}}_{\boldsymbol{b}}(\boldsymbol{i}) - \widehat{\boldsymbol{T}}_{\boldsymbol{adj}}(\boldsymbol{i})}{\boldsymbol{SD}_{\widehat{\boldsymbol{T}}}(\boldsymbol{i})}\right) \tag{6.8}$$

To estimate the standard error for the adjusted percent above cut estimate, researchers added and subtracted the adjusted estimate of the standard error of the TIMSS mean estimate that included model error (see table 6.49). The standard error estimate is half of the difference between their corresponding percentiles based on the normal distribution with adjusted mean. In sum, the SE adjustment comprised two steps:

1. Convert the adjusted percentile described above back into the TIMSS scale score metric using the inverse cumulative normal distribution with the projected NAEP mean and standard deviation. This is just $\widehat{\boldsymbol{C}}_{\boldsymbol{b}}(\boldsymbol{i})$ obtained in Step 1.

2. Add and subtract the adjusted estimate of the ***standard error*** of the state mean described in equation 6.6 and convert each value back into the percentile metric using the cumulative normal distribution function. The standard error estimate, $\boldsymbol{SE}_{\widehat{\boldsymbol{P}}_{\boldsymbol{adj,b}}}(\boldsymbol{i})$, is half of the difference between the percentiles derived from adding and from subtracting the adjusted standard error.

$$\boldsymbol{SE}_{\widehat{\boldsymbol{P}}_{\boldsymbol{adj,b}}}(\boldsymbol{i}) = 0.5 \times \left[\boldsymbol{F}\left(\frac{\widehat{\boldsymbol{C}}_{\boldsymbol{b}}(\boldsymbol{i}) + \boldsymbol{SE}'_{\widehat{\boldsymbol{T}}_{\boldsymbol{adj}(\boldsymbol{i})}} - \widehat{\boldsymbol{T}}_{\boldsymbol{adj}(\boldsymbol{i})}}{\boldsymbol{SD}_{\widehat{\boldsymbol{T}}}(\boldsymbol{i})} \right) - \boldsymbol{F}\left(\frac{\widehat{\boldsymbol{C}}_{\boldsymbol{b}}(\boldsymbol{i}) - \boldsymbol{SE}'_{\widehat{\boldsymbol{T}}_{\boldsymbol{adj}(\boldsymbol{i})}} - \widehat{\boldsymbol{T}}_{\boldsymbol{adj}(\boldsymbol{i})}}{\boldsymbol{SD}_{\widehat{\boldsymbol{T}}}(\boldsymbol{i})} \right) \right] \quad (6.9)$$

Direct Adjustment

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 achievement level. The adjusted projected percent-above-cut estimates were obtained by adding the adjustment to the original projected percent-above-cut estimate. The common form of the adjustment equations for mathematics and science are shown below (equations 6.10 and 6.11), where $\widehat{\boldsymbol{P}}_{\boldsymbol{adj,b}}(\boldsymbol{i})$ and $\widehat{\boldsymbol{P}}_{\boldsymbol{b}}(\boldsymbol{i})$ are the adjusted and unadjusted percent-above-cut estimate, respectively, $\boldsymbol{\beta}_{\boldsymbol{AccAdj,b}}$ is the adjustment coefficient for achievement benchmark level *b*.

- For mathematics:

$$\widehat{\boldsymbol{P}}_{\boldsymbol{adj,b}}(\boldsymbol{i}) = \widehat{\boldsymbol{P}}_{\boldsymbol{b}}(\boldsymbol{i}) + \boldsymbol{\beta}_{\boldsymbol{AccAdj,b}} \times (\%\, \boldsymbol{Acc}_{\boldsymbol{i}} - \mathbf{9.68}) \quad (6.10)$$

- For science:

$$\widehat{\boldsymbol{P}}_{\boldsymbol{adj,b}}(\boldsymbol{i}) = \widehat{\boldsymbol{P}}_{\boldsymbol{b}}(\boldsymbol{i}) + \boldsymbol{\beta}_{\boldsymbol{AccAdj,b}} \times (\%\, \boldsymbol{Acc}_{\boldsymbol{i}} - \mathbf{10.59}) \quad (6.11)$$

Corresponding adjusted standard errors were obtained in the same fashion as in accommodation adjustment for the mean. An unbiased estimate of the mean squared error was obtained by dividing the sum of the squared difference between adjusted NAEP projected and empirical TIMSS percent above cut estimates by eight degrees of freedom. Researchers 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 projected percent-above-cut estimate is the square root of the sum of the model error and the original variance. In equation 6.12 below, $SE'_{\hat{P}_{adj,b}(i)}$ is the adjusted standard error for projected percent-above-cut estimate; $s^2_{P_b(i),s}$, $s^2_{P_b(i),m}$, and $s^2_{P_b(i),l}$ are estimated variance components from sampling, measurement, and linking errors; and $s^2_{P_b(\text{model error})}$ is estimated model error.

$$SE'_{\hat{P}_{adj,b}(i)} = \sqrt{s^2_{\hat{P}_b(i),s} + s^2_{\hat{P}_b(i),m} + s^2_{\hat{P}_b(i),l} + s^2_{\hat{P}_b(\text{model error})}} \quad (6.12)$$

Table 6.49 shows the mean squared errors (MSEs) for the unadjusted NAEP projected percent above cut estimates and the two adjusted estimates. Tables 6.50 through 6.53 compare the two percent-above-cut adjustments (AccADJ_Normal and AccADJ_Direct) with model error to the unadjusted estimates without model error. As seen in tables 6.50 through 6.53, the AccADJ_Direct resulted in one negative estimate (table 6.53) and negative model errors for three of the benchmark levels (tables 6.50 – 6.52).

