IES Institute of Education Sciences

A Publication of the National Center for Education Statistics

NCES 2020-038 U.S. DEPARTMENT OF EDUCATION

Estimating Student Achievement at the Topic Level in TIMSS Using IRT-Based Domain Scoring

Methodological Report

Estimating Student Achievement at the Topic Level in TIMSS Using IRT-Based Domain Scoring

Methodological Report

FEBRUARY 2020

Stephen Provasnik
Enis Dogan
National Center for Education Statistics

Ebru Erberber
Xiaying Zheng
American Institutes for Research

NCES 2020-038

U.S. DEPARTMENT OF EDUCATION

IES Institute of Education Sciences

U.S. Department of Education
Betsy DeVos
Secretary

Institute of Education Sciences
Mark Schneider
Director

National Center for Education Statistics
James L. Woodworth
Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity responsible for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high-priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high-quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public. Unless specifically noted, all information contained herein is in the public domain.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to

NCES, IES, U.S. Department of Education
Potomac Center Plaza (PCP)
550 12th Street SW
Washington, DC 20202

February 2020

The NCES Home Page address is https://nces.ed.gov.
The NCES Publications and Products address is https://nces.ed.gov/pubsearch.

This publication is only available online. To download, view, and print the report as a PDF file, go to the NCES Publications and Products address shown above.

This report was prepared under Contract No. ED-IES-12-D-0002 with the American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

Suggested Citation
Provasnik, S., Dogan, E., Erberber, E., and Zheng, X. (2020). *Estimating Student Achievement at the Topic Level in TIMSS Using IRT-Based Domain Scoring: Methodological Report* (NCES 2020-038). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved [date] from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2020038.

Content Contacts
Stephen Provasnik
(202) 245-6442
stephen.provasnik@ed.gov

Enis Dogan
(202) 245-6970
enis.dogan@ed.gov

Contents

List of Tables

List of Figures

List of Exhibits

Section 1: Introduction

What is the value of reporting results below the content area level?

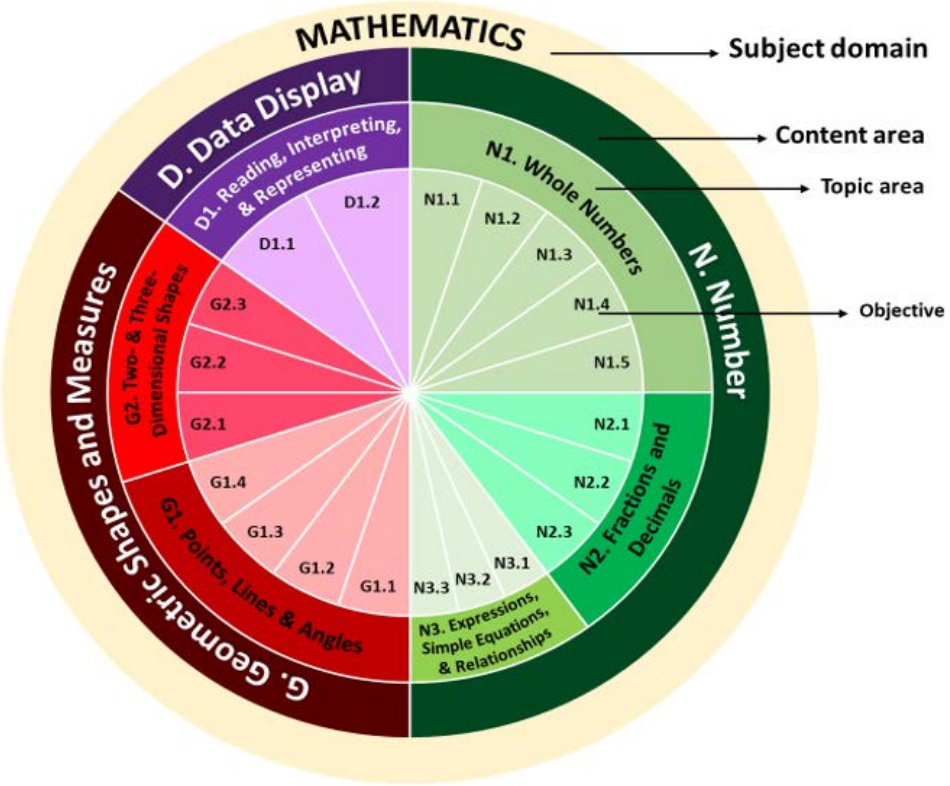
Large-scale assessment programs, such as the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS), employ item response theory (IRT) and marginal estimation methods to estimate student proficiency in specific subjects such as mathematics, science, or reading. Each of these subjects is defined by a framework that specifies the subject's content areas and, within each content area, the topics that are covered. In TIMSS mathematics, the framework specifies three content areas at grade 4 broken down into six topics, as illustrated in exhibit 1. TIMSS reports on student proficiency in a subject in two ways: with overall average scale scores and with the percentage of students who reach defined performance benchmarks (sometimes referred to as achievement levels in other large-scale assessments). TIMSS also reports on student performance for content areas within subjects. However, no estimate of student achievement is reported for any level below the content area.

The purpose of this report is to introduce a method for analyzing student achievement at the topic level using data from TIMSS assessments. Reporting topic-level scores in large-scale assessments, such as TIMSS, would be tremendously valuable from both a policy and pedagogical perspective. Topic-level scores could allow one to pinpoint areas in which students are excelling or struggling more specifically than is possible when one looks at just content area scores. Analysis of topic-level scores could reveal the strengths and weaknesses of students in relation to the intended and actualized curricula.[1] Trends in topic-level scores across years and between countries could also help more fully realize the promise of monitoring system-level performance in an "actionable" way. In other words, instead of learning, for example, that U.S. fourth-graders consistently lag behind Korean students in mathematics, or even that they are weaker in geometry compared to their international peers, topic-level scores would allow one to identify the particular topic(s) within geometry in which U.S. students are comparatively weaker and stronger.