These results, combined with comparisons of the MSEs suggest that the normal approximation adjustment was as good as or better than the direct adjustment across all benchmark levels. The normal approximation was also the more parsimonious method because it required one adjustment equation as opposed to four separate adjustment equations by achievement level used in the direct approach. For these reasons, researchers recommended the normal approximation method to adjust the percent above cut estimates for differences in NAEP accommodation rates.

Summary of Recommendations

Recommendation 1: Use estimates from the statistical moderation linkages.

SUPPORTING REFERENCES:

- Tables 6.2 – 6.5, 6.10 – 6.13
- Figures 6.2 – 6.3

While results indicated slight improvements in estimates using the calibration (CAL) approach, the differences do not justify using this approach in future years because of the extra effort and expense associated with it. In addition, all of the assumptions of the calibration approach have not been fully investigated in this study, 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 mean estimates.

SUPPORTING REFERENCES:

- Tables 6.24 – 6.26, 6.36 – 6.37, 6.43 – 6.44
- Figure 6.14
- Equations 6.3 – 6.4

The other adjustments examined in this study 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 clear reduction in differences between the linkage-based and empirical estimates of validation state means.

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

SUPPORTING REFERENCES:

- Tables 6.45 – 6.48
- Figures 6.15 – 6.16
- Equations 6.5 – 6.6, 6.9

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 was 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 6.49 – 6.53
- Equations 6.7 – 6.8

As described above, the normal approximation approach avoided negative estimates of the percent of students above a cut point and was more parsimonious in that it used the same single adjustment equation as the TIMSS mean score estimates rather than four separate adjustment equations for each subject.

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 were presented. Readers could construe differences among states or between states and countries/education systems where no true differences exist. The contractor strongly encourages the inclusion of confidence intervals and/or error estimates in all reporting to minimize misinterpretation of the information by end users.

Tables

Table 6.1. ETS and AIR preliminary approaches to address differences in NAEP and TIMSS

Assessment process	Differences in…	ETS (CAL and PRO)	AIR (MOD)
Content	• Content coverage • Slight differences in item format • Test administration time	• CAL: Examine correlations between NAEP and TIMSS to determine the extent to which both assessments measure the same constructs. • PRO: Does not require the same construct, although differences in content will affect the validity of specific interpretations based on linkage results.	• The content does not have to be the same and there is no way to statistically adjust for these differences. Content differences will affect the interpretations, but it is up to the reader or the user to take these differences into account when making interpretations.
Sampling	• Sampling method • Sample size • Minimum acceptable participation rate	• TIMSS scores projected from NAEP operational sample using both calibration and statistical projection linking approaches will include sampling error resulting from NAEP sampling design and measurement error computed from NAEP plausible value (PV) scores.	• Differences in sampling methods will be accounted for by standard errors of the linking parameters, which will involve sampling and measurement errors from both assessments.
Administration	• Administration timing (time of year)	• Differences in timing cannot be statistically accounted for, but both samples will be examined to assess impact of timing.	• Timing is not a huge problem for linking, as one would capture the differences in the linking.
Inclusion and accommodation	• Accommodation policy • Exclusion policy	• Differences in inclusion rates between states may affect linking results. AIR and ETS might want to exclude accommodated students for all approaches to determine what, if any, effect it has on the linking results.	• Differences in inclusion rates and accommodations are not a problem when linking.
Analysis and scaling	• Conditioning model • Treatment of not-reached items • Establishing trend	• CAL: In IRT estimation stage use univariate scale for NAEP mathematics and science and in conditioning stage use separate univariate models for NAEP mathematics and science and joint bivariate model for TIMSS mathematics and science. • PRO: Use univariate NAEP mathematics and science scores to estimate regression equation.	• Differences in analysis and scaling method details are not directly relevant to statistical moderation linking approach.
Reporting	• Benchmarks • Scale (score range, mean, SD)	• Reporting differences will not affect the linking.	• Reporting differences will not affect the linking.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.2. Differences in estimates of TIMSS scale score means for each validation state – mathematics

State	Actual TIMSS	Mean estimates using:			Projected error (predicted – actual)		
		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.

Table 6.3. Differences in estimates of TIMSS scale score means for each validation state – science

State	Actual TIMSS	Mean estimates using:			Projected error (predicted – actual)		
		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.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.4. Statistical significance of differences in estimates of TIMSS scale score means – mathematics

	Actual TIMSS		[A]Moderation				Projection				Calibration			
			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	6.06	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.

Table 6.5. Statistical significance of differences in estimates of TIMSS scale score means – science

	Actual TIMSS		[A]Moderation				Projection				Calibration			
			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	6.46	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.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.6. Differences in estimates of TIMSS scale score SDs for each validation state – mathematics

State	Actual TIMSS	Mean estimates using:			Projected error (predicted – actual)		
		MOD	PRO	CAL	MOD	PRO	CAL
Nation	76.04	76.04	77.38	77.43	0.00	1.34	1.39
9-MA	73.27	72.09	72.37	72.34	-1.18	-0.91	-0.93
8-MN	72.08	73.71	74.25	74.00	1.63	2.17	1.92
7-CO	76.12	75.24	74.85	75.08	-0.88	-1.28	-1.04
6-CT	83.95	74.22	75.11	75.22	-9.73	-8.84	-8.73
5-NC	80.18	75.17	76.52	76.50	-5.01	-3.66	-3.68
4-IN	70.75	68.78	70.90	70.16	-1.96	0.15	-0.58
3-FL	76.36	73.42	76.38	75.73	-2.94	0.02	-0.63
2-CA	80.63	82.76	83.24	84.34	2.14	2.62	3.71
1-AL	79.23	74.68	78.16	77.67	-4.54	-1.06	-1.56
				Root mean square error:	**4.25**	**3.44**	**3.53**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