For countries with a single national curriculum that has remained the same or comparable, such comparisons would open up the possibility of examining trends in student performance in relation to the curriculum over time. Comparisons between all participating countries could also become more meaningful because instead of researchers comparing how mathematics is taught in general, researchers could focus their efforts on comparing specific topic areas of mathematics instruction—for example, looking at how "fractions and decimals" or "patterns and relationships" are taught—with data on student performance in that specific topic to ground their analysis. Moreover, although these examples and this report have been based on TIMSS, the method introduced here for estimating student proficiency at a "topic" level could be applied to other large-scale trend assessments that are curricular-based, such as PIRLS or the National Assessment of Educational Progress (NAEP).

[1] TIMSS collects and reports contextual information about both the intended curriculum at the topic level and which topics teachers report that students have been taught.

Exhibit 1. Illustration of hierarchical structure for TIMSS 2015 grade 4 mathematics framework that specifies the subject's content areas and, for each content area, the topics that are covered



A new approach for estimating student achievement at the topic level

Researchers have long recognized the value of estimating proficiency at the topic level but have not had a reliable or viable method to do so for large-scale assessments. Among potential methods for estimating student proficiency at the topic level, the simplest approach is to calculate the proportion of items correctly answered in each topic. Using this approach, one can divide the number of items per topic that students answered correctly by the total number of items per topic, and thereby determine the percentage of correct responses per topic. While viable when dealing with a very large number of items, this simple method lacks reliability at the topic level for large-scale assessments because the number of items measuring a topic is relatively small. As Haladyna and Roid (1983) have noted, computing an observed percentage correct score for a test with fewer than 20 items can lead to unreasonable estimates. In addition, with the simple number/proportion correct method, the topic-level scores would need to be computed one test form (or "booklet") at a time. Most large-scale assessments feature multiple forms, and these are not equated on the number/proportion correct metric. Therefore, observed performance in a topic on a given form would not necessarily reflect what the performance would be on a larger and more representative set of items.

Given the limitations of the percentage correct scoring approach, IRT is the next logical method to consider. In IRT-based "domain scoring" [2] (Skorupski 2008; Bock, Thissen, and Zimowski 1997; Pommerich 2006), the probability of a correct answer (or weighted sum of probability of correct and partially correct answers in the case of polytomous items, which are items that have more than two possible scores) on a set of items is estimated based on the student's "ability" and item parameter estimates. This probability is known as model-based item scores (also known as expected item scores in the IRT literature). Summing these probabilities across a "domain" of interest provides a domain score. When expressed as a percentage, the domain score can be thought of as "the index of the proportion of the domain mastered" (Bock, Thissen, and Zimowski 1997, p. 197).

The meaningfulness of a domain score evidently depends on how large and representative of the domain the items in the given domain are. In the case of TIMSS, in any given assessment year, there are only between 12 and 30 items per topic. Furthermore, not all students get all the items in each topic area. Because of booklet spiraling, students may get only a few items for a given topic in the particular test booklet they receive. Thus, with only a very small pool of items completed by each student in each topic in each administration of a large-scale assessment, estimating student performance at the topic level with a domain score approach might not yield reliable and meaningful scores. The tacit assumption behind this conclusion, however, is that *topic-level performance has to be calculated within cross-sectional data from a single administration*. This is not correct. One of the strengths of IRT is that it allows estimation of the probability of a correct answer on an item for a student with an ability estimate, even if the test form the student took did not include the given item, as long as the parameter estimates of the item are on the same metric as the items that made up the form that the student did take (Lord 1980). Therefore, domain scores at the topic level can be estimated using the larger pool of items accumulated across years and booklets. In other words, just as IRT allows item information gathered from across all administered booklets within a single administration to be used to estimate student performance on items that were not in a student's test form, IRT allows estimation of student performance on items that were not administered in the year they received the assessment.

Topic-level scores have never been reported in TIMSS because there have been too few items at the topic level in any single administration to yield meaningful scores. However, the total number of TIMSS items in each topic has increased with each administration of TIMSS, that is, every 4 years. This study's new approach takes advantage of this now extant larger item pool (with its deeper breadth of topic-level items). In this study, probabilities of a correct answer are computed for the U.S. fourth-grade sample that participated in the 2011 grade 4 TIMSS mathematics assessments, with their expected item scores computed for all items in the 2007 and 2011 grade 4 TIMSS mathematics assessments combined. Per Bock, Thissen, and Zimowski (1997), when these expected item scores are summed across items within a topic, the resulting score can be thought of as a student's degree of mastery of the topic.

[2] Domain score estimates refer to the estimates of student proficiency across the domain of interest, which may be the subject domain (e.g., mathematics) or a subsidiary level of the subject domain (e.g., the topics that make up the mathematics domain).

There are several technical conditions that need to be met in order to generate an estimated domain score this way. As discussed in Bock, Thissen, and Zimowski (1997), these conditions include the following:

- The relative weights of items in the item pool in defining the domain of interest are known.
- Item parameters have been estimated in large samples of examinees from a specified population.[3]
- The item response models show satisfactory item and model fit.
- The test is composed of items from an item pool that cover the domain of interest.

All of these conditions are in fact met in the case of TIMSS data because these are the same conditions that permit the conventional reporting of TIMSS scores by the TIMSS International Study Center (ISC). The relative weight of items in the TIMSS item pool at the topic level can easily be computed using the number of points associated with each subtopic under the given topic. The size of the sample used in item parameter estimation well exceeds the minimum 500 recommended by Bock, Thissen, and Zimowski (1997). TIMSS 2011 item parameters were estimated from a total student sample of over 300,000 at grade 4 (with more than 42,000 student responses per item) (Mullis et al. 2012). Model and item fit are routinely checked in the scaling of the TIMSS student achievement data by comparing the item response function curves generated using the item parameters estimated from the data with the empirical item response functions (Foy, Brossman, and Galia 2013). Finally, since a relatively large number of items, accumulated across two administrations,[4] are used in analyzing achievement in each topic, we assume that the fourth condition listed above is also satisfied in this study.