Table 6.7. Differences in estimates of TIMSS scale score SDs for each validation state – science

State	Actual TIMSS	Mean estimates using:			Projected error (predicted – actual)		
		MOD	PRO	CAL	MOD	PRO	CAL
Nation	80.42	80.42	81.89	81.90	0.00	1.47	1.48
9-MA	81.04	79.64	81.03	81.32	-1.40	-0.01	0.28
8-MN	72.17	73.76	76.34	75.60	1.59	4.16	3.43
7-CO	77.55	75.91	77.63	77.49	-1.64	0.08	-0.06
6-CT	87.64	78.74	80.16	80.32	-8.90	-7.48	-7.32
5-IN	75.62	74.54	77.05	75.49	-1.08	1.43	-0.13
4-FL	85.07	80.53	81.68	82.01	-4.54	-3.38	-3.06
3-NC	81.71	78.47	80.25	80.82	-3.24	-1.46	-0.90
2-CA	84.25	87.63	88.36	88.49	3.37	4.10	4.24
1-AL	87.56	82.18	83.59	83.90	-5.38	-3.97	-3.66
				Root mean square error:	**4.20**	**3.67**	**3.45**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.8. Statistical significance of differences in estimates of TIMSS scale score SDs – mathematics

	Actual TIMSS		[A]Moderation				Projection				Calibration			
			Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	SD	SE	SD	SE	*t*	Sig.	SD	SE	*t*	Sig.	SD	SE	*t*	Sig.
Nation	76.04	1.59	76.04	0.37	0.00	1.000	77.38	0.30	0.83	0.407	77.43	0.34	0.86	0.391
9-MA	73.27	2.68	72.09	1.17	-0.40	0.686	72.37	1.62	-0.29	0.772	72.34	1.31	-0.31	0.755
8-MN	72.08	2.83	73.71	1.34	0.52	0.603	74.25	1.36	0.69	0.489	74.00	1.40	0.61	0.542
7-CO	76.12	2.18	75.24	1.51	-0.33	0.741	74.85	1.31	-0.50	0.616	75.08	1.23	-0.41	0.679
6-CT	83.95	2.86	**74.22**	**1.32**	**-3.08**	**0.002**	**75.11**	**1.41**	**-2.77**	**0.006**	**75.22**	**1.28**	**-2.78**	**0.005**
5-NC	80.18	4.12	75.17	1.45	-1.15	0.251	76.52	1.40	-0.84	0.401	76.50	1.63	-0.83	0.406
4-IN	70.75	1.86	68.78	1.35	-0.85	0.393	70.90	1.26	0.07	0.947	70.16	1.36	-0.25	0.800
3-FL	76.36	3.06	73.42	1.01	-0.91	0.361	76.38	1.20	0.01	0.995	75.73	1.10	-0.19	0.846
2-CA	80.63	2.82	82.76	1.68	0.65	0.515	83.24	1.33	0.84	0.401	84.34	1.54	1.16	0.248
1-AL	79.23	3.15	74.68	1.51	-1.30	0.194	78.16	1.51	-0.30	0.761	77.67	1.65	-0.44	0.662

[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 6.9. Statistical significance of differences in estimates of TIMSS scale score SDs – science

	Actual TIMSS		[A]Moderation				Projection				Calibration			
			Projected		Error		Projected		[B]Error		Projected		[B]Error	
State	SD	SE	SD	SE	*t*	Sig.	SD	SE	*t*	Sig.	SD	SE	*t*	Sig.
Nation	80.42	1.43	80.42	0.39	0.00	1.000	81.89	0.39	0.99	0.321	81.90	0.40	1.00	0.319
9-MA	81.04	2.41	79.64	2.01	-0.45	0.656	81.03	2.22	0.00	0.997	81.32	1.94	0.09	0.928
8-MN	72.17	2.70	73.76	1.62	0.50	0.614	76.34	1.58	1.33	0.184	75.60	1.73	1.07	0.285
7-CO	77.55	2.10	75.91	1.59	-0.62	0.534	77.63	1.98	0.03	0.978	77.49	1.70	-0.02	0.983
6-CT	87.64	2.91	**78.74**	**1.44**	**-2.74**	**0.006**	**80.16**	**1.70**	**-2.22**	**0.026**	**80.32**	**1.75**	**-2.16**	**0.031**
5-IN	75.62	1.98	74.54	1.74	-0.41	0.682	77.05	1.80	0.54	0.591	75.49	1.70	-0.05	0.960
4-FL	85.07	3.10	80.53	1.75	-1.27	0.203	81.68	1.83	-0.94	0.346	82.01	1.57	-0.88	0.379
3-NC	81.71	3.40	78.47	1.91	-0.83	0.406	80.25	1.94	-0.37	0.709	80.82	2.05	-0.23	0.822
2-CA	84.25	2.49	87.63	1.86	1.09	0.278	88.36	1.74	1.35	0.176	88.49	1.97	1.34	0.182
1-AL	87.56	2.68	82.18	1.79	-1.67	0.096	83.59	1.96	-1.19	0.233	83.90	2.01	-1.09	0.275

[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.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.10. Statistical significance of differences in estimates of percent above low TIMSS benchmark level cutoffs

			[A]Moderation			Projection			Calibration		
Mathematics	Actual TIMSS		Projected		[B]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**

			[A]Moderation			Projection			Calibration		
Science	Actual TIMSS		Projected		[B]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 means are statistically significant from the actual means.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.11. Statistical significance of differences in estimates of percent above intermediate TIMSS benchmark level cutoffs

			[A]Moderation			Projection			Calibration		
Mathematics	Actual TIMSS		Projected		[B]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**

			[A]Moderation			Projection			Calibration		
Science	Actual TIMSS		Projected		[B]Error	Projected		[B]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 means are statistically significant from the actual means.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.12. Statistical significance of differences in estimates of percent above high TIMSS benchmark level cutoffs

			[A]Moderation			Projection			Calibration		
Mathematics	Actual TIMSS		Projected		[B]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

			[A]Moderation			Projection			Calibration		
Science	Actual TIMSS		Projected		[B]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 means are statistically significant from the actual means.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.13. Statistical significance of differences in estimates of percent above advanced TIMSS benchmark level cutoffs

			[A]Moderation			Projection			Calibration		
Mathematics	Actual TIMSS		Projected		[B]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

			[A]Moderation			Projection			Calibration		
Science	Actual TIMSS		Projected		[B]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 means are statistically significant from the actual means.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.14. Statistical significance of differences in estimates of TIMSS scale score means for male and female students – mathematics

Male students			[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

Female students			[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.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.15. Statistical significance of differences in estimates of TIMSS scale score means for male and female students – science

Male students			[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

Female students			[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.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.16. Statistical significance of differences in estimates of TIMSS scale score means for White students – 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	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 6.17. Statistical significance of differences in estimates of TIMSS scale score means for African-American students – 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	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.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.18. Statistical significance of differences in estimates of TIMSS scale score means for Hispanic students – 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	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 6.19. Statistical significance of differences in estimates of TIMSS scale score means for Asian students – 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	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.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.20. Statistical significance of differences in estimates of TIMSS scale score means for White students – 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	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 6.21. Statistical significance of differences in estimates of TIMSS scale score means for African-American students – 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	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.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.22. Statistical significance of differences in estimates of TIMSS scale score means for Hispanic students – 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	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 6.23. Statistical significance of differences in estimates of TIMSS scale score means for Asian students – 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	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.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.24. NAEP and TIMSS exclusion and accommodation rates – mathematics

	2011 NAEP/TIMSS mathematics: Exclusion and accommodation percentages				
	NAEP			TIMSS	Diff. (T-N)
State	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 6.25. NAEP and TIMSS exclusion and accommodation rates – science

	2011 NAEP/TIMSS science: Exclusion and accommodation percentages				
	NAEP			TIMSS	Diff. (T-N)
State	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.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.26. 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
Mathematics	MOD	.39	-.72
Mathematics	PRO	.37	-.74
Mathematics	CAL	.39	-.72
Science	MOD	.45	-.79
Science	PRO	.38	-.81
Science	CAL	.48	-.78

NOTE: Estimation errors were computed as the predicted TIMSS mean minus the observed TIMSS mean. N-T = NAEP minus TIMSS. MOD = Moderation; PRO = Projection, CAL = Calibration.

Table 6.27. Differences in estimates of TIMSS scale score means: NoSDE – mathematics

State	Actual TIMSS	Mean estimates using:			Projected error (predicted – actual)		
		MOD	PRO	CAL	MOD	PRO	CAL
Nation	506.89	519.85	519.58	520.39	12.97	12.70	13.50
9-MA	560.58	556.16	553.07	556.70	-4.42	-7.51	-3.88
8-MN	544.73	545.71	544.02	546.51	0.98	-0.70	1.78
7-CO	517.79	541.22	539.55	541.49	23.42	21.75	23.69
6-CT	517.62	527.05	526.57	527.87	9.43	8.95	10.25
5-NC	536.90	526.99	525.88	527.75	-9.91	-11.02	-9.15
4-IN	521.51	522.15	521.95	523.17	0.65	0.45	1.66
3-FL	513.30	508.83	508.49	508.86	-4.48	-4.82	-4.44
2-CA	492.62	508.00	508.22	508.56	15.38	15.60	15.94
1-AL	465.93	488.98	489.99	488.86	23.05	24.06	22.94
				Root mean square error:	**13.10**	**13.24**	**13.21**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.28. Differences in estimates of TIMSS scale score means: NoSDE – science

State	Actual TIMSS	Mean estimates using:			Projected error (predicted – actual)		
		MOD	PRO	CAL	MOD	PRO	CAL
Nation	522.19	536.26	535.39	536.45	14.07	13.20	14.26
9-MA	566.78	561.48	558.63	562.44	-5.30	-8.14	-4.34
8-MN	553.27	559.43	556.62	559.80	6.15	3.34	6.52
7-CO	541.95	560.22	557.32	561.06	18.27	15.37	19.11
6-CT	531.60	542.68	541.83	542.98	11.08	10.23	11.38
5-IN	532.80	537.88	536.31	537.86	5.08	3.51	5.05
4-FL	529.89	530.07	529.98	530.68	0.18	0.09	0.80
3-NC	531.53	526.46	526.81	526.20	-5.08	-4.72	-5.34
2-CA	498.52	521.92	521.81	521.74	23.40	23.30	23.22
1-AL	485.37	507.94	510.22	507.66	22.57	24.85	22.29
				Root mean square error:	**13.46**	**13.39**	**13.53**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

Table 6.29. Differences in estimates of TIMSS scale score means: NoACC – mathematics

State	Actual TIMSS	Mean estimates using:			Projected error (predicted – actual)		
		MOD	PRO	CAL	MOD	PRO	CAL
Nation	506.89	515.95	515.90	516.41	9.06	9.01	9.52
9-MA	560.58	553.33	550.43	553.86	-7.25	-10.15	-6.72
8-MN	544.73	541.49	540.14	542.37	-3.23	-4.58	-2.35
7-CO	517.79	535.82	534.51	536.06	18.02	16.71	18.27
6-CT	517.62	525.31	524.93	526.15	7.69	7.31	8.53
5-NC	536.90	525.21	524.25	525.98	-11.68	-12.64	-10.91
4-IN	521.51	521.39	521.36	522.45	-0.11	-0.14	0.95
3-FL	513.30	508.71	508.32	508.74	-4.60	-4.99	-4.57
2-CA	492.62	493.77	494.79	493.97	1.15	2.17	1.35
1-AL	465.93	482.76	483.94	482.35	16.83	18.01	16.42
				Root mean square error:	**9.94**	**10.39**	**9.83**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.30. Differences in estimates of TIMSS scale score means: NoACC – science