Besides these technical reasons that warrant using an IRT-based domain score approach to estimate student proficiency at the topic level, another compelling reason for using this approach is that for secondary users using standard statistical software, replicating the process of generating IRT-based domain scores at the topic level is relatively straightforward: these users only need item weights[5] in the domain, item parameter estimates, and student ability estimates on the theta metric.[6] TIMSS routinely releases all item parameters as part of its technical reports. In addition, public-use TIMSS data files include, for each student, five plausible values per content area that can easily be converted to a theta metric using the linear transformation coefficients that TIMSS uses in generating the plausible values, which also are available in TIMSS technical reports.

Moreover, using an IRT-based domain score approach allows one to generate scores that are relatively simple to interpret, since they can be expressed in terms of the *number of points earned* in the given domain (e.g., 20 out 25 points) or as a *percentage of points earned* in the given domain (e.g., 80 percent). "Domain scores offer the possibility of facilitating interpretation and evaluation of performance, provided

[3] It is important to keep in mind that the condition of a "specified population" for IRT may encompass more than demographic parameters for a population. For example, items administered in different years of TIMSS may have been under different educational conditions or curricular requirements, which would effectively mean that responses from different cycles were collected from nonequivalent populations for the purposes of policy analysis.
[4] Although the items come from two administrations of TIMSS (2007 and 2011), they represent items written to the 2003, 2007, and 2011 framework because TIMSS retains about one-half of its items from the previous round for measuring trends. Note that to apply this methodology, care must be taken to ensure that the framework remains the same for all used items or, if the framework is refreshed, it cannot have been substantively changed for items to cover a consistent domain of interest.
[5] In section 3, we explain how item weights can be modified for topic-level analysis.
[6] Analysis must also take into account design variables, including sampling weights. Today, most statistical software programs (such as SAS, SPSS, STATA, and R) can handle such design variables.

the domain has been well defined" (Pommerich, Nicewander, and Hanson 1999, p. 199). Bock, Thissen, and Zimowski (1997) argue that "the meaning of a reported percent of domain attainment requires little further explanation" (p. 208) if the content of the domain has been adequately described. In the case of TIMSS, the framework and the released items help define the domain, satisfying this condition.

Section 2: TIMSS Grade 4 Mathematics Assessments

To understand the method presented in this report, it is important to understand how TIMSS mathematics assessment items are classified, scored, and benchmarked in general. This section explains this key background information.

Items and their classification

The TIMSS mathematics framework describes the mathematics content to be assessed at the target grades by classifying the content into specific categories. For example, exhibit 1 displays how fourth-grade mathematics content is organized in a hierarchical manner based on the most current version of the framework: the TIMSS 2015 mathematics framework.

In TIMSS, items are administered using a matrix sampling design so that each student receives 1 of 14 test booklets. Each booklet contains two mathematics item blocks as well as two science blocks. Across the 14 test booklets in the TIMSS 2011 grade 4 mathematics assessment, there were a total of 14 mathematics item blocks. Six of these 14 item blocks were composed of new operational items for 2011; the remaining 8 blocks were composed of items carried forward from the previous TIMSS assessment cycles, called *trend items*. For each cycle of TIMSS, trend items are the basis for linking the TIMSS assessments and maintaining trends in performance measures over time (Foy, Brossman, and Galia 2013).

Table 1 shows the number of mathematics items used in this study, i.e., all items in the 2007 and 2011 TIMSS grade 4 mathematics assessments that have published item difficulty estimates that were calibrated to a common metric. As displayed in the table, the total number of score points[7] in the TIMSS 2007 and 2011 grade 4 mathematics assessments ranged from 23 in Points, Lines, and Angles (G1) to 72 in Whole Numbers (N1) at the topic level (see gray columns).

[7] Dichotomously scored items feature one point. Polytomously scored items can feature two or more points, depending on the rubric used in scoring these items.

Table 1. Number of TIMSS grade 4 mathematics items and score points used in this study, by content area, topic area, and TIMSS assessment year: TIMSS 2007 and 2011

Content area	Topic area	Total item pool for topic-level estimates		TIMSS 2007 items released		Trend items TIMSS 2007 and TIMSS 2011[1]		New items introduced in TIMSS 2011	
		Items	Score points	Items	Score points	Items	Score points	Items	Score points
Number	N1: Whole Numbers	66	72	20	22	28	31	18	19
	N2: Fractions and Decimals	34	34	9	9	12	12	13	13
	N3: Expressions, Simple Equations, and Relationships	27	27	9	9	13	13	5	5
	Total	**127**	**133**	**38**	**40**	**53**	**56**	**36**	**37**
Geometric Shapes and Measures	G1: Points, Lines, and Angles	23	23	4	4	10	10	9	9
	G2: Two- and Three-Dimensional Shapes	62	66	20	20	26	30	16	16
	Total	**85**	**89**	**24**	**24**	**36**	**40**	**25**	**25**
Data Display	D1: Reading, Interpreting, and Representing	37	39	12	13	14	15	11	11
Total		**249**	**261**	**74**	**77**	**103**	**111**	**72**	**73**

[1] Note that items used in both TIMSS 2007 and TIMSS 2011 are trend items that could have been written originally for the TIMSS 2003 mathematics framework or for the TIMSS 2007 mathematics framework, but which were classified by the TIMSS International Study Center (ISC) according to the 2011 framework for the purposes of scoring TIMSS 2011.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS) 2007 and 2011.

The TIMSS mathematics framework is modified slightly for every new administration in order to keep the framework educationally relevant and current. During such updates, some topics and/or objectives may be modified, moved, or replaced. As a result of such changes between 2007 and 2011, TIMSS reclassified some of the 2007 items into the new schema of topics and objectives so that they could be used in the 2011 cycle. However, items that were released in 2007 did not get reclassified by TIMSS because they were no longer being administered in 2011. In this study, all items—released, trend, and new 2011 items—were classified according to the most recent TIMSS framework (2015) by two mathematics experts so that they could be readily analyzed in the future with the TIMSS 2015 items when they are released.[8]

Current reporting in TIMSS

TIMSS mathematics scale and international benchmarks

Overall scores and content area scores in TIMSS mathematics assessments are reported on a 0 to 1,000 scale. Because TIMSS uses a matrix-sampling booklet design where each student is administered only a subset of the entire item pool, TIMSS does not report scores at the individual student level. Student achievement is estimated at the group level by the use of IRT scaling and marginal estimation in the following four steps (Mullis et al. 2012):

1. Student responses to individual items are scored, and all items are calibrated and placed onto a common theta (θ) scale.

[8] See appendix A for full details on the process of reclassification. For the purposes of topic-level analyses, all items were reclassified because even the items released after 2007 are of value as they have a *theta,* or measure of student performance, on a topic that was tested in a TIMSS cycle. Reclassifying these items, in the same way that trend items are reclassified by the ISC, allows these 2007-released items to provide data to support estimates of subscores.

2. Student achievement distribution is estimated on the theta (θ) metric for the overall mathematics scale and the content areas using the item parameter estimates and the conditioning variables. Design variables (e.g., sampling weights) are incorporated in this estimation.
3. Multiple plausible values[9] are drawn from student achievement distributions.
4. The plausible values on the theta (θ) metric are linearly transformed to the reporting metric and recorded in databases for secondary users of TIMSS data. A different set of transformation constants is used for each of the five plausible values (table 2) such that, for student i and draw p, $PV_{ip} = A_p + B_p * \theta_{ip}$.

Table 2. Linear transformation constants for the TIMSS 2011 fourth-grade mathematics assessment

Draw (p)	A_p	B_p
1	516.32438	94.09515
2	516.41272	94.32281
3	516.60847	93.17783
4	516.19085	94.03151
5	515.99456	94.10840

SOURCE: Foy, P., Brossman, B., and Galia, J. (2013).

In addition to estimating the scale scores, TIMSS also estimates and reports the percentage of students at or above four international benchmarks, or achievement levels: Advanced, High, Intermediate, and Low.

Section 3: An Illustration of an IRT-Based Domain Scoring Approach for Estimating Topic-Level Achievement

Steps for estimating topic-level achievement

Once the items in each topic are properly classified (as described in appendix A), the next task is to compute a statistic to summarize student performance for each topic. As mentioned earlier, this is accomplished by estimation of the probability of a correct (and partially correct for polytomous items) answer on all items in a given topic and summing these across items to generate a topic-level score. The following steps, originally proposed by Bock, Thissen, and Zimowski (1997), explain the process in more detail:

1. Compute the weight of each item, in terms of its contribution to the topic-level score, according to the objective and topic it belongs to.
2. Convert each of the five plausible values for each content area for each student to the theta metric (θ_{ip}) using the linear transformation coefficients displayed in table 2.

[9] For a detailed discussion of plausible values, see Mislevy (1991) and von Davier, Gonzalez, and Mislevy (2009).

3. For each item, compute an expected item score using
    a. IRT parameter estimates for the given item; and
    b. the theta (θ_{ip}) equivalent of each of the five plausible values obtained in step 2, above.
4. Multiply each of these five expected item scores by the weight computed in step 1 for the respective item.
5. For each topic, sum each of the five weighted expected item scores, one at a time, from the previous step across all items within the topic. This summation produces five model-based topic scores (MTS_{ip}). Dividing each MTS_{ip} by the sum of item weights for the given topic and multiplying that product by 100 produces model-based topic percent scores ($MTPS_{ip}$).

After computing five MTS and five MTPS for each student for each topic, the standard plausible values methodology is applied to these to compute average MTS and average MTPS for the desired student group. A detailed description of these five steps follows.

Step 1: Computing item weights

As discussed earlier, the TIMSS 2007 and 2011 fourth-grade mathematics assessments include three content areas (Number, Geometric Shapes and Measures, and Data Display). Each content area includes multiple topics, and each topic is composed of multiple objectives. To compute the weight of an item that measures a specific objective, the first step is to calculate the ratio of the total points in the item's respective objective to the total points in its respective topic in 2011. For example, of the 19 points available in topic G1 (table 3), 4 were from objective G1.1, corresponding to a ratio of 4/19 = .21 for this objective. We kept this ratio constant even after items from 2007 were brought into the G1 item pool. Then, the weight of each item in a given objective (w_j) within a given topic was computed such

$$\frac{w_j * n_{(2007+2011)}}{N_{(2007+2011)}} = \frac{n_{(2011)}}{N_{(2011)}},$$

where $n_{(year)}$ is the number of points in the given objective for the given administration year(s), and $N_{(year)}$ is the total number of points in the given topic that the objective belongs to, for the given administration year(s).

All items that measure the same objective receive the same weight. Table 3 shows the weights for items within topic G1 to illustrate this computation.

Table 3. Computation of item weights for topic G1 (Points, Lines, and Angles): TIMSS 2007 and 2011

Objective	Points		Total points	2011 ratio	Item weight
	2007	2011			
G1.1	4	4	8	0.21	0.61
G1.2	0	4	4	0.21	1.21
G1.3	0	5	5	0.26	1.21
G1.4	0	6	6	0.32	1.21
G1 Total	**4**	**19**	**23**	**†**	**†**

† Not applicable.
NOTE: The item weights and the 2011 ratios are rounded results, and the item weights are calculated from the original values used to calculate the ratios and not from the rounded ratios.
SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007 and 2011.

Step 2: Conversion of plausible values to the theta (θ) metric

In this second step, each plausible value for each content area for each student is converted to the theta (θ) metric by solving for Θ_{ip}, in the following equation:

$$PV_{ip} = A_p + B_p * \Theta_{ip},$$

Where PV_{ip} is the plausible value p (p = 1, …, 5) for student i in a given content area, and A_p and B_p are the published linear transformation coefficients (table 2) for each plausible value.

This results in a total of 15 Θ_{ip} for each student, 5 from each of the three content areas.

Step 3: Computing expected item scores

The third step is to compute an expected item score, $E(X_{ipj} \mid \Theta_{ip})$, for each student i for each item j within each topic, based on Θ_{ip}, the item parameter estimates, and the item response model for each item:

$$E(X_{ipj} \mid \Theta_{ip}) = \sum_{h=0}^{m_j} h\, P(X_{ipj} = h \mid \Theta_{ip}),$$

where h denotes the score category (0 or 1 in dichotomous items and 0, 1, or 2 in polytomous items[10]), m_j is the maximum score attainable for item j, and $P(X_{ipj} = h \mid \Theta_{ip})$ is the probability of a score of h on item j, given Θ_{ip}.