State	Actual TIMSS	Mean estimates using:			Projected error (predicted – actual)		
		MOD	PRO	CAL	MOD	PRO	CAL
Nation	522.19	531.77	531.26	531.97	9.59	9.07	9.79
9-MA	566.78	557.83	555.27	558.74	-8.95	-11.51	-8.03
8-MN	553.27	554.32	551.92	554.76	1.05	-1.35	1.49
7-CO	541.95	555.30	552.85	556.03	13.35	10.90	14.08
6-CT	531.60	541.88	541.08	542.02	10.28	9.48	10.42
5-NC	532.80	536.93	535.47	536.91	4.12	2.66	4.10
4-IN	529.89	529.81	529.72	530.45	-0.08	-0.17	0.56
3-FL	531.53	524.26	524.79	523.89	-7.27	-6.75	-7.65
2-CA	498.52	506.37	507.56	506.49	7.85	9.04	7.97
1-AL	485.37	502.00	504.66	501.66	16.63	19.29	16.28
				Root mean square error:	**9.27**	**9.72**	**9.03**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

Table 6.31. Differences in estimates of TIMSS scale score means: AccRW – mathematics

State	Actual TIMSS	Mean estimates using:			Projected error (predicted – actual)		
		MOD	PRO	CAL	MOD	PRO	CAL
Nation	506.89	508.28	508.62	508.55	1.39	1.73	1.67
9-MA	560.58	549.44	546.83	549.92	-11.14	-13.75	-10.67
8-MN	544.73	539.06	537.84	539.89	-5.67	-6.89	-4.83
7-CO	517.79	532.30	531.30	532.60	14.51	13.50	14.81
6-CT	517.62	519.46	519.31	520.11	1.83	1.68	2.49
5-NC	536.90	516.54	516.18	517.26	-20.36	-20.72	-19.64
4-IN	521.51	518.12	518.28	519.12	-3.39	-3.23	-2.39
3-FL	513.30	504.43	504.20	504.34	-8.88	-9.10	-8.96
2-CA	492.62	488.92	490.22	489.00	-3.70	-2.40	-3.62
1-AL	465.93	478.53	479.83	477.96	12.60	13.91	12.03
				Root mean square error:	**10.79**	**11.27**	**10.50**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.32. Differences in estimates of TIMSS scale score means: AccRW – science

State	Actual TIMSS	Mean estimates using:			Projected error (predicted – actual)		
		MOD	PRO	CAL	MOD	PRO	CAL
Nation	522.19	523.69	523.81	523.81	1.51	1.63	1.62
9-MA	566.78	554.10	551.86	554.95	-12.68	-14.91	-11.82
8-MN	553.27	551.83	549.60	552.24	-1.45	-3.67	-1.04
7-CO	541.95	551.95	549.79	552.66	10.00	7.85	10.71
6-CT	531.60	535.50	535.17	535.67	3.90	3.57	4.07
5-NC	532.80	532.89	531.80	532.79	0.09	-1.00	-0.01
4-IN	529.89	524.71	525.02	525.20	-5.18	-4.87	-4.69
3-FL	531.53	516.73	517.82	516.15	-14.80	-13.71	-15.39
2-CA	498.52	502.12	503.64	502.25	3.60	5.13	3.73
1-AL	485.37	497.80	500.77	497.26	12.43	15.40	11.89
				Root mean square error:	**8.77**	**9.35**	**8.73**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.33. Impact of differential accommodation reweighting: AccDRW – mathematics

Accommodation	N_Uwgt	N_Wgt	Mean	Overall percent	
				Before	After
1. Calculator	197	4,076.7	237.1	0.1	0
2. Spanish assistance	313	7,532.5	237.1	0.2	0
3. Reading accommodation	8,240	164,757.1	242.7	4.7	0.3
4. Braille	9	136.6	237.4	0	0
5. Testing accommodation	8,547	170,177.8	248.5	4.9	4.9
6. Others	314	8,207.9	251.8	0.2	0.2
7. No accommodation	146,777	3,060,487.7	287	87.4	87.4
Total Included				97.5	92.8
Total Excluded				2.5	7.2
Overall NAEP mean				**282.7**	**284.8**
NAEP mean from uniform reweighting					**284.8**

Table 6.34. Impact of differential accommodation reweighting: AccDRW – science

Accommodation	N_Uwgt	N_Wgt	Mean	Overall percent	
				Before	After
1. Calculator	11	176.0	99.6	0	0
2. Spanish assistance	200	7,747.8	100.2	0.2	0
3. Reading accommodation	7,077	192,580.0	117.7	5.5	0.1
4. Braille	2	42.0	153	0	0
5. Testing accommodation	5,946	177,620.7	120	5.1	5.1
6. Others	245	10,167.7	115.5	0.3	0.3
7. No accommodation	106,166	3,060,181.4	154.9	87.3	87.3
Total Included				98.4	92.8
Total Excluded				2.6	7.2
Overall NAEP mean				**150.7**	**152.8**
NAEP mean from uniform reweighting					**152.6**

Table 6.35. AccADJ coefficients – mathematics and science

Subject	MOD	PRO	CAL
Mathematics	2.65	2.59	2.80
Science	2.21	2.16	2.39

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.36. Differences in estimates of TIMSS scale score means: AccADJ – mathematics