Note that since $E(X_{ipj} \mid \Theta_{ip})$ ranges from 0 to 2 for polytomous items, these items contribute twice as much to the model-based topic score compared to a dichotomous item. In computing the expected item scores, we used the same item response model that TIMSS used in calibrating the item:

- a three-parameter logistic model (3PL) was used with dichotomously scored multiple-choice items;

[10] Polytomous items can theoretically assume higher values; however, no polytomous item in our data assumed a higher value than 2.

- a two-parameter logistic model (2PL) was used with dichotomously scored constructed-response items; and
- a partial credit model was used with polytomous constructed-response items.

Note that the item parameter estimates used in the study were estimated and published by TIMSS as part of their operational work. In order to maintain trend, TIMSS estimates the item parameters for the items in the current assessment through a concurrent calibration of the data from the current assessment and from the previous assessment. In 2011, TIMSS concurrently calibrated the 2011 and 2007 items, putting them on the same metric (Foy, Brossman, and Galia 2013).

Step 4: Weighting each expected item score
In the fourth step, each expected item score is simply multiplied by its weight from step 1.

Step 5: Computing model-based topic scores (MTS) and model-based topic percent scores (MTPS)
In the last step, the products from step 4 are summed across all items within a topic. This summation produces five model-based topic scores (MTS_{ip}) (p = 1, …, 5) within each topic:

$$MTS_{ip} = \sum_{j=1}^{N} w_j E\left(X_{ipj} | \theta_{ip}\right),$$

where w_j denotes the weight of item j in the given topic, N is the number of items in the given topic, and $E\left(X_{ipj} \mid \theta_{ip}\right)$ is the expected item score for student i for item j, given θ_{ip}.

Dividing each MTS_{ip} by the sum of item weights for the given topic and multiplying that product by 100 produces five model-based topic percent scores ($MTSP_{ip}$) (p = 1, …, 5):

$$MTPS_{ip} = 100 * \frac{MTS_{ip}}{\sum_{j=1}^{n} w_j}.$$

In computing mean MTS and mean MTPS for the desired student group for a given topic, the standard plausible values methodology was applied. The mean of interest was computed five times, using each of the five MTS_{ip} and $MTSP_{ip}$ values. The average of these five means produces the mean MTS and mean MTPS, respectively. The variance of this statistic has two components: sampling variance and imputation variance.[11] The standard error of the estimate for the mean is the square root of the sum of these two variance components. The same procedure was followed in computing the mean MTPS for the desired student group for a given topic and its associated standard error.

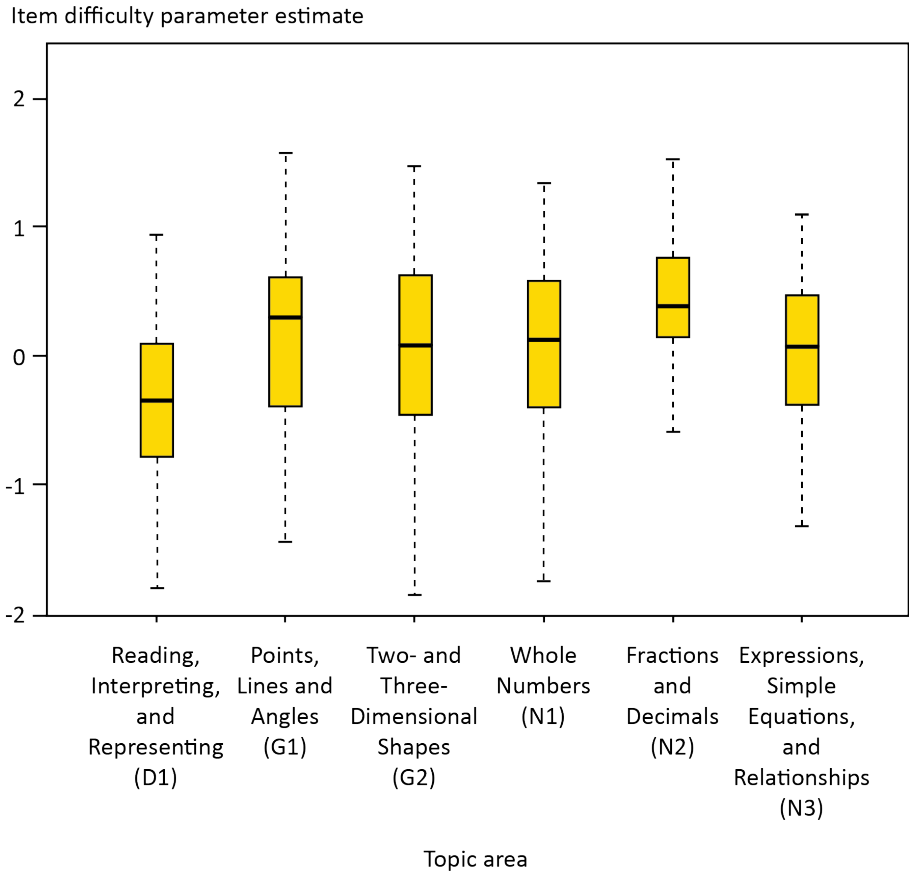
Results

Because the model-based topic scores are based on item parameters in each topic, we first examined the distribution of the difficulty parameter (b) estimates by topic. Mean estimates for item difficulty parameters ranged from -0.38 (Reading, Interpreting, and Representing) to 0.42 (Fractions and Decimals), with an overall mean of 0.03 and a standard deviation of 0.71. Figure 1 displays the range of the topic-

[11] See equation 5.2 in Beaton et al. (2011) for the computation of imputation variance.

level parameter estimates. An omnibus test across the six topics revealed statistically significant differences ($F_{5, 249} = 4.87$, $p < .05$). In addition, the mean estimate for Fractions and Decimals was higher than the mean estimates for Whole Numbers (t_{90}, 2.98, $p < .05$) and for Expressions, Simple Equations, and Relationships (t_{59}, 2.69, $p < .05$). These differences indicated that it is reasonable to expect differences in topic-level performance using the MTS and MTPS metrics.