State	Actual TIMSS	Mean estimates using:			Projected error (predicted – actual)		
		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	554.07	552.96	554.10	-6.51	-7.62	-6.48
8-MN	544.73	529.80	528.77	530.57	-14.93	-15.96	-14.16
7-CO	517.79	526.52	526.12	526.91	8.73	8.32	9.11
6-CT	517.62	522.68	523.05	523.07	5.06	5.43	5.44
5-NC	536.90	521.41	521.60	521.96	-15.49	-15.30	-14.94
4-IN	521.51	518.22	519.11	518.94	-3.29	-2.39	-2.57
3-FL	513.30	513.72	514.74	513.04	0.42	1.43	-0.27
2-CA	492.62	480.13	481.27	480.27	-12.49	-11.35	-12.35
1-AL	465.93	462.14	462.55	461.93	-3.79	-3.38	-4.00
				Root mean square error:	**9.93**	**9.96**	**9.71**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

Table 6.37. Differences in estimates of TIMSS scale score means: AccADJ – science

State	Actual TIMSS	Mean estimates using:			Projected error (predicted – actual)		
		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	558.53	557.92	558.97	-8.25	-8.85	-7.80
8-MN	553.27	541.18	539.00	541.65	-12.09	-14.28	-11.63
7-CO	541.95	544.50	542.90	545.21	2.55	0.96	3.26
6-CT	531.60	535.68	536.01	535.76	4.08	4.41	4.16
5-NC	532.80	532.51	532.35	532.17	-0.30	-0.45	-0.63
4-IN	529.89	529.29	531.27	529.25	-0.60	1.38	-0.64
3-FL	531.53	518.54	520.02	517.83	-12.99	-11.51	-13.71
2-CA	498.52	491.86	493.19	492.15	-6.66	-5.33	-6.37
1-AL	485.37	482.80	484.65	482.58	-2.57	-0.72	-2.79
				Root mean square error:	**7.56**	**7.64**	**7.59**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.38. Differences in estimates of TIMSS scale score means: RaceADJ – mathematics

State	Actual TIMSS	Reweighted actual TIMSS	Original mean estimates using:			Projected error (projected – reweighted actual)		
			MOD	PRO	CAL	MOD	PRO	CAL
Nation	506.89	506.22	506.89	507.30	507.14	0.66	1.08	0.91
9-MA	560.58	557.05	540.00	538.10	540.34	-17.05	-18.95	-16.71
8-MN	544.73	542.25	532.52	531.63	533.22	-9.73	-10.62	-9.03
7-CO	517.79	519.08	525.80	525.35	526.20	6.72	6.27	7.12
6-CT	517.62	514.56	515.85	515.83	516.39	1.29	1.28	1.83
5-NC	536.90	535.98	514.31	514.11	515.02	-21.67	-21.87	-20.95
4-IN	521.51	517.76	511.66	512.19	512.53	-6.10	-5.57	-5.23
3-FL	513.30	512.47	496.63	496.69	496.34	-15.84	-15.78	-16.14
2-CA	492.62	492.75	486.00	487.47	486.01	-6.74	-5.27	-6.73
1-AL	465.93	467.20	478.30	479.61	477.72	11.09	12.41	10.51
					Root mean square error:	**12.29**	**12.71**	**12.01**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

Table 6.39. Differences in estimates of TIMSS scale score means: RaceADJ – science

State	Actual TIMSS	Reweighted actual TIMSS	Original mean estimates using:			Projected error (projected – reweighted actual)		
			MOD	PRO	CAL	MOD	PRO	CAL
Nation	522.19	520.92	522.19	522.43	522.29	1.27	1.51	1.37
9-MA	566.78	562.09	546.63	545.06	547.37	-15.46	-17.04	-14.72
8-MN	553.27	550.91	545.86	544.05	546.21	-5.05	-6.86	-4.70
7-CO	541.95	542.52	545.12	543.57	545.81	2.60	1.05	3.29
6-CT	531.60	528.39	531.34	531.32	531.53	2.95	2.93	3.13
5-NC	532.80	527.47	527.35	526.77	527.14	-0.13	-0.70	-0.33
4-IN	529.89	529.94	516.71	517.66	516.98	-13.23	-12.28	-12.96
3-FL	531.53	530.47	515.16	516.37	514.53	-15.31	-14.10	-15.94
2-CA	498.52	497.95	498.12	499.96	498.25	0.16	2.00	0.30
1-AL	485.37	486.98	497.10	500.11	496.52	10.12	13.13	9.54
					Root mean square error:	**9.38**	**9.85**	**9.27**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.40. Differences in estimates of TIMSS scale score means: RaceAccADJ – mathematics

State	Actual TIMSS	Reweighted actual TIMSS	AccADJ mean estimates using:			Projected error (projected – reweighted actual)		
			MOD	PRO	CAL	MOD	PRO	CAL
Nation	506.89	506.22	506.89	507.30	507.14	0.66	1.08	0.91
9-MA	560.58	557.05	554.07	552.96	554.10	-2.98	-4.09	-2.95
8-MN	544.73	542.25	529.80	528.77	530.57	-12.45	-13.48	-11.68
7-CO	517.79	519.08	526.52	526.12	526.91	7.45	7.04	7.83
6-CT	517.62	514.56	522.68	523.05	523.07	8.13	8.49	8.51
5-NC	536.90	535.98	521.41	521.60	521.96	-14.57	-14.38	-14.02
4-IN	521.51	517.76	518.22	519.11	518.94	0.45	1.35	1.17
3-FL	513.30	512.47	513.72	514.74	513.04	1.25	2.26	0.56
2-CA	492.62	492.75	480.13	481.27	480.27	-12.62	-11.47	-12.47
1-AL	465.93	467.20	462.14	462.55	461.93	-5.06	-4.65	-5.27
					Root mean square error:	**9.25**	**9.27**	**9.09**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