Figure 1. Item difficulty parameter estimates, by topic area: 2007 and 2011 TIMSS mathematics item pool



SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007 and 2011.

Achievement profile of U.S. students in terms of topic-level performance

The mean MTS and MTPS were computed for the U.S. sample that took the 2011 assessment (table 4).[12] A total of 249 items, corresponding to 261 points in the combined item pool (i.e., all items from TIMSS 2007 and TIMSS 2011), were used to compute these scores. MTS are additive across topics since each item belongs to a single topic. The average MTS across all topics add up to 155.9, indicating that the 2011 U.S. sample would have earned, on average, 59.7 percent of 261 points had they been given the entire pool of 249 items. The mean MTPS add up to 70.3, 59.1, and 57.0 in Data, Geometry, and Numbers, respectively, when aggregated within each content area. At the topic level, the mean MTPS

[12] All design variables, including sampling weights, were incorporated in these analyses.

ranged from 52.4 in Fractions and Decimals to 70.3 in Reading, Interpreting, and Representing, showing considerable variation across topics.

The mean MTPS were also compared by gender (table B-1) and race/ethnicity (table B-2). The largest difference between the two groups in mean MTPS was observed in Fractions and Decimals (N2), where the means were 53.7 and 50.8 for male and female students, respectively, corresponding to an effect size of 0.09, which is small according to Cohen's (1988) convention for interpreting effect sizes.[13]

Table 4. Mean model-based topic score and model-based topic percent score for U.S. students, by content area and topic area: 2011 TIMSS mathematics assessment

[Standard errors appear in parentheses]

Content area and topic area	Number of Items	Number of Points	Model-based topic score	Model-based topic percent score
Numbers	127	133	75.8 (0.67)	57.0 (0.50)
Whole Numbers (N1)	66	72	41.6 (0.35)	57.8 (0.48)
Fractions and Decimals (N2)	34	34	17.8 (0.19)	52.4 (0.56)
Expressions, Simple Equations, and Relationships (N3)	27	27	16.4 (0.13)	60.8 (0.49)
Geometry	85	89	52.6 (0.42)	59.1 (0.47)
Points, Lines, and Angles (G1)	23	23	14.1 (0.11)	61.5 (0.46)
Two- and Three-Dimensional Shapes (G2)	62	66	38.5 (0.32)	58.3 (0.48)
Data Display	37	39	27.4 (0.14)	70.3 (0.36)
Reading, Interpreting, and Representing (D1)	37	39	27.4 (0.14)	70.3 (0.36)
Total	**249**	**261**	**155.9 (1.22)**	**59.7 (0.47)**

SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2011.

When analyzed by race/ethnicity, the MTPS add up to a maximum of 70.2 (Asian) and a minimum of 46.6 (Black) across all six topics (table B-2). As with gender, the largest differences in MTPS means were in the topic of Fractions and Decimals, where the difference in means between White and Asian students was 8.9 in favor of Asian students, with a small effect size of 0.29, and the difference in means between White and Black students was 19.1 in favor of White students with a medium effect size of 0.66.

Section 4: Conclusion

Topic-level scores for an assessment such as TIMSS can be generated by estimating how a particular cohort would have performed on items pooled across multiple administrations of the assessment if (a) the items are based on the same framework, (b) the populations across years are equivalent, and (c) the items are calibrated to the same scale. Such analysis opens up the possibility of using IRT-based domain scoring to estimate topic-level scores and to consider the relationships between specific areas of achievement and what is taught (or "opportunity to learn") across subgroups as well as across different school settings. An advantage offered by this method is that it is relatively straightforward for a secondary user employing standard statistical software to use it to replicate and produce scores that are relatively

[13] Effect size is computed as the difference in mean MTPS divided by the square root of the pooled variance of the MTPS. Note that statistical significance tests (i.e., *t* tests) were not conducted for group differences on the MTPS because the subgroups' MTS and MTPS distributions were found to be highly nonnormal, which could be skewed, multimodal, or uniform. Attempts to transform these scores to more normal distributions were not successful. Due to the serious violation of distributional assumption, significance tests were not carried out on the group differences.

simple to interpret, given that they can be expressed in terms of MTS (e.g., 20 out 25 points) or MTPS (e.g., 80 percent). This method also enhances the interpretation of group differences in terms of achievement in more specific terms. For instance, in the case of differences between White and Black students, several studies, including TIMSS, have demonstrated significant gaps. However, we know now, for the first time, where that gap is largest (Fractions and Decimals) and how wide that gap is: 57 percent versus 38 percent mastery of the topic, corresponding to a medium effect size of 0.66. The method explained in this paper has been illustrated with TIMSS 2011 grade 4 mathematics data; however, the same method can be applied to other grades, subjects, and assessments that meet the conditions described in section 1. An inherent limitation to the approach introduced in the study is that domain scores are correlated to the same degree that the theta estimates used to produce them are correlated. This does not, however, change the fact that students' level of mastery of each topic is different.

Appendix A: Crosswalk Analysis

In order to compute topic scores, it is necessary to sort the pool of available items into the proper topic classification. (For this study, the pool consists of the TIMSS 2007 and TIMSS 2011 grade 4 mathematics items shown in table 1.)

To do this, the first step is to create a crosswalk between the various TIMSS mathematics frameworks to which the items have been classified. This is done by subject-matter experts conducting a crosswalk analysis, which entails analyzing how items written for one framework would be classified within a new framework's schema of content area, topic area, and objective. When there are no changes between two versions of the framework, the resulting crosswalk has a one-to-one correspondence between every topic area and objective. When changes occur between different versions of the framework, the resulting crosswalk indicates how items classified for one version of the framework would be reclassified within the revised version of the framework. For reliability's sake, experts conduct a crosswalk analysis independently and then adjudicate any cases where they have a slightly different mapping of topic area or objective.

Once a crosswalk has been fully mapped between two versions of the TIMSS framework, it is a simple matter to sort available items by using the crosswalk to "translate" the items' original classifications into the classification schema of the target version of the framework. All items that the crosswalk indicates should be reclassified into a new objective, however, should be checked to be sure that they do actually fit.

The results of the mapping, or crosswalk, analysis for this paper are presented below in table A-1. This crosswalk analysis was conducted by two mathematics experts who each have over 10 years of experience working with both the NAEP and TIMSS mathematics frameworks and items. Specifically, they have both undertaken similar exercises in the past with the NAEP mathematics items. Additionally, one of them has extensive experience in developing TIMSS frameworks and writing TIMSS items, while the other has substantive experience in developing NAEP mathematics items.

For most of the objectives in the 2011 and 2007 mathematics grade 4 frameworks, the experts were able to find a matching objective in the respective 2015 framework. However, in a few cases, an objective matched more than one objective/topic area/content area or did not match an objective. These few exceptions are shown in italics in table A-1. In these instances, the experts reviewed the items that fell under those objectives and coded each item into a specific objective–topic of the 2015 framework based on the major focus of the item. For example, the experts reviewed the 12 items that were identified under N1.3 (third objective under the first topic of the Number content area) in the TIMSS 2011 international item information sheet, mapped each item to a specific objective of the respective 2015 framework, and thereby verified that all of those 12 items are indeed categorized under N1 (Whole Numbers—first topic of the Number content area).

Table A-1. Number of TIMSS 2007 and 2011 grade 4 mathematics items classified into TIMSS 2015 framework by content area, topic area, and objective

TIMSS 2007 Content area, topic area, & objective	Number of items	**TIMSS 2011** Content area, topic area, & objective	Number of items
Number		**Number**	
N1: Whole Numbers		N1: Whole Numbers	
N1.1 to N1.1	1	N1.1 to N1.1	5
N1.2 to N1.1	—	N1.2 to N1.2	3
N1.3 to N1.2	1	*N1.3 to N1.3 or No match*	12
N1.4 to N1.3	5	N1.4 to N1.5	4
N1.5 to N1.5 or D1.1	1	N1.5 to N1.4	21
N1.6 to No match	—		
N1.7 to N1.4	11		
N1.8 to N1.4	1		
Total N1 to N1	**20**	**Total N1 to N1**	**45**
N2: Fractions and Decimals		N2: Fractions and Decimals	
N2.1 to N2.1	1	N2.1 to N2.1	7
N2.2 to N2.1	1	N2.2 to N2.2	5
N2.3 to N2.2	1	N2.3 to N2.2	4
N2.4 to N2.2	2	N2.4 to N2.3	3
N2.5 to N2.3	1	N2.5 to N2.3	3
N2.6 to N2.3	2	*N2.6 to N2.2 or N2.3*	3
N2.7 to N2.2 or N2.3	1		
Total N2 to N2	**9**	**Total N2 to N2**	**25**
N3: Number Sentences		N3: Number Sentences	
N3.1 to N3.1	2	N3.1 to N3.1	4
N3.2 to N3.2	2	N3.2 to N3.2	3
N4: Patterns and Relationships		N4: Patterns and Relationships	
N4.1 to N3.3	2	N4.1 to N3.3	8
N4.2 to N3.3	—	N4.2 to N3.3	3
N4.3 to N3.3	—		
N4.4 to N3.3	3		
Total N3 and N4 to N3	**9**	**Total N3 and N4 to N3**	**18**
Geometric Shapes and Measures		**Geometric Shapes and Measures**	
G1: Lines and Angles		G1: Points, Lines, and Angles	
G1.1 to G1.1	4	G1.1 to G1.1	4
G1.2 to G1.2	—	G1.2 to G1.2	4
G1.3 to G1.3	—	G1.3 to G1.3	5
		G1.4 to G1.4	6
Total G1 to G1	**4**	**Total G1 to G1**	**19**
G2: 2- and 3-Dimensional Shapes		G2: 2- and 3-Dimensional Shapes	
G2.1 to G2.1	1	G2.1 to G2.1	6
G2.2 to G2.1	6	G2.2 to G2.1	18
G2.3 to G2.1	3	G2.3 to G2.2	5
G2.4 to G2.2	1	G2.4 to G2.3	13
G2.5 to G2.3	2		
G2.6 to G2.3	4		
G3: Location and Movement			
G3.1 to G1.4	—		
G3.2 to G2.1	2		
G3.3 to G2.1	1		
Total G2 and G3 to G2	**20**	**Total G2 to G2**	**42**

See notes at end of table.

Table A-1. Number of TIMSS 2007 and 2011 grade 4 mathematics items classified into TIMSS 2015 framework by content area, topic area, and objective —Continued

Data Display		Data Display	
D1: Reading and Interpreting		D1: Reading and Interpreting	
D1.1 to D1.1	1	D1.1 to D1.1	5[1]
D1.2 to D1.2	2	D1.2 to D1.2	3
D1.3 to D1.2	1	D1.3 to D1.2	10
D2: Organizing and Representing		D2: Organizing and Representing	
D2.1 to D1.1	2	D2.1 to D1.1	3
D2.2 to D1.1	6	D2.2 to D1.1	4
Total D1 and D2 to D1	**12**	**Total D1 and D2 to D1**	**25**

— Not available.

[1] M031294 was identified as a D1.1 item in the TIMSS 2011 international item information file, but it was mapped into topic N1 (Whole Numbers) of the TIMSS 2015 grade 4 mathematics framework during this content analysis.

NOTE: During the mapping analysis, most objectives in the 2011 and 2007 mathematics grade 4 frameworks were matched with an objective in the 2015 mathematics grade 4 framework. If an objective matched more than one objective/topic area/content area or did not match an objective, it was shown in italics in table A-1.

SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007 and 2011.