Table 6.41. Differences in estimates of TIMSS scale score means: RaceAccADJ – science

State	Actual TIMSS	Reweighted actual TIMSS	AccADJ mean estimates using:			Projected error (projected – reweighted actual)		
			MOD	PRO	CAL	MOD	PRO	CAL
Nation	522.19	520.92	522.19	522.43	522.29	1.27	1.51	1.37
9-MA	566.78	562.09	558.53	557.92	558.97	-3.57	-4.17	-3.12
8-MN	553.27	550.91	541.18	539.00	541.65	-9.73	-11.92	-9.26
7-CO	541.95	542.52	544.50	542.90	545.21	1.98	0.38	2.69
6-CT	531.60	528.39	535.68	536.01	535.76	7.29	7.62	7.37
5-NC	532.80	527.47	532.51	532.35	532.17	5.03	4.88	4.70
4-IN	529.89	529.94	529.29	531.27	529.25	-0.65	1.33	-0.69
3-FL	531.53	530.47	518.54	520.02	517.83	-11.93	-10.45	-12.64
2-CA	498.52	497.95	491.86	493.19	492.15	-6.09	-4.77	-5.81
1-AL	485.37	486.98	482.80	484.65	482.58	-4.18	-2.33	-4.40
					Root mean square error:	**6.96**	**6.89**	**7.00**

NOTE: MOD = Moderation; PRO = Projection; CAL = Calibration.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.42. Tests for state by item-type interaction for mathematics and science

Mathematics

Source	DF	Type III SS	Mean square	F Value	Pr > F
Assess	1	0.58	0.58	17.88	<.0001
State	8	7.28	0.91	27.93	<.0001
Itype	1	4.71	4.71	144.69	<.0001
State*Itype	8	0.13	0.02	0.5	0.8545
Assess*Itype	1	0.01	0.01	0.38	0.5376
Assess*State	8	0.32	0.04	1.24	0.2706

Science

Source	DF	Type III SS	Mean square	F Value	Pr > F
Assess	1	4.42	4.42	173.89	<.0001
State	8	2.75	0.34	13.56	<.0001
Itype	1	8.41	8.41	331.3	<.0001
State*Itype	8	0.08	0.01	0.41	0.9135
Assess*Itype	1	0.92	0.92	36.32	<.0001
Assess*State	8	0.2	0.02	0.97	0.4572

Table 6.43. Predicted state mean estimates for the statistical moderation using AccADJ – mathematics

		Mean estimates using:		Projected error (predicted – actual)	
State	Actual TIMSS	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.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.44. Predicted state mean estimates for the statistical moderation using AccADJ – science

State	Actual TIMSS	Mean estimates using:		Projected error (predicted – actual)	
		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.

Table 6.45. Estimation of model error variance 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 using df=8.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.46. Estimation of model error variance for the AccADJ statistical moderation linkage – science

	Total error		Variances in MOD estimates				Variances in TIMSS		
State	Error	Error^2	Total	Sample	Meas.	Link.	Total	Samp.	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 =	**13.58**	**(Total error - NAEP estimate variance - TIMSS variance)**						

NOTE: MOD = Moderation; Meas. = Measurement; Link. = Linking; MSE = Mean square error using df=8.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.47. Statistical significance of differences in estimates of TIMSS scale score means for unadjusted means (without model error added to SEs) and adjusted means (with model error added to SEs) for the statistical moderation linkage – mathematics

	Actual TIMSS		Unadjusted with no model error				AccADJ with model error			
			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

† Not applicable.
NOTE: Bold font indicates predicted means are statistically significant from the actual means.

Table 6.48. Statistical significance of differences in estimates of TIMSS scale score means for unadjusted means (without model error added to SEs) and adjusted means (with model error added to SEs) for the statistical moderation linkage – science

	Actual TIMSS		Unadjusted with no model error				AccADJ with model error			
			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	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	6.46	497.10	4.25	-11.73	1.52	482.80	5.59	2.57	-0.30

† Not applicable.
NOTE: Bold font indicates predicted means are statistically significant from the actual means.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.49. MSEs for unadjusted (without model error) and adjusted (with model error) percent-above-cut estimates for the statistical moderation linkage

	Mathematics			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

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

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.50. 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

			Unadj. (without model error)			AccADJ_Normal (with model error)			AccADJ_Direct (with model error)		
Mathematics	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	-1.10	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	3.20	**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	-2.10	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

			Unadj. (without model error)			AccADJ_Normal (with model error)			AccADJ_Direct (with model error)		
Science	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	0.70	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	1.00	-0.82	95.04	0.68	-0.07	95.03	**1.00**	-0.08
4-FL	93.48	1.49	91.28	1.39	-2.20	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	-1.40	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: P-A = Predicted minus Actual. 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. Unadj. - No adjustment for % accommodated. AccADJ_Normal - Using adjustment to the mean and the normal approximation. AccADJ_Direct - Direct adjustment using a separate regression equation for each cutoff.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.51. 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

			Unadj. (without model error)			AccADJ_Normal (with model error)			AccADJ_Direct (with model error)		
Mathematics	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	70.40	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	62.20	2.50	-5.40	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	3.20	53.60	3.15	7.85	44.99	4.49	-0.77	47.71	3.47	1.96

			Unadj. (without model error)			AccADJ_Normal (with model error)			AccADJ_Direct (with model error)		
Science	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	2.00	77.66	2.40	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	-2.70	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	2.70	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: P-A = Predicted minus Actual. 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. Unadj. - No adjustment for % accommodated. AccADJ_Normal - Using adjustment to the mean and the normal approximation. AccADJ_Direct - Direct adjustment using a separate regression equation for each cutoff.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.52. 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

			Unadj. (without model error)			AccADJ_Normal (with model error)			AccADJ_Direct (with model error)		
Mathematics	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	38.70	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	-3.20	36.74	4.19	0.22	36.40	4.63	-0.12
5-NC	44.24	3.60	**32.40**	**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

			Unadj. (without model error)			AccADJ_Normal (with model error)			AccADJ_Direct (with model error)		
Science	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	3.20	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: P-A = Predicted minus Actual. 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. Unadj. - No adjustment for % accommodated. AccADJ_Normal - Using adjustment to the mean and the normal approximation. AccADJ_Direct - Direct adjustment using a separate regression equation for each cutoff.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Table 6.53. 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

			Unadj. (without model error)			AccADJ_Normal (with model error)			AccADJ_Direct (with model error)		
Mathematics	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	1.60	-3.25	9.21	1.85	-3.87	9.08	2.89	-4.00
7-CO	7.70	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	4.40	1.06	-0.41	3.78	0.83	-1.04	2.78	2.63	-2.04
1-AL	2.10	0.77	1.91	0.67	-0.19	1.10	0.33	-1.00	**-2.56**	2.50	-4.66

			Unadj. (without model error)			AccADJ_Normal (with model error)			AccADJ_Direct (with model error)		
Science	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	-3.20	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: P-A = Predicted minus Actual. Bold font indicates predicted estimates are statistically significant from the actual estimates. Unadj. - No adjustment for % accommodated. AccADJ_Normal - Using adjustment to the mean and the normal approximation. AccADJ_Direct - Direct adjustment using a separate regression equation for each cutoff.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Mathematics and Science Assessments; and International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

Chapter 7: Summary and Conclusions

With the 2011 NAEP-TIMSS linking study, NCES produced TIMSS scores for the U.S. states that participated in the 2011 NAEP mathematics and science assessment of students at grade 8. Three linking methods (*calibration*, *statistical projection*, and *statistical moderation*) were used in this study to examine how best to use states' NAEP scores to predict performance on TIMSS.

Study of Calibration Linking

In this study, average TIMSS scores and percentages of students reaching each of the TIMSS international benchmark levels were established based on countries that participated in TIMSS. This method of linking resulted in predicted state TIMSS results that were of similar magnitude to the moderation and projection linking methods. It was also observed that there were sizeable discrepancies between predicted and actual state results for more than half of the validation states. After adjustments were made, some impact was detected but these corrections were considered to be ad hoc and experimental in nature and did not fully account for many other sources of bias. However, it was suggested that such analyses are insightful to assess what level of prediction bias that can be potentially reduced by taking into account the accommodations and exclusion differences between the two assessment programs.

Study of Statistical Projection

This linking study was designed so that there were two braided-booklet samples, one in each NAEP and TIMSS assessment windows. The projection functions derived separately from the individual braided-booklet samples were used to predict state TIMSS results from states' NAEP scores. Another focus was whether the braided-booklet samples administered in the two assessment windows were both necessary for carrying out a projection type of linkage. The empirical evidence suggests that the predicted state TIMSS results (before linear adjustments) were relatively more accurate when based on the projection function derived from the NAEP window braided-booklet sample as compared to the projection function derived from the TIMSS window braided-booklet sample. After linear adjustments were applied, the two sets of predicted results were comparable and the adjustments generally reduced the differences between the predicted and actual state TIMSS results. However, the relationship between NAEP and TIMSS might not be the same within the states as it is in the country as a whole. Additionally, the braided-booklet sample design does not support the estimation of state-specific projection functions.

Study of Statistical Moderation

Statistical moderation employed the grade 8 U.S. national NAEP and TIMSS samples that were validated in the nine states in which NAEP and TIMSS were administered: Alabama, California, Colorado, Connecticut, Florida, Indiana, Massachusetts, Minnesota, and North Carolina. After the statistical link was established between NAEP and TIMSS, the link was applied to the remaining states in the study in order to estimate TIMSS performance. After accommodation adjustments were applied to the nine states, results indicated that the accommodation adjustments in both mathematics and science should improve projections.

Overall Findings from the Linking Studies

Based on the results from the linking studies, it was found that it is possible to express data from NAEP in the metric of TIMSS. By expressing both assessments in the same metric, two major results were realized: (1) the estimation of the state-TIMSS mean that each state might have obtained had that state actually taken TIMSS, and (2) the comparison of state results to international results. In addition, all three linking methods, where linkages were performed on the U.S. national population, yielded comparable projected national TIMSS results by student groups as well as projected state TIMSS results. The consistency of the predicted results among the linking methods provides evidence that the two assessments are strongly related and supports using statistical moderation linking, which does not require braided-booklet samples.

However, when the results were applied to separate validation states, differences were observed between the linkage-based predicted state TIMSS results and their actual reported TIMSS scores. A two-stage adjustment procedure to the linkage-based state results—first by state NAEP accommodation rates, then by a state-level projection linkage—was applied to reduce such observed differences. Therefore, the projected state TIMSS results presented in this report were estimated from the moderation linking and the two-stage adjustment procedure.

Recommendations

Recommendations from the evaluation of the three linking methods are as follows:

1. Use estimates from the statistical moderation linkages.
2. Use the adjustment based on percent of students accommodated to improve linkage-based mean estimates.
3. Include an estimate of model error in standard error estimates and confidence bounds for linkage-based estimates.
4. Use normal approximations to adjust estimates of percent above cut points for consistency with the adjustment based on the percent of students accommodated for state mean estimates.
5. Include confidence bounds in all reporting.