Appendix B: Topic Scores by Gender and Race/Ethnicity

Table B-1. Mean model-based topic score and model-based topic percent score for U.S. students, by gender, content area, and topic area: 2011 TIMSS mathematics assessment

[Standard errors appear in parentheses]

Content area and topic area	Number of		Model-based topic score		Model-based topic percent score	
	Items	Points	Male	Female	Male	Female
Numbers	127	133	77.3 (0.94)	73.9 (1.11)	58.1 (0.71)	55.6 (0.84)
Whole Numbers (N1)	66	72	42.3 (0.49)	40.6 (0.58)	58.8 (0.68)	56.4 (0.80)
Fractions and Decimals (N2)	34	34	18.3 (0.27)	17.3 (0.31)	53.7 (0.79)	50.8 (0.92)
Expressions, Simple Equations, and Relationships (N3)	27	27	16.7 (0.19)	16.0 (0.22)	61.9 (0.69)	59.4 (0.82)
Geometry	85	89	53.4 (0.60)	51.9 (0.70)	60.0 (0.67)	58.3 (0.79)
Points, Lines, and Angles (G1)	23	23	14.3 (0.15)	13.9 (0.18)	62.3 (0.65)	60.6 (0.77)
Two- and Three-Dimensional Shapes (G2)	62	66	39.1 (0.45)	37.9 (0.53)	59.2 (0.68)	57.4 (0.80)
Data Display	37	39	27.6 (0.21)	27.2 (0.28)	70.7 (0.53)	69.7 (0.71)
Reading, Interpreting, and Representing (D1)	37	39	27.6 (0.21)	27.2 (0.28)	70.7 (0.53)	69.7 (0.71)
Total	**249**	**261**	**158.3 (1.68)**	**152.9 (2.01)**	**60.6 (0.64)**	**58.6 (0.77)**

NOTE: The unweighted sample size for Male was 6,227 and for Female was 6,342. The population size represented by these samples was 1,759,426 for Male and 1,808,420 for Female. All design variables, including sampling weights, were incorporated in estimating mean scores.

SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007 and 2011.

Table B-2. Mean model-based topic percent score for U.S. students, by race/ethnicity, content area, and topic area: 2011 TIMSS mathematics assessment

[Standard errors appear in parentheses]

Content area and topic area	Number of Items	Number of Points	Model-based topic percent score White	Black	Hispanic	Asian	Two or more races	Other
Numbers	127	133	61.4 (0.87)	43.8 (1.13)	51.8 (0.79)	69.1 (2.17)	59.8 (1.61)	51.9 (1.85)
Whole Numbers (N1)	66	72	62.0 (0.83)	44.9 (1.12)	52.8 (0.76)	69.3 (2.05)	60.4 (1.55)	52.9 (1.80)
Fractions and Decimals (N2)	34	34	57.2 (0.99)	38.1 (1.13)	46.6 (0.85)	66.1 (2.48)	55.4 (1.77)	46.6 (1.96)
Expressions, Simple Equations, and Relationships (N3)	27	27	65.2 (0.84)	47.7 (1.15)	55.7 (0.78)	72.5 (2.08)	63.6 (1.58)	55.8 (1.84)
Geometry	85	89	64.0 (0.79)	45.2 (1.21)	54.1 (0.75)	68.4 (2.02)	64.0 (1.54)	54.6 (1.44)
Points, Lines, and Angles (G1)	23	23	66.3 (0.76)	47.8 (1.24)	56.5 (0.75)	70.4 (1.93)	66.3 (1.49)	57.1 (1.43)
Two- and Three-Dimensional Shapes (G2)	62	66	63.3 (0.80)	44.4 (1.20)	53.2 (0.76)	67.6 (2.05)	63.2 (1.55)	53.8 (1.45)
Data Display	37	39	74.7 (0.60)	59.7 (1.25)	65.0 (0.80)	78.0 (1.93)	73.3 (1.62)	67.4 (2.14)
Reading, Interpreting, and Representing (D1)	37	39	74.7 (0.60)	59.7 (1.25)	65.0 (0.80)	78.0 (1.93)	73.3 (1.62)	67.4 (2.14)
Total	**249**	**261**	**64.3 (0.77)**	**46.6 (1.11)**	**54.5 (0.72)**	**70.2 (1.91)**	**63.3 (1.40)**	**55.1 (1.58)**

NOTE: The "Other" category consisted of the small numbers of students indicating that they were "American Indian or Alaska Native" or "Native Hawaiian or Other Pacific Islander." Race categories exclude students of Hispanic ethnicity. The unweighted sample size for White was 6,137; for Black, 1,408; for Hispanic, 3,325; for Asian, 524; for Two or more races, 482; and for Other, 452. The population size represented by these samples was 1,851,583 for White; 396,028 for Black; 851,257 for Hispanic; 136,398 for Asian; 142,871 for Two or more races; and 130,481 for Other. All design variables, including sampling weights, were incorporated in estimating mean scores.

SOURCE: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2007 and 2011.

References

Beaton, A.E., Rogers, A.M., Gonzalez, E., Hanly, M.B., Kolstad, A., Rust, K.F., Sikali, E., Stokes, L., and Jia, Y. (2011). *The NAEP Primer* (NCES 2011-463). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Bock, R.D., Thissen, D., and Zimowski, M.F. (1997). IRT Estimation of Domain Scores. *Journal of Educational Measurement, 34*(3): 197–211.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* New York: Academic Press.

Foy, P., Brossman, B., and Galia, J. (2013). Scaling the TIMSS and PIRLS 2011 Achievement Data. In M. Martin and I. Mullis (Eds.), *Methods and Procedures in TIMSS and PIRLS 2011* (pp. 1–28). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Haladyna, T.M., and Roid, G.H. (1983). A Comparison of Two Approaches to Criterion-Referenced Test Construction. *Journal of Educational Measurement*, *20*(3): 271–282.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Erlbaum.

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

Mullis, I.V.S., Martin, M.O., Foy, P., and Arora, A. (2012). *TIMSS 2011 International Results in Mathematics.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Pommerich, M. (2006). Validation of Group Domain Score Estimates Using a Test of Domain. *Journal of Educational Measurement, 43*(2), 97–111.

Pommerich, M., Nicewander, W.A., and Hanson, B.A. (1999). Estimating Average Domain Scores. *Journal of Educational Measurement, 36*(3): 199–216.

Skorupski, W.P. (2008, August). *A Review and Empirical Comparison of Approaches for Improving the Reliability of Objective Level Scores.* Paper presented at the annual meeting of the Council of Chief State School Officers.

von Davier, M., Gonzalez, E., and Mislevy, R.J. (2009). What Are Plausible Values and Why Are They Useful? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments.* Retrieved June 21, 2018, from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf.