

International Journal of Testing

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/hijt20>

A Cognitive Diagnostic Modeling of Attribute Mastery in Massachusetts, Minnesota, and the U.S. National Sample Using the TIMSS 2007

Young-Sun Lee ^a, Yoon Soo Park ^a & Didem Taylan ^b

^a Department of Human Development, Teachers College, Columbia University

^b Department of Learning, Teaching, & Curriculum, University of Missouri, Columbia

Available online: 29 Apr 2011

To cite this article: Young-Sun Lee, Yoon Soo Park & Didem Taylan (2011): A Cognitive Diagnostic Modeling of Attribute Mastery in Massachusetts, Minnesota, and the U.S. National Sample Using the TIMSS 2007, *International Journal of Testing*, 11:2, 144-177

To link to this article: <http://dx.doi.org/10.1080/15305058.2010.534571>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.tandfonline.com/page/terms-and-conditions>

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to

date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

A Cognitive Diagnostic Modeling of Attribute Mastery in Massachusetts, Minnesota, and the U.S. National Sample Using the TIMSS 2007

Young-Sun Lee and Yoon Soo Park
*Department of Human Development, Teachers College,
Columbia University*

Didem Taylan
*Department of Learning, Teaching, & Curriculum,
University of Missouri, Columbia*

Studies of international mathematics achievement such as the Trends in Mathematics and Science Study (TIMSS) have employed classical test theory and item response theory to rank individuals within a latent ability continuum. Although these approaches have provided insights into comparisons between countries, they have yet to examine how specific attribute mastery affects student performance and how they can provide information for curricular instruction. In the 2007 administration of TIMSS, two benchmark participants—Massachusetts and Minnesota—were tested following the same procedural methods, providing an opportunity for comparison within and across the United States. Overall comparison of their performance showed Massachusetts and Minnesota to significantly outperform the United States. However, this article shows that there is a greater wealth of fine-grained information that can be translated directly for classroom application at the attribute level when a cognitive diagnostic model (CDM) such as the deterministic, inputs, noisy, “and” gate (Junker & Sijtsma, 2001) model is used. Results showed a significant disparity between proportions of correctly answering and mastering skills required to solve an item. Advantages of CDMs are discussed as well as a CDM-based method to filter distractor response categories that can aid instructors to diagnose a student’s attribute mastery.

Keywords: attribute mastery, cognitive diagnostic modeling, DINA, TIMSS

In mathematics education, researchers evaluate student achievement as well as their mastery of curricular instruction through international comparative studies such as the Trends in International Mathematics and Science Study (TIMSS) and the Organization for Economic Cooperation and Development Programme for International Student Assessment (PISA). While TIMSS is more interested in assessing students' knowledge of curricular topics in mathematics and science (i.e., reproduction of knowledge), PISA is designed to assess students' ability to apply what they learned in school to their daily activities (i.e., abilities of making connections to real-life situations; Hutchison & Schagen, 2007). As such, the specific aims of these international comparative studies provide opportunities and challenges for mathematics education researchers interested in using their findings, instruments, and theoretical perspectives as catalysts for secondary analysis and additional research (Ferrini-Mundy & Schmidt, 2005; Lemke et al., 2004).

In particular, the TIMSS, a quadrennial assessment administered by the International Association for the Evaluation of Educational Achievement (IEA) since 1995, evaluates the mathematics and science abilities of fourth and eighth graders. The significance of information from such sources is the production of reliable and timely data on American students' achievement from an international context and the possibility to analyze trends in student progress that can provide feedback for future improvement in areas that need further instruction.

An important distinction in TIMSS 2007 from its predecessors is the revival of two American states—Massachusetts and Minnesota—as benchmark participants, regional entities that follow the same assessment procedures as the countries. Educational researchers have often questioned the overall effectiveness of mathematics education in the United States, because American performance in international assessments has not dominated its international peers (American Federation of Teachers [AFT], 1999). According to a report to the US Department of Education, the US performance was below average for low- and high-level skills and also below average for items with low- and high-level difficulty when compared to other countries that had participated in the TIMSS 2003 and the PISA (Ginsburg et al., 2005).

However, results from TIMSS 2007 showed that this gap in performance was not uniform across the US. Students from both Massachusetts and Minnesota placed well in the international mathematics and science categories. Massachusetts ranked fourth and Minnesota ranked sixth overall in the forth-grade mathematics assessments (Table 1). Historically, the US ranked seventh out of 26 countries and twelfth out of 25 countries, when excluding benchmark participants in 1995 and 2003, respectively; TIMSS was not administered to fourth graders in 1999. Although the US demonstrated an overall improvement in performance since 1995 (e.g., the average scaled score for the US increased significantly

TABLE 1
Mean Proficiency Statistics for the TIMSS 2007 Fourth Grade Participants

| | Country | Sample Size | Mean Proficiency | Standard Error |
|----|----------------------------|-------------|------------------|----------------|
| 1 | Hong Kong SAR | 3791 | 606.80 | 3.58 |
| 2 | Singapore | 5041 | 599.41 | 3.74 |
| 3 | Chinese Taipei | 4131 | 575.82 | 1.73 |
| 4 | <u>Massachusetts, USA*</u> | <u>1747</u> | <u>572.48</u> | <u>3.51</u> |
| 5 | Japan | 4487 | 568.16 | 2.12 |
| 6 | <u>Minnesota, USA*</u> | <u>1846</u> | <u>554.12</u> | <u>5.86</u> |
| 7 | Kazakhstan | 3990 | 549.35 | 7.15 |
| 8 | Russian Federation | 4464 | 544.05 | 4.91 |
| 9 | England | 4316 | 541.47 | 2.88 |
| 10 | Latvia | 3908 | 537.20 | 2.31 |
| 11 | Netherlands | 3349 | 534.95 | 2.15 |
| 12 | Lithuania | 3980 | 529.80 | 2.37 |
| 13 | <u>United States</u> | <u>7896</u> | <u>529.01</u> | <u>2.45</u> |
| 14 | Germany | 5200 | 525.16 | 2.25 |
| 15 | Denmark | 3519 | 523.11 | 2.40 |
| 16 | Quebec, Canada* | 3885 | 519.10 | 3.03 |
| 17 | Australia | 4108 | 516.06 | 3.51 |
| 18 | Ontario, Canada* | 3496 | 511.61 | 3.10 |
| 19 | Hungary | 4048 | 509.72 | 3.55 |
| 20 | Italy | 4470 | 506.75 | 3.14 |
| 21 | Austria | 4859 | 505.39 | 2.01 |
| 22 | Alberta, Canada* | 4037 | 505.32 | 2.95 |
| 23 | British Columbia, Canada* | 4153 | 505.22 | 2.75 |
| 24 | Sweden | 4676 | 502.57 | 2.53 |
| 25 | Slovenia | 4351 | 501.84 | 1.81 |
| 26 | Armenia | 4079 | 499.51 | 4.29 |
| 27 | Slovak Republic | 4963 | 495.98 | 4.47 |
| 28 | Scotland | 3929 | 494.45 | 2.21 |
| 29 | New Zealand | 4940 | 492.48 | 2.31 |
| 30 | Czech Republic | 4235 | 486.40 | 2.78 |
| 31 | Norway | 4108 | 473.22 | 2.54 |
| 32 | Ukraine | 4292 | 469.00 | 2.91 |
| 33 | Dubai, UAE* | 3064 | 444.33 | 2.14 |
| 34 | Georgia | 4108 | 438.46 | 4.21 |
| 35 | Iran, Islamic Rep. of | 3833 | 402.42 | 4.05 |
| 36 | Algeria | 4223 | 377.65 | 5.18 |
| 37 | Colombia | 4801 | 355.45 | 4.97 |
| 38 | Morocco | 3894 | 341.31 | 4.67 |
| 39 | El Salvador | 4166 | 329.91 | 4.10 |
| 40 | Tunisia | 4,34 | 327.44 | 4.47 |
| 41 | Kuwait | 3803 | 315.54 | 3.65 |
| 42 | Qatar | 7019 | 296.27 | 1.04 |
| 43 | Yemen | 5811 | 223.68 | 5.97 |

Note: Regions underlined represent samples used in this study.

*Benchmark Participants, based on TIMSS 2007 Technical Report (Olson, Martin, & Mullis, 2009).

from 518 to 529 between 1995 to 2007), it ranked thirteenth overall. This result was also evident in science and also in the performance of eighth graders. The mean scaled scores for Massachusetts, Minnesota, and the US were 572.48 ($SE = 3.51$), 554.12 ($SE = 5.86$), and 529.01 ($SE = 2.45$), respectively. Although the scores of the two benchmark participants were marginally different, there was a significant difference between their performances to the US. Another phenomenon of note is the large variability in Minnesota's mean scaled score. The standard error of Minnesota's mean scaled score was 5.86 and was second highest among the 43 participating countries and regional entities. In comparison, the standard errors of Massachusetts and the US were 3.51 and 2.45, respectively.

Based solely on these results, it becomes natural to wonder about factors that relate to the wide variability in performance within the US. In light of such unexplainable variability, there is also the concern that simple comparisons based on overall performance may only provide limited practical implications that teachers or education practitioners can directly utilize in their classroom instruction.

Although various studies have scrutinized the TIMSS (Olson, Martin, & Mullis, 2009) using methods that employ classical test theory (CTT), item response theory (IRT), and generalizability theory, there has been a general concern within the mathematics education community that the information garnered through the TIMSS is not directly applicable at the classroom level (Holliday & Holliday, 2003; Wang, 2001). That is, estimating single overall scores that indicate a student's relative position on an ability continuum scale does not provide diagnostic information about items and students. To overcome these limitations, Lee, Choi, and Park (2009) proposed employing a cognitive diagnostic model (CDM) approach that was designed to provide additional information to researchers and educators on cognitive skills or attributes that are required to solve a particular item. Consequently, diagnostic information can be applied to various instructional practices by identifying the presence or the absence (i.e., mastery and non-mastery) of specific, fine-grained skills or attributes.

The purpose of this study is twofold—(1) to employ a CDM to identify item characteristics such as discrimination, slip, and guessing parameters, and (2) to examine students' mastery of attributes and whether they were able to judiciously utilize them during an exam setting. In doing so, it can help investigate what similarities and differences exist due to their attribute mastery in the United States and the two benchmark states. These facets extend beyond traditional means of item analysis and student performance, because it allows an attribute-level investigation that can generate inferences on whether a student has truly mastered all the required skills and whether a student can execute combinations of these skills appropriately to solve an item correctly. If a student that mastered all required skills fails to solve the item or if a student successfully solves an item without

having mastered all required skills, more can be inferred about the item and the curricular paradigm that led to such a result. This is the motivational basis of CDMs and the aim of this study—to examine diagnostic extensions of solving an item at the attribute level so that more information can be drawn from examinees' responses to explain differences among regional entities that traditional methods using simple overall scores cannot.

This comparative study of the US and the two benchmark states will shed light on information relevant to specific classroom instruction and student learning that are pertinent to a regional entity (i.e., state) versus a nation as a whole; in this case, we identified Massachusetts as a high-performing benchmark participant based on its overall score. Furthermore, there may also be skills identified to be equally present in the US and the two benchmark states, contradicting an intuitive notion that students from a better-performing state also possess a significantly greater amount of necessary skills required to solve a mathematics item. In short, the framework of CDM analysis provides more valuable diagnostic information about how well students performed on underlying skills and cognitive processes required in answering items, and it goes beyond simply identifying strengths and weaknesses in the performance of test takers of the US and the two benchmark states.

COGNITIVE DIAGNOSTIC MODELING—THE DINA MODEL

Cognitive diagnostic models were developed to provide more targeted information in the form of score profiles that resolve the limitation of IRT models (de la Torre, 2009). Various formations of CDMs have been proposed in the measurement literature. Holistically, these models cover a variety of situations (i.e., types of construct, response, and dimensionality) that would be of interest to researchers in psychometrics and in cognitive and learning sciences. Due to the prevalence of these models there are now studies among researchers to understand and unify similar and related CDMs. For example, the generalized DINA (de la Torre, *in press*) model, the general diagnostic model (von Davier, 2005), and the loglinear cognitive diagnostic model (Burke & Henson, 2008; Henson, Templin, & Willse, 2009) demonstrate this trend in the literature. These efforts delineate the growing popularity and the development among scholars to make CDMs more accessible to applied researchers.

Among CDMs in the literature, the Rule Space Methodology (RSM; Tatsuoka, 1983) has been used to analyze international comparative assessments such as the TIMSS (e.g., Birenbaum, Tatsuoka, & Yamada, 2004; Dogan & Tatsuoka, 2008; Tatsuoka, Corter, & Tatsuoka, 2004; Um et al., 2003). All studies utilized attributes developed by Corter and Tatsuoka (2002; in Tatsuoka et al., 2004) that included 27 attributes from which 6 were content attributes (e.g., algebra, geometry), 10 were

process attributes (e.g., apply rules of algebra, quantitative and logical reading), and 11 were skill attributes (e.g., unit conversion, proportional reasoning). RSM is a pattern recognition and classification method; it assumes that a defined set of attributes is used to form a pattern of attribute mastery, known as knowledge states. Because both attributes and knowledge states are latent in nature, RSM allows a transformation of these variables to its manifest state via attribute mastery probabilities (Birenbaum et al., 2004).

This study focuses on the deterministic, inputs, noisy, “and” gate (DINA; Junker & Sijtsma, 2001) model, which is arguably one of the most parsimonious and interpretable CDMs developed. Regardless of the number of attributes considered in the entire assessment, the DINA model only requires the estimation of two parameters for each item. Furthermore, when attributes required for an item are considered equally important, the DINA model is deemed appropriate. Various studies have investigated the DINA model; a thorough discussion of the DINA model including applications and related latent classification models can be found in de la Torre and Douglas (2008), Doignon and Falmagne (1999), Haertel (1989), Junker and Sijtsma (2001), Macready and Dayton (1977), and Tatsuoka (2002).

The difference between the RSM and DINA is that the latter model incorporates guessing and slip parameters, which can be used to identify items where students do not perform well. For example, when an item has a high parameter value, the item has a high probability that either examinees without mastery of all required attributes guess and correctly solves the problem (i.e., high guessing parameter) or examinees with mastery of all required attributes slip and solve the problem incorrectly (i.e., high slip parameter). These item parameters can be mapped to the Q-matrix, which indicates attributes required to solve the item. As such, educational researchers and classroom teachers can derive diagnostic information for instructional purposes as item parameters relate to attributes specified in the Q-matrix. Furthermore, the DINA item parameters preserves item-level information that can be generated from both CTT and IRT. In a study by de la Torre and Lee (2008), it was found that item parameters from CTT (e.g., difficulty and discrimination) and from IRT (e.g., difficulty, discrimination, and pseudo-guessing) were significantly associated with DINA item guessing and slip parameters.

The construction of a Q-Matrix (Tatsuoka, 1983) is fundamental to the identification of attributes required to answer an item in the DINA modeling framework. If a specific attribute is under- or overspecified, an examinee’s skill set and performance will not correspond with the design of the items, leading to incorrect inferences of model parameters. More formally, the Q-Matrix can be defined as follows. Let X_{ij} be examinee i ’s response for item j , such that $i = \{1, 2, \dots, I\}$ and $j = \{1, 2, \dots, J\}$, and let $\underline{\alpha}_i = \{\alpha_{ik}\}$ be a vector of examinee i ’s skills for attribute k , such that $k = \{1, 2, \dots, K\}$. It follows that the vector $\underline{\alpha}_i$ becomes a binary vector with the value “1,” signaling the presence of an attribute, skill, or cognitive process for the k th element, whereas the value “0” on the k th element

represents the lack of such skill. This framework leads to a binary $J \times K$ matrix, where q_{jk} , the element in the j th row and k th column of the matrix, corresponds to whether the k th skill is required to solve the j th item correctly.

The DINA model consists of two components: (1) a deterministic or error-free process and (2) a stochastic process. The deterministic process is represented by the latent response vector ($\underline{\eta}_i$) that is calculated using the estimated vector $\underline{\alpha}_i$; in other words, $\underline{\eta}_i = \{\eta_{ij}\}$, such that $\eta_{ij} = \prod_{k=1}^K \alpha_{ik}^{q_{jk}}$. The value “1” to the latent response represents examinee i ’s possession of all attributes required to solve item j , whereas the value “0” corresponds to an examinee lacking one of the required attributes for the item, not just any of the K attributes. The DINA model also incorporates a stochastic component in the model, because student responses often carry non-systematic noise. The “and” gate portion of the model derives its name from the conjunctive nature of η_{ij} (de la Torre, 2009). As such, two parameters define the noise aspect: the *slip* and the *guessing* parameters. The DINA model defines the slip parameter as $s_j = P(X_{ij} = 0 | \eta_{ij} = 1)$ and the guessing parameter as $g_j = P(X_{ij} = 1 | \eta_{ij} = 0)$ for item j . Unlike the non-stochastic setting, students who do not possess all the attributes sometimes guess (i.e., *guessing*) and correctly answer an item, and students who possess all the attributes sometimes slip (i.e., *slip*) and incorrectly answer an item. Combining the two parameters, the DINA model calculates the probability that examinee i solves item j correctly given the skills vector $\underline{\alpha}_i$ as

$$P_j(\underline{\alpha}_i) = P(X_{ij} = 1 | \underline{\alpha}_i) = g_j^{1-\eta_{ij}} (1 - s_j)^{\eta_{ij}}. \quad (1)$$

Hence, to answer an item correctly, the equation (1) requires that an examinee possessing all attributes necessary to solve item j as defined by the Q-Matrix to avoid slip, and for an examinee that lacks a specific attribute to guess correctly. Furthermore, if both guessing and slip do not exist, an examinee’s response becomes completely deterministic, and the results rely solely on the interaction between the $\underline{\alpha}$ and the Q-vector for the item (de la Torre, 2009). The estimation of both the guessing and the slip effects can identify inaccurate relationships between a student’s mastery of a required attribute and answering an item. Consequently, mathematics educators and teachers are able to comprehend whether a certain attribute or skill that is required in a given grade level is mastered by a student or a group of students. Details of the EM algorithm for the DINA model—estimation of item parameters and computation of standard error—and implementation of the algorithms are available in de la Torre (2009).

Based on guessing and slip parameters, the DINA-based discrimination index (de la Torre, 2008), $\delta_j = 1 - s_j - g_j$ can be computed in order to represent the difference in probabilities of correct response between $\eta = 0$ and $\eta = 1$. To illustrate the representation of the DINA discrimination index, Figure 1 shows two

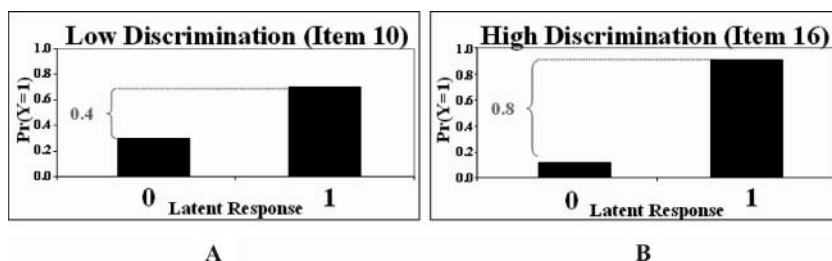


FIGURE 1

Two high- and low-discrimination items from the TIMSS 2007 fourth grade mathematics.

fourth grade mathematics items analyzed in the study: one with low discrimination (Panel A, $\delta = 0.4$ with both guessing and slip parameters estimates 0.2) and the other with high discrimination (Panel B, $\delta = 0.8$ with both guessing and slip parameters estimates 0.1) to contrast the effects. This demonstrates that as discrimination increases, the association between the latent group membership of an examinee and item response becomes robust, because the difference in probability of correct responses between the two latent classification $\eta = 0$ and $\eta = 1$ increases.

METHODS

Data

Data were taken from booklets 4 and 5 of TIMSS 2007 fourth grade mathematics assessment, which consists of 25 items with 15 multiple choice items and 10 constructed response items (Foy & Olson, 2009). The TIMSS releases selected items and groups of examinees in its released data set. Booklets 4 and 5 were chosen because they encompass the greatest number of dichotomously scored items, which the DINA model requires. Consequently, constructed response items with polytomous responses were dichotomized by treating responses with partial credit as incorrect and responses with full credit as correct. Out of the total 25 items used for this study, only 2 items (Item ID M041275 and M031247) were originally scored polytomously with a maximum score of 2 rather than 1; the remaining constructed response items had a maximum score of 1, which was naturally dichotomized. Omitted or unreachable items were also scored as incorrect.

All 11 items released for booklet 4 were new items developed for TIMSS 2007. The remaining 14 items from booklet 5 were previously administered during TIMSS 2003. Each booklet in TIMSS was developed to create a balance with respect to item format as well as content and cognitive domains that are closely tied to the targeted specifications in the TIMSS 2007 framework (Mullis et al.,

2005). The TIMSS mathematics items were designed to reflect the international mathematics curriculum by disseminating surveys to each participating country regarding their assessment objectives and whether they aligned with the design implemented by TIMSS. As a result, TIMSS 2007 mathematics items were developed with two main domains: (1) content domains, which are dedicated to identifying areas or subject matter that evaluates understanding of mathematics; and (2) cognitive domains, which describe the thinking processes that students encounter as they deal with mathematics content. The Number, Geometric Shapes and Measures, and Data Display content domains were each targeted to account for 50%, 35%, and 15% of the assessment, respectively; the Knowing, Applying, and Reasoning cognitive domains were developed to target 40%, 40%, and 20%, respectively. Among the 25 items in booklets 4 and 5, 11 items (44%) targeted the Number; 8 items (32%) targeted the Geometric Shapes and Measures; and 6 items (24%) targeted the Data Display content domains. Likewise, 10 (40%), 10 (40%), and 5 (20%) items used in this study were designed for Knowing, Applying, and Reasoning cognitive domains, respectively. This shows that the booklet selected for this study conforms to the overall criterion created by the test developers.

This study analyzed the fourth grade students' responses from the United States and the two benchmark states to compare DINA parameter estimates and attribute mastery prevalence using the attributes or skills as prescribed by TIMSS (see Table 2). Based on the country code, which includes benchmark identifiers, the data from 823 students were used. This includes 564 students from 255 schools that comprised the US sample, 132 students from 48 schools from Minnesota, and 127 students from 47 schools from Massachusetts.

Analysis

Attributes used to define skills required to solve a particular item in this study were developed based on the 2007 TIMSS Mathematics Framework (Mullis et al., 2005). These formed the basis of the attribute list, because TIMSS test developers had set each item to target a specific content subdomain (i.e., topic areas and objectives); three content domains—Number, Geometric Shapes and Measures, and Data Display—were further described by a number of topic areas, and each topic area was described by a list of objectives that were derived from the mathematics curricula of participating countries (for more details, see Mullis et al., 2005). The TIMSS 2007 framework identified 38 unique objectives; using these objectives, we examined the attributes required for each item. To determine the ultimate set of attributes or skills required, three mathematics educators with mathematics education degrees and teaching experience at the fourth-grade mathematics level and two domain-expert researchers identified the attributes required for each item. They independently coded the 25 items to correspond to the 38 objectives. However, due to the comprehensive nature of the objectives set by TIMSS, some

TABLE 2
Attributes Developed from the 2007 TIMSS Framework for Fourth Grade Mathematics

| Content Domain | Attributes | # of Times Specified |
|-------------------------------------|--|----------------------|
| Number (N) | <i>Whole Numbers (4)</i> | |
| | 1. Representing, comparing, and ordering whole numbers as well as demonstrating knowledge of place value. | 6 |
| | 2. Recognize multiples, computing with whole numbers using the four operations, and estimating computations. | 16 |
| | 3. Solve problems, including those set in real life contexts (for example, measurement and money problems). | 11 |
| | 4. Solve problems involving proportions. | 3 |
| | <i>Fractions and Decimals (2)</i> | |
| | 5. Recognize, represent, and understand fractions and decimals as parts of a whole and their equivalents. | 3 |
| | 6. Solve problems involving simple fractions and decimals including their addition and subtraction. | 2 |
| | <i>Number Sentences with Whole Numbers (1)</i> | |
| | 7. Find the missing number or operation and model simple situations involving unknowns in number sentence or expressions. | 2 |
| Geometric Shapes & Measurement (GM) | <i>Patterns and Relationships (1)</i> | |
| | 8. Describe relationships in patterns and their extensions; generate pairs of whole numbers by a given rule and identify a rule for every relationship given pairs of whole numbers. | 3 |
| | <i>Lines and Angles (1)</i> | |
| | 9. Measure, estimate, and understand properties of lines and angles and be able to draw them. | 3 |
| | <i>Two- and Three-dimensional Shapes (2)</i> | |
| | 10. Classify, compare, and recognize geometric figures and shapes and their relationships and elementary properties. | 7 |
| | 11. Calculate and estimate perimeters, area, and volume. | 2 |
| Data & Display (DD) | <i>Location and Movement (1)</i> | |
| | 12. Locate points in an informal coordinate to recognize and draw figures and their movement. | 3 |
| | <i>Reading and Interpreting (2)</i> | |
| | 13. Read data from tables, pictographs, bar graphs, and pie charts. | 4 |
| | 14. Comparing and understanding how to use information from data. | 3 |
| | <i>Organizing and Representing (1)</i> | |
| | 15. Understanding different representations and organizing data using tables, pictographs, and bar graphs. | 2 |

Note: The italicized headings in the attributes column designates the Topic Areas within the Content Domains as indicated in the 2007 TIMSS framework (Mullis et al., 2005).

fine-grained attributes were never identified in our sample of 25 items. Although these items covered the breadth of content domains, it did not cover all 38 objectives. Therefore, through discussion, the mathematics education researchers and the measurement expert modified or combined the TIMSS objectives so that mathematics educators can deliver reinforced instruction based on findings from this study.

A list of attributes that was required for the development of the Q-matrix for this study is presented in Table 2, which also includes the number of times each attribute was specified in the Q-matrix. Although the list of objectives was combined, the topic areas were preserved. Subsequently, five separate coding sheets from the coders were combined to form a final Q-matrix by discussing discrepancies until they reached agreement.

When an item can be solved using different strategies, the most dominant attributes or skills were used to define the Q-matrix. This process ensured the validity and usefulness of the attributes and the Q-matrix. Furthermore, following the estimation of guessing and slip parameters, items with notably high estimates were reviewed by mathematics education experts and domain-expert researchers to further validate the Q-matrix by assessing whether there were any surplus or shortage of attributes specified. After a thorough examination, content experts concluded that there were no misspecifications of attributes in the Q-matrix—the specified Q-matrix conformed to the necessary topics as indicated by what field experts saw as the most dominant strategy required for solving the items.

The Q-matrix presented in Table 3 was specifically constructed for the TIMSS fourth grade mathematics test used in this study. Note that attributes in the Q-matrix are independently manifested during the process that an item is solved; in other words, each step in solving a problem requires the mastery of a specific attribute that is mapped to the Q-matrix.


Figure 2 shows an exemplar item from the TIMSS 2007 fourth grade mathematics assessment along with its attribute specifications. This item asks a student to find the number of house figures () that represent the number of houses on a street based on the given pictograph. The answer to this problem is “A,” or four house figures. In order to solve this item correctly, the student needs to master four attributes, which were validated to be the most generic and dominant method by the five coders. First, the student should be able to read data from the pictograph, which requires Attribute 13. Next, the student should also be able to understand that each house figure in the pictograph represents five houses (i.e., a whole number), which requires Attribute 1. In addition, the student should be able to compute and find the number of house figures that represent 20 houses on Hill Street, which requires the application of Attribute 2. Finally, the student should also be able to understand that *four house figures* is only a different representation of the 20 houses cited in the problem, which can be solved by Attribute 15.











TABLE 3
TIMSS 2007 Fourth Grade Mathematics Q-matrix



| Item | | Number | | | | | | | | Geometric Shapes & Measures | | | | Data Display | | |
|------|----------|--------|---|---|---|---|---|---|---|--------------------------------|----|----|----|--------------|----|----|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 1 | M041052 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | M041056 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | M041069 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | M041076 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | M041281 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | M041164 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 7 | M041146 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 8 | M041152 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 9 | M041258A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 10 | M041258B | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 11 | M041131 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | M041275 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 13 | M041186 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 14 | M041336 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 15 | M031303 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | M031309 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | M031245 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | M031242A | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | M031242B | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 20 | M031242C | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 21 | M031247 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | M031219 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| 23 | M031173 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24 | M031085 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 25 | M031172 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |

Given the Q-matrix, the data were fit separately for the US and the two benchmark states using the EM algorithm of the DINA model to estimate both guessing and slip parameters and attribute mastery prevalence. The computer program Ox (Doornik, 2002), available free of charge for academic research and teaching purposes, was implemented to run the EM algorithm for the DINA model, which was originally developed in de la Torre (2009).

Model fit was also evaluated using information criteria measures that used the estimated log-likelihood values for the US and its regional entities. These statistics were compared to the 1PL, 2PL, and the 3PL item response theory (IRT) models; items used in each administration of the TIMSS assessment follow a rigorous IRT-based calibration and scaling methodology as outlined in the *TIMSS 2007 Technical Report* (Olson et al., 2009, p. 193). As such, general assumptions

Item 25: M031172 (Data Display: Organizing & Representing)

| Street | Number of houses |
|--------|---|
| Main |      |
| Center |   |
| First |    |
| Hill | |

Mary is making a chart to show the number of houses on some streets.
Every  stands for 5 houses. There are 20 houses on Hill Street.
How many  should Mary put in the chart beside Hill Street?

- (A) 4
- (B) 5
- (C) 15
- (D) 20

Copyright © 2008 International Association for the Evaluation of Educational Achievement (IEA). All rights reserved.

| Item | Number | | | | | | | | Geometric Shapes & Measures | | | | Data Display | | |
|------|--------|---|---|---|---|---|---|---|--------------------------------|----|----|----|-----------------|----|----|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 25 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |

FIGURE 2
An item from TIMSS 2007 fourth grade mathematics.

of IRT models including conditional local independence have been previously investigated and satisfied (Olson et al., 2009, p. 228). Comparing IRT models to the DINA model examines a method of evaluating the model fit from examining the data using the current model to the model used in this study; a better fit to the model implies that the selected model explains the data better.

In addition to the DINA model parameters (i.e., guessing and slip) and the attribute mastery level, proportion mastered (η_j ; proportion of masters) was compared to proportion correct for each item. Proportion mastered was derived from

the DINA model as a byproduct of the EM algorithm (de la Torre, 2009), which is the posterior probability of mastering all attributes required for the item; this posterior probability represents the proportion of students that mastered all attributes specified in the Q-matrix. As noted earlier, if a student lacked the mastery of a single attribute, η_{ij} would be 0, and if a student mastered all the required attributes, η_{ij} would be 1. Proportion mastered examines the proportion of all examinees that mastered the required attributes. In this study, the proportion mastered statistic is contrasted to the proportion correct statistic, which refers to the proportion of the population that answered the item correctly.

Detailed comparisons between the US and the two benchmark states—Massachusetts and Minnesota—were made for items that revealed differences in parameter estimates in order to identify strengths and weaknesses in performance of test takers. In addition to the comparison of item parameter estimates, proportions of attribute prevalence (i.e., attribute mastery level) between the US and the two benchmark states obtained from the DINA model were statistically tested; attribute prevalence differs from proportion mastered in that the former refers to the proportion of individuals who have mastered a particular attribute. Finally, a CDM-based distractor analysis was conducted to evaluate the utility of attribute mastery on distractors and to understand their interaction. This information can help mathematics educators and researchers determine areas that lack the mastery of specific fine-grained attributes, which can be used to examine students that may have endorsed the wrong response options and to enhance the instruction of specific attributes required to solve an item correctly.

RESULTS

To make inferences about the results derived from the DINA model, an evaluation of model fit is presented between IRT models and the DINA model. Such comparison between IRT models and the DINA model is meaningful, because as noted earlier, TIMSS uses the former to calibrate and scale student performance. Moreover, examining how the DINA model performs in relation to IRT provides credibility to the interpretation and meaning of the results. Table 4 shows the comparison of the DINA model to 1PL, 2PL, and 3PL IRT models via information criteria-based statistics.

Results showed that the DINA model had a better model fit for all three regions as noted by the lower *AIC* and *BIC* statistics; that is, the DINA model was preferred over the IRT models. The better fit of the DINA model provided grounds to further examine the results.

TABLE 4
Information Criteria for Model Comparison Between IRT and DINA Models

| Region | Model | <i>−2LL</i> | <i>AIC</i> | <i>BIC</i> |
|---------------|-------|-------------|------------|------------|
| United States | 1PL | 15815.14 | 15867.14 | 15979.85 |
| | 2PL | 15575.52 | 15675.52 | 15892.27 |
| | 3PL | 15526.72 | 15676.72 | 16001.85 |
| | DINA | 15532.67 | 15632.67 | 15849.42 |
| Massachusetts | 1PL | 3323.34 | 3375.34 | 3449.29 |
| | 2PL | 3249.57 | 3349.57 | 3491.78 |
| | 3PL | 3216.34 | 3366.34 | 3579.65 |
| | DINA | 3145.51 | 3245.51 | 3387.72 |
| Minnesota | 1PL | 3448.04 | 3500.04 | 3574.99 |
| | 2PL | 3396.07 | 3496.07 | 3640.21 |
| | 3PL | 3366.52 | 3516.52 | 3732.73 |
| | DINA | 3313.03 | 3413.03 | 3557.17 |

Proportion Correct and Proportion Mastered

Figure 3 illustrates the proportion of correctly answered items, which indicates the performance for the 25 items used for this study. The proportion of examinees that mastered all required attributes as defined by the Q-matrix is also included in the figure. Figure 3 shows a disparity between proportion correct and proportion mastered—some items exhibited greater proportion correct than proportion mastered and others had greater proportion mastered than proportion correct.

The average proportion correct for Massachusetts, Minnesota, and the United States was 0.673, 0.638, and 0.571, respectively. However, the average proportion mastered for the three regions was 0.557, 0.610, and 0.602, respectively—the proportion correct was highest for Massachusetts, yet its proportion mastered was the lowest, and in contrast, the United States had the lowest proportion correct, but its proportion mastered was greater than Massachusetts. Figure 3 shows an item-by-item contrast between these two proportions that relates to the efficacy and the motivation of this study. In the Number domain, the proportion correct statistic was greater than the proportion mastered for items 1, 5, 15, and 18; however, for items 2, 17, and 21, proportion mastered was greater than proportion correct. For items 3, 4, 16, and 23, the proportions between mastery and correct were very similar. In the Geometric Shapes and Measures domain, only item 6 had proportion correct that was greater than proportion mastered for all three regions; for items 11, 22, and 24, proportion mastered was greater than proportion correct, and for items 7, 8, 9, and 10, the proportions were similar. Finally, in the Data Display domain, items 12, 13, 14, 20, and 25 had higher proportion correct than mastered,

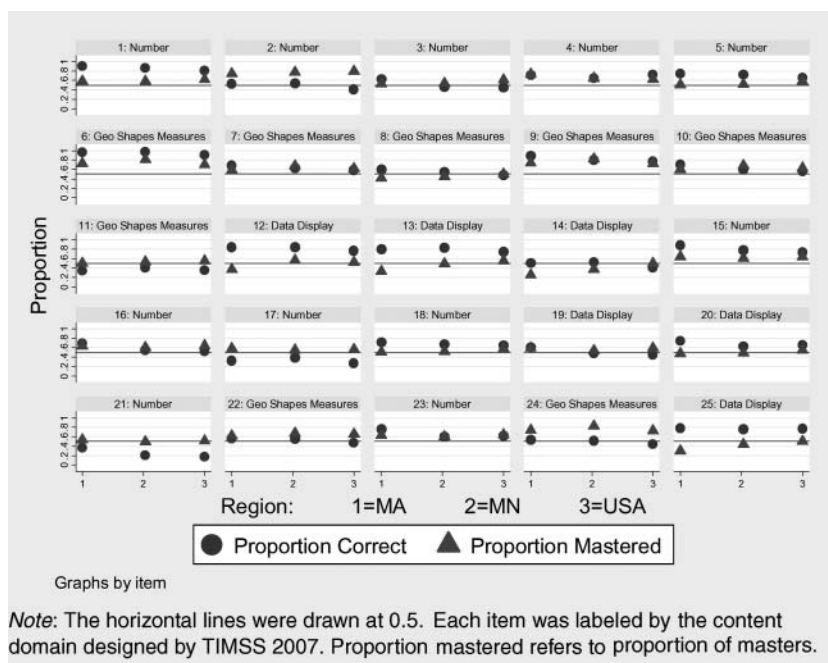


FIGURE 3
Proportion correct and proportion mastered.

and the proportion mastered was greater than proportion correct for item 14 of the US cohort; only item 19 had similar proportions.

From the Number domain, 36% of the items were high guessing items and 27% of the items were high slip items; from the Geometric Shapes and Measures domain, only 13% of the items were high guessing, and the remaining 38% were high slip; finally, from the Data Display domain, 67% of the items were high guessing, and only 17% were high slip items. This showed that in general, there was a greater tendency for guessing in the Number and Data Display domains and a higher tendency for slip in the Geometric Shapes and Measures domain.

To capture the contrast in performance between the three regions, Table 5 shows a comparison of the proportion correct and the proportion mastered using two separate logistic regressions. Two helmert-contrast variables were created; the first variable contrasted the combined effect of the two benchmark participants (positive) to the US (negative), and the second variable contrasted the effect among the two benchmark participants (MA: positive vs. MN: negative), excluding the effect of the US. Therefore, an odds ratio greater than one signaled greater odds for the positively coded (i.e., benchmark participants or Massachusetts) region to

TABLE 5
Logistic Regression to Contrast the Performance and the Mastery for the Three Regions

| Item | Cognitive Domain | Parameters | Correct | | Mastery | |
|------|-------------------------------|------------------|------------|-------|------------|-------|
| | | | Odds Ratio | SE | Odds Ratio | SE |
| 1 | Number | Benchmark vs. US | 1.278** | 0.097 | 0.966 | 0.049 |
| | | MA vs. MN | 1.150 | 0.232 | 0.968 | 0.122 |
| 2 | Number | Benchmark vs. US | 1.204*** | 0.061 | 0.952 | 0.057 |
| | | MA vs. MN | 0.936 | 0.117 | 0.856 | 0.127 |
| 3 | Number | Benchmark vs. US | 1.144** | 0.058 | 0.916 | 0.046 |
| | | MA vs. MN | 1.343* | 0.170 | 0.936 | 0.117 |
| 4 | Number | Benchmark vs. US | 0.961 | 0.052 | 1.117* | 0.060 |
| | | MA vs. MN | 1.103 | 0.148 | 1.189 | 0.162 |
| 5 | Number | Benchmark vs. US | 1.144* | 0.064 | 0.935 | 0.047 |
| | | MA vs. MN | 1.014 | 0.143 | 0.949 | 0.118 |
| 6 | Geometric Shapes and Measures | Benchmark vs. US | 1.656** | 0.255 | 1.159* | 0.069 |
| | | MA vs. MN | 0.688 | 0.301 | 0.740* | 0.114 |
| 7 | Geometric Shapes and Measures | Benchmark vs. US | 1.120* | 0.059 | 1.041 | 0.054 |
| | | MA vs. MN | 1.097 | 0.145 | 0.806 | 0.105 |
| 8 | Geometric Shapes and Measures | Benchmark vs. US | 1.149** | 0.058 | 0.925 | 0.047 |
| | | MA vs. MN | 1.099 | 0.139 | 0.899 | 0.113 |
| 9 | Geometric Shapes and Measures | Benchmark vs. US | 1.222** | 0.084 | 1.147* | 0.070 |
| | | MA vs. MN | 1.339 | 0.243 | 0.734* | 0.115 |
| 10 | Geometric Shapes and Measures | Benchmark vs. US | 1.159** | 0.061 | 1.040 | 0.054 |
| | | MA vs. MN | 1.218 | 0.161 | 0.805 | 0.105 |
| 11 | Geometric Shapes and Measures | Benchmark vs. US | 1.046 | 0.054 | 0.967 | 0.049 |
| | | MA vs. MN | 0.834 | 0.107 | 0.906 | 0.113 |
| 12 | Data Display | Benchmark vs. US | 1.211** | 0.081 | 0.948 | 0.048 |
| | | MA vs. MN | 0.948 | 0.165 | 0.659** | 0.084 |
| 13 | Data Display | Benchmark vs. US | 1.183** | 0.075 | 0.842** | 0.043 |
| | | MA vs. MN | 0.854 | 0.141 | 0.692** | 0.089 |
| 14 | Data Display | Benchmark vs. US | 1.185** | 0.060 | 0.785*** | 0.042 |
| | | MA vs. MN | 0.934 | 0.116 | 0.735* | 0.099 |
| 15 | Number | Benchmark vs. US | 1.250** | 0.083 | 0.988 | 0.051 |
| | | MA vs. MN | 1.386 | 0.242 | 1.036 | 0.133 |
| 16 | Number | Benchmark vs. US | 1.168** | 0.060 | 0.988 | 0.051 |
| | | MA vs. MN | 1.310* | 0.171 | 1.036 | 0.133 |
| 17 | Number | Benchmark vs. US | 1.163** | 0.062 | 1.021 | 0.052 |
| | | MA vs. MN | 0.858 | 0.111 | 1.030 | 0.130 |
| 18 | Number | Benchmark vs. US | 1.100 | 0.059 | 0.935 | 0.047 |
| | | MA vs. MN | 1.067 | 0.145 | 0.949 | 0.118 |
| 19 | Data Display | Benchmark vs. US | 1.143** | 0.058 | 0.983 | 0.050 |
| | | MA vs. MN | 1.241 | 0.156 | 1.097 | 0.138 |
| 20 | Data Display | Benchmark vs. US | 1.074 | 0.058 | 0.935 | 0.047 |
| | | MA vs. MN | 1.255 | 0.171 | 0.961 | 0.120 |
| 21 | Number | Benchmark vs. US | 1.274*** | 0.076 | 1.012 | 0.051 |
| | | MA vs. MN | 1.390* | 0.193 | 1.057 | 0.132 |

(Continued)

TABLE 5
(Continued)

| Item | Cognitive Domain | Parameters | Correct | | Mastery | |
|------|-------------------------------|------------------|------------|-------|------------|-------|
| | | | Odds Ratio | SE | Odds Ratio | SE |
| 22 | Geometric Shapes and Measures | Benchmark vs. US | 1.146** | 0.058 | 1.016 | 0.054 |
| | | MA vs. MN | 0.982 | 0.123 | 0.861 | 0.113 |
| 23 | Number | Benchmark vs. US | 1.135* | 0.061 | 0.988 | 0.051 |
| | | MA vs. MN | 1.404* | 0.193 | 1.036 | 0.133 |
| 24 | Geometric Shapes and Measures | Benchmark vs. US | 1.138* | 0.057 | 1.147* | 0.070 |
| | | MA vs. MN | 1.010 | 0.126 | 0.734* | 0.115 |
| 25 | Data Display | Benchmark vs. US | 1.034 | 0.062 | 0.857** | 0.044 |
| | | MA vs. MN | 1.044 | 0.157 | 0.718* | 0.093 |

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

get the particular item correct or master all required attributes for the particular item.

The results showed an unintuitive pattern. For nearly half of all items (i.e., items 1, 2, 3, 5, 7, 8, 10, 15, 16, 17, 19, 21, and 22), there were statistically significant differences in the odds that the contrasted region answered an item correctly, but there were no significant differences in the odds for mastering the attributes required to solve the item. Similarly, for items 4 and 25, there were significant differences in attribute mastery ($OR = 1.117$ and 0.857 , $p < 0.05$, respectively) among the contrasted regions, whereas in their performance, there were no significant differences. Item 4 targets the Number domain, and its attribute mastery was significantly greater for the benchmark countries than the US.

Likewise, for items 6, 9, 12, 13, 14, and 24, there were significant differences in attribute mastery, but not in their performance between Massachusetts and Minnesota. For item 25, which targets the Data Display domain, both the US and Minnesota had significantly greater attribute mastery than the benchmark participants and Massachusetts, respectively, whereas there were no differences in their performance. Here, an odds ratio less than one signals greater odds for the negatively coded region (i.e., the US or Minnesota) to get a particular item correct or master all required attributes for the particular item. In addition, for items 13 and 14 of the Data Display content domain, the US had greater odds of attribute mastery ($OR = 0.842$ and 0.785 , $p < 0.05$, respectively) than the benchmark participants, yet the odds of correctly answering the item ($OR = 1.183$ and 1.185 , $p < 0.05$, respectively) was significantly greater than the US.

DINA Model Item Parameters

The DINA item parameters represent the guessing and the slip probabilities for each item; in addition, the guessing and the slip parameter estimates can be used to calculate the discrimination parameter—the probability of correctly solving an item without the effect of guessing and slip (i.e., $\delta = 1 - g - s$). Table 6 presents the estimates for the guessing, slip, and discrimination parameters. However, to ease contrast of parameter estimates and multiple comparisons, Figure 4 was created to show the Bonferroni-corrected guessing and slip confidence intervals, which extended to 98.3% (i.e., derived using the alpha level of $0.05/3 = 0.0167$). This correction was made to test three hypotheses simultaneously—Massachusetts versus Minnesota, Massachusetts versus the United States, and Minnesota versus the United States. Although they convey the same information, the visualization of both guessing and slip parameters on the same graph in Figure 4 allows readers to easily grasp the item characteristics from a CDM perspective.

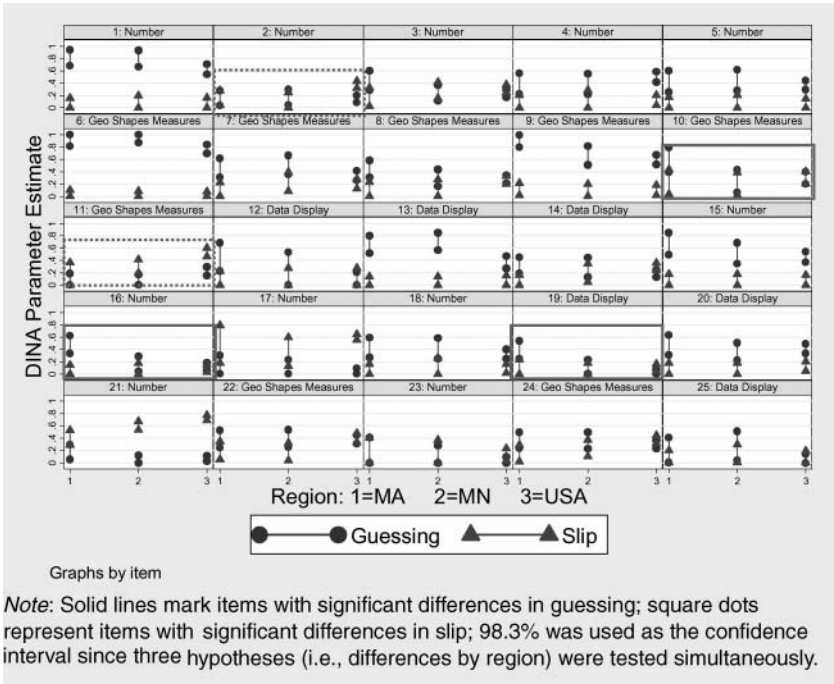


FIGURE 4
Bonferroni-adjusted confidence intervals for DINA parameter estimates.

TABLE 6
DINA Item Parameter Estimates and SEs

| Item | Content Domain | MA | | MN | | U.S. | | MA | | MN | | US | | MA | | MN | | US | |
|------|----------------|----------|----------------|----------|----------------|----------|----------------|----------|----------------|----------|----------------|----------|----------------|----------|----------------|----------|----------------|----------|----------------|
| | | <i>g</i> | SE(<i>g</i>) | <i>g</i> | SE(<i>g</i>) | <i>g</i> | SE(<i>g</i>) | <i>s</i> | SE(<i>s</i>) | <i>s</i> | SE(<i>s</i>) | <i>s</i> | SE(<i>s</i>) | <i>s</i> | SE(<i>s</i>) | <i>s</i> | SE(<i>s</i>) | <i>s</i> | SE(<i>s</i>) |
| 1 | N | 0.808 | 0.055 | 0.794 | 0.056 | 0.621 | 0.034 | 0.020 | 0.019 | 0.058 | 0.029 | 0.082 | 0.016 | 0.172 | 0.017 | 0.147 | 0.018 | 0.297 | 0.019 |
| 2 | N | 0.154 | 0.051 | 0.169 | 0.053 | 0.142 | 0.024 | 0.156 | 0.047 | 0.120 | 0.040 | 0.373 | 0.029 | 0.690 | 0.029 | 0.711 | 0.030 | 0.486 | 0.031 |
| 3 | N | 0.443 | 0.064 | 0.235 | 0.055 | 0.227 | 0.027 | 0.172 | 0.055 | 0.284 | 0.058 | 0.307 | 0.031 | 0.385 | 0.031 | 0.481 | 0.032 | 0.466 | 0.033 |
| 4 | N | 0.388 | 0.069 | 0.384 | 0.067 | 0.496 | 0.034 | 0.034 | 0.026 | 0.122 | 0.039 | 0.118 | 0.020 | 0.578 | 0.020 | 0.494 | 0.021 | 0.386 | 0.021 |
| 5 | N | 0.425 | 0.071 | 0.445 | 0.068 | 0.368 | 0.032 | 0.000 | 0.111 | 0.041 | 0.024 | 0.068 | 0.018 | 0.575 | 0.018 | 0.514 | 0.019 | 0.565 | 0.020 |
| 6 | GM | 0.913 | 0.042 | 0.952 | 0.034 | 0.769 | 0.030 | 0.000 | 0.116 | 0.000 | 0.112 | 0.006 | 0.005 | 0.087 | 0.005 | 0.048 | 0.006 | 0.225 | 0.007 |
| 7 | GM | 0.462 | 0.064 | 0.512 | 0.065 | 0.338 | 0.032 | 0.071 | 0.037 | 0.238 | 0.053 | 0.206 | 0.026 | 0.467 | 0.026 | 0.250 | 0.027 | 0.456 | 0.028 |
| 8 | GM | 0.448 | 0.057 | 0.300 | 0.059 | 0.277 | 0.027 | 0.098 | 0.052 | 0.129 | 0.048 | 0.267 | 0.030 | 0.455 | 0.030 | 0.571 | 0.031 | 0.455 | 0.032 |
| 9 | GM | 0.895 | 0.040 | 0.663 | 0.063 | 0.597 | 0.033 | 0.116 | 0.042 | 0.050 | 0.028 | 0.101 | 0.018 | 0.000 | 0.018 | 0.287 | 0.019 | 0.301 | 0.020 |
| 10 | GM | 0.589 | 0.083 | 0.245 | 0.077 | 0.299 | 0.040 | 0.234 | 0.048 | 0.202 | 0.045 | 0.305 | 0.026 | 0.177 | 0.026 | 0.553 | 0.027 | 0.396 | 0.028 |
| 11 | GM | 0.000 | 0.079 | 0.058 | 0.047 | 0.225 | 0.029 | 0.179 | 0.084 | 0.308 | 0.060 | 0.532 | 0.032 | 0.821 | 0.032 | 0.634 | 0.033 | 0.243 | 0.034 |
| 12 | DD | 0.462 | 0.095 | 0.258 | 0.115 | 0.019 | 0.081 | 0.000 | 0.026 | 0.000 | 0.018 | 0.087 | 0.017 | 0.538 | 0.017 | 0.742 | 0.018 | 0.894 | 0.019 |
| 13 | DD | 0.661 | 0.059 | 0.709 | 0.058 | 0.370 | 0.042 | 0.000 | 0.141 | 0.000 | 0.038 | 0.057 | 0.015 | 0.339 | 0.015 | 0.291 | 0.016 | 0.574 | 0.017 |
| 14 | DD | 0.322 | 0.053 | 0.289 | 0.063 | 0.191 | 0.024 | 0.036 | 0.036 | 0.201 | 0.059 | 0.304 | 0.034 | 0.642 | 0.034 | 0.510 | 0.035 | 0.505 | 0.036 |
| 15 | N | 0.672 | 0.075 | 0.520 | 0.071 | 0.457 | 0.037 | 0.000 | 0.107 | 0.009 | 0.017 | 0.073 | 0.017 | 0.328 | 0.017 | 0.471 | 0.018 | 0.470 | 0.019 |
| 16 | N | 0.477 | 0.060 | 0.167 | 0.051 | 0.120 | 0.025 | 0.000 | 0.133 | 0.052 | 0.030 | 0.095 | 0.021 | 0.523 | 0.021 | 0.782 | 0.022 | 0.785 | 0.023 |
| 17 | N | 0.000 | 0.126 | 0.000 | 0.097 | 0.043 | 0.019 | 0.485 | 0.060 | 0.360 | 0.057 | 0.598 | 0.027 | 0.515 | 0.027 | 0.640 | 0.028 | 0.359 | 0.029 |
| 18 | N | 0.432 | 0.066 | 0.410 | 0.069 | 0.325 | 0.032 | 0.000 | 0.058 | 0.091 | 0.038 | 0.089 | 0.019 | 0.568 | 0.019 | 0.499 | 0.020 | 0.586 | 0.021 |
| 19 | DD | 0.392 | 0.061 | 0.117 | 0.048 | 0.000 | 0.027 | 0.085 | 0.050 | 0.063 | 0.039 | 0.098 | 0.026 | 0.523 | 0.026 | 0.821 | 0.027 | 0.902 | 0.028 |
| 20 | DD | 0.471 | 0.067 | 0.366 | 0.060 | 0.412 | 0.032 | 0.016 | 0.017 | 0.029 | 0.028 | 0.126 | 0.022 | 0.513 | 0.022 | 0.605 | 0.023 | 0.462 | 0.024 |
| 21 | N | 0.172 | 0.048 | 0.052 | 0.029 | 0.071 | 0.016 | 0.410 | 0.069 | 0.600 | 0.061 | 0.732 | 0.028 | 0.419 | 0.028 | 0.348 | 0.029 | 0.197 | 0.030 |
| 22 | GM | 0.386 | 0.061 | 0.397 | 0.059 | 0.375 | 0.029 | 0.201 | 0.060 | 0.182 | 0.057 | 0.403 | 0.035 | 0.413 | 0.035 | 0.421 | 0.036 | 0.222 | 0.037 |
| 23 | N | 0.000 | 0.169 | 0.000 | 0.116 | 0.000 | 0.040 | 0.000 | 0.097 | 0.093 | 0.035 | 0.132 | 0.020 | 1.000 | 0.020 | 0.907 | 0.021 | 0.868 | 0.022 |
| 24 | GM | 0.366 | 0.056 | 0.361 | 0.057 | 0.290 | 0.027 | 0.154 | 0.064 | 0.236 | 0.061 | 0.381 | 0.033 | 0.480 | 0.033 | 0.404 | 0.034 | 0.329 | 0.035 |
| 25 | DD | 0.204 | 0.083 | 0.276 | 0.099 | 0.000 | 0.059 | 0.000 | 0.033 | 0.058 | 0.029 | 0.048 | 0.013 | 0.796 | 0.013 | 0.666 | 0.014 | 0.951 | 0.015 |

Note: Acronyms for the content domain represents the following: Number (N), Geometric Shapes and Measures (GM), and Data Display (DD).

In general, the results showed four general patterns of items for further scrutiny: high-guessing, high-slip, high-guess-high-slip, and low-guess-low-slip items. Item 1 (Number domain) can be classified as a high-guessing item. Massachusetts and Minnesota had guessing parameter estimates of 0.808 and 0.794, respectively, whereas the US had an estimate of 0.621. For this item, the slip parameters were low with estimates of 0.020, 0.058, and 0.082, respectively. On the other hand, item 17 (Number domain) can be classified as a high-slip item, with estimates of 0.485, 0.360, and 0.598, respectively, while the guessing parameters were low with estimates of 0.000, 0.000, and 0.043, respectively. There were also items with equally high guessing and slip estimates (i.e., high-guess-high-slip items).

Item 10 (Geometric Shapes and Measures domain) had guessing estimates of 0.589, 0.245, and 0.299, respectively, while its slip parameters were 0.234, 0.202, and 0.305, respectively. Finally, there were items with low-guessing and low-slip estimates (i.e., low-guess-low-slip item). Item 23 (Number domain) had guessing parameter estimates of 0.000, 0.000, and 0.000, respectively, and its slip parameters were 0.000, 0.093, and 0.132, respectively.

Figure 4 shows items that have significantly different guessing and slip parameter estimates from other regions. In items 10 (Geometric Shapes and Measures domain), 16 (Data Display domain), and 19 (Data Display domain), Massachusetts had a significantly greater item guessing estimate than both Minnesota and the US. Furthermore, in item 13 (Data Display domain), the US had a significantly lower item guessing estimate than the two benchmark participants. For item slip, the US had a significantly higher item slip than Minnesota and Massachusetts for items 2 (Number domain) and 11 (Geometric Shapes and Measures domain).

Figure 5 illustrates the item discrimination estimates presented in Table 6 to aid visualization of the variability in parameter estimates. Items with high discrimination represents a greater difference in probabilities of correct responses between examinees that were classified to $\eta = 1$ and $\eta = 0$. Among the 25 items, items 1 (Number domain), 6 (Geometric Shapes and Measures domain), and 9 (Geometric Shapes and Measures domain) had low discrimination indices due to their high guessing parameter estimates; high guessing indicates that the students did not master a required attribute for this item. For comparison between regions, Massachusetts has a notably lower discrimination than both the US and Minnesota for items 9 (Geometric Shapes and Measures domain) and 19 (Data Display domain). The low discrimination in Massachusetts can be attributed to its high guessing relative to other regions, which were 0.895 and 0.392 for items 9 and 19, respectively; Minnesota had guessing estimates of 0.663 and 0.117, respectively, and the US had estimates of 0.597 and 0.000, respectively. For item 11 (Geometric Shapes and Measures domain), the US had a significantly lower item discrimination than both benchmark participants, which can be attributed to

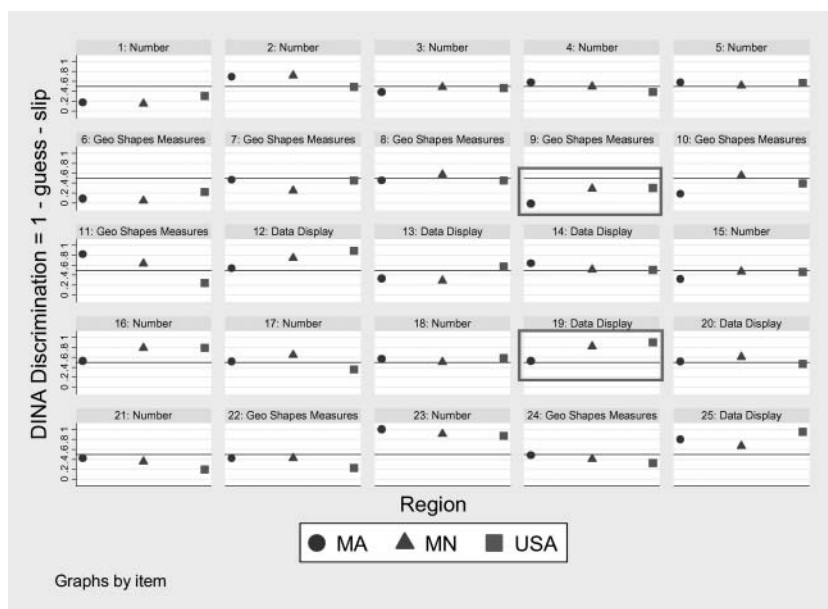


FIGURE 5
DINA discrimination parameters.

the high slip estimate of 0.532, compared to 0.179 and 0.308 for Massachusetts and Minnesota, respectively.

The results from Table 6 can be accompanied with Table 5, because it explains the disparity between an examinee's performance and their mastery of skills. For example, in item 1 (Number domain), it was noted that there were significant differences between the performance of the US and the two benchmark participants; however, there were no differences in their mastery (see Table 4), because the guessing estimate for the two benchmark participants were higher than the US as shown in Table 6.

Similarly, for items 13 (Data Display domain) and 25 (Data Display domain), the lower-performing regions had a significantly greater proportion of attribute mastery. In other words, the US had a higher proportion of attribute mastery than the benchmark participants, and Minnesota had a higher proportion of attribute mastery than Massachusetts. However, even with low attribute mastery, Massachusetts was able to attain a higher performance, because the items were not difficult and students were able to guess to get them correctly, which was shown from previously obtained estimates. Likewise, for items with higher proportion mastered than proportion correct, the estimates for the slip parameter in items 2

(Number domain) and 21 (Number domain) of the US were notably large; they also have a greater or almost equal proportion mastered when compared to the benchmark participants. In addition to cross referencing item parameter estimates to proportion correct and proportion mastered, the information gathered can again be compared to the attribute list specified by the Q-matrix in Table 3. Items 8 and 22 require Attribute 11, and were designed to target the Geometric Shapes and Measures domain. However, for both items, the proportion mastered was higher than the proportion correct for the US, and the slip parameter was also higher than the other two benchmark participants. Although American students in general mastered the skill to calculate and estimate perimeters, area, and volume (Attribute 11), they tended to slip more than students in Massachusetts and Minnesota.

Attribute Prevalence and Attribute Pattern Probability

The DINA model also provides a platform to estimate attribute prevalence in the population. Figure 6 presents the prevalence of attributes mastery of the 15

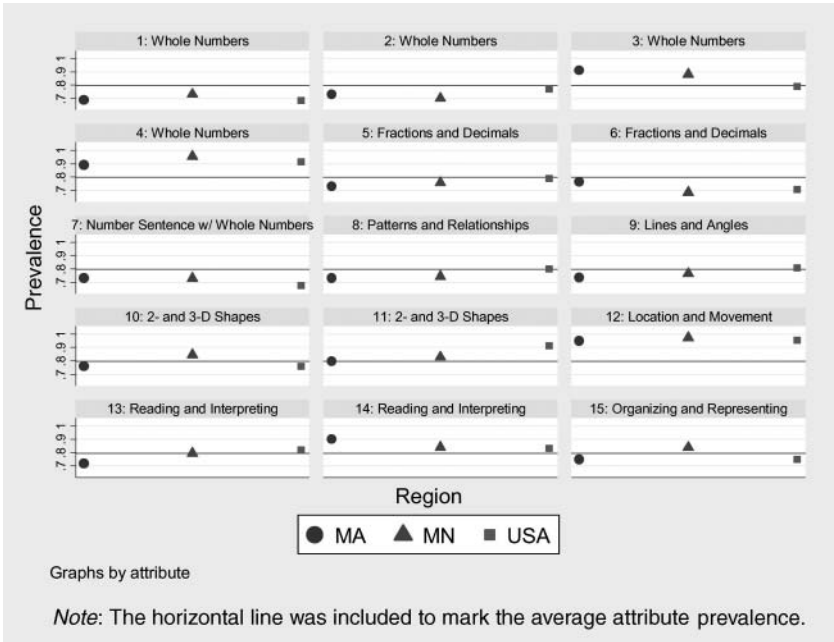


FIGURE 6
Attribute prevalence.

attributes labeled by the topic areas derived from the Q-matrix. Among the 15 attributes, the US had a higher attribute prevalence for Attributes 2, 5, 8, 9, 11, and 13; Minnesota had a higher attribute prevalence for Attributes 1, 4, 10, 12, and 15; and Massachusetts had a higher attribute prevalence for Attributes 3, 6, 7, and 14. This was a notable result, because Massachusetts had the highest overall scaled score, yet it had the lowest number of dominating attributes.

To compare the attribute prevalence between the US and two benchmark states, a logistic regression was used with helmert-contrast variables, which compared the benchmark participants to the US and also compared the effects between Massachusetts and Minnesota, without the effect of the US. The results in Table 7 show that attribute prevalence for the combined effect of the benchmark participants was significantly greater for Attributes 1, 3, 7, 10, and 15. However, for Attributes 11 (calculate and estimate perimeters, area, and volume) and 13 (read data from tables, pictographs, bar graphs, and pie charts), the US had a significantly greater prevalence in attribute mastery than both Massachusetts and Minnesota. Furthermore, for items 4 and 10, Minnesota had a greater attribute prevalence than Massachusetts.

Based on attribute mastery, latent class sizes were also calculated. Since there were 15 attributes used for this study, there were $2^{15} = 32,768$ possible combinations for the mastery of each attribute. Table 8 shows the latent classes and their sizes categorized by the three content domains in the TIMSS 2007 assessment. There were eight attributes in the Number domain; four attributes in the Geometric Shapes and Measures domain; and three attributes in the Data Display domain. Consequently, there were 256, 16, and 8 combinations, respectively. The results from the previous sections examined each attribute independently, whereas this table shows the mastery patterns by content domain. Table 8 shows the top 10 ranked latent classes.

In the latent class column, a binary indicator was used to sequentially specify whether the latent class mastered the attribute or not (e.g., 01010000 indicates a class that mastered Attributes 2 and 4 only). The results show that students in the US had a comparable latent class size to the other benchmark participants. In fact, for the size of content mastery of Number, Geometric Shapes and Measures, and Data Display, it had the highest latent class size of 0.415, 0.411, and 0.457. This was in contrast to the sizes from the benchmark participants, which were 0.284, 0.252, and 0.362 for Massachusetts, and 0.364, 0.417, and 0.364 for Minnesota, respectively.

A comparison between the benchmark participants showed that Minnesota had a greater attribute prevalence for Attributes 4 (solve problems involving proportions) and 10 (classify, compare, and recognize geometric figures and shapes and their relationships and elementary properties), while Massachusetts only had a greater attribute prevalence for Attribute 14 (comparing and understanding how to use information from data). Items 3, 11, and 13 require Attribute 4, and items 6, 7, 8,

TABLE 7
Logistic Regression to Contrast Attribute Prevalence

| Attribute | Topic Areas | Parameter | Odds Ratio | SE |
|-----------|-------------------------------------|------------------|------------|-------|
| 1 | Whole Numbers | Benchmark vs. US | 1.116* | 0.062 |
| | | MA vs. MN | 0.884 | 0.123 |
| 2 | Whole Numbers | Benchmark vs. US | 0.942 | 0.053 |
| | | MA vs. MN | 1.071 | 0.148 |
| 3 | Whole Numbers | Benchmark vs. US | 1.221** | 0.094 |
| | | MA vs. MN | 1.150 | 0.232 |
| 4 | Whole Numbers | Benchmark vs. US | 1.149 | 0.116 |
| | | MA vs. MN | 0.564* | 0.151 |
| 5 | Fractions and Decimals | Benchmark vs. US | 0.952 | 0.057 |
| | | MA vs. MN | 0.856 | 0.127 |
| 6 | Fractions and Decimals | Benchmark vs. US | 1.043 | 0.058 |
| | | MA vs. MN | 1.228 | 0.172 |
| 7 | Number Sentences with Whole Numbers | Benchmark vs. US | 1.151* | 0.064 |
| | | MA vs. MN | 1.034 | 0.145 |
| 8 | Patterns and Relationships | Benchmark vs. US | 0.934 | 0.055 |
| | | MA vs. MN | 0.955 | 0.138 |
| 9 | Lines and Angles | Benchmark vs. US | 0.937 | 0.057 |
| | | MA vs. MN | 0.913 | 0.135 |
| 10 | Two- and Three-dimensional shapes | Benchmark vs. US | 1.147* | 0.070 |
| | | MA vs. MN | 0.734* | 0.115 |
| 11 | Two- and Three-dimensional shapes | Benchmark vs. US | 0.779** | 0.057 |
| | | MA vs. MN | 0.951 | 0.156 |
| 12 | Location and Movement | Benchmark vs. US | 0.862 | 0.129 |
| | | MA vs. MN | 0.631 | 0.222 |
| 13 | Reading and Interpreting | Benchmark vs. US | 0.819** | 0.054 |
| | | MA vs. MN | 0.848 | 0.130 |
| 14 | Reading and Interpreting | Benchmark vs. US | 1.136 | 0.095 |
| | | MA vs. MN | 1.630* | 0.357 |
| 15 | Organizing and Representing | Benchmark vs. US | 1.196** | 0.078 |
| | | MA vs. MN | 0.758 | 0.129 |

Note: * $p < 0.05$, ** $p < 0.01$.

9, 10, 22, and 24 require Attribute 10. More specifically, item 3 was designed by TIMSS to target the Number domain, which is classified under Attribute 4.

However, even though Minnesota had a significantly greater attribute prevalence for this attribute, the odds that students in Massachusetts answered this item correctly was 1.343 times greater than students from Minnesota, which shows a high-level of guessing. In addition, although Massachusetts had a significantly higher prevalence than Minnesota for Attribute 14 there were no significant differences in student performance between the benchmark participants among items 14, 19, and 20.

TABLE 8
Top 10 Ranked Latent Classes

| Rank | Massachusetts | | Minnesota | | United States | |
|--|---------------|-------|--------------|-------|---------------|-------|
| | Latent Class | Size | Latent Class | Size | Latent Class | Size |
| Number (8 attributes) | | | | | | |
| 1 | 11111111 | 0.284 | 11111111 | 0.364 | 11111111 | 0.415 |
| 2 | 01010000 | 0.071 | 01000110 | 0.099 | 01000000 | 0.082 |
| 3 | 01010110 | 0.063 | 01000100 | 0.068 | 00000000 | 0.053 |
| 4 | 00010100 | 0.055 | 00010110 | 0.053 | 00000100 | 0.050 |
| 5 | 01100110 | 0.055 | 10011111 | 0.053 | 11100000 | 0.046 |
| 6 | 10000000 | 0.047 | 01000000 | 0.030 | 10011111 | 0.036 |
| 7 | 01101000 | 0.039 | 11100110 | 0.030 | 01010110 | 0.030 |
| 8 | 01000100 | 0.032 | 01010000 | 0.023 | 01100110 | 0.030 |
| 9 | 10010000 | 0.024 | 01111110 | 0.023 | 01000110 | 0.023 |
| 10 | 11011111 | 0.024 | 11100000 | 0.023 | 01111000 | 0.021 |
| Geometric Shapes and Measures (4 attributes) | | | | | | |
| 1 | 1111 | 0.252 | 1111 | 0.417 | 1111 | 0.411 |
| 2 | 1110 | 0.173 | 1100 | 0.136 | 0000 | 0.176 |
| 3 | 0000 | 0.150 | 0000 | 0.091 | 1100 | 0.099 |
| 4 | 1000 | 0.118 | 1110 | 0.091 | 1110 | 0.094 |
| 5 | 1100 | 0.110 | 1000 | 0.076 | 1000 | 0.083 |
| 6 | 0010 | 0.071 | 1001 | 0.068 | 0001 | 0.057 |
| 7 | 1101 | 0.063 | 1101 | 0.053 | 0010 | 0.028 |
| 8 | 0001 | 0.032 | 0001 | 0.023 | 0011 | 0.020 |
| 9 | 1001 | 0.024 | 0010 | 0.023 | 1101 | 0.018 |
| 10 | 0011 | 0.008 | 0011 | 0.023 | 1001 | 0.014 |
| Data Display (3 attributes) | | | | | | |
| 1 | 000 | 0.362 | 111 | 0.364 | 111 | 0.457 |
| 2 | 001 | 0.307 | 000 | 0.333 | 000 | 0.317 |
| 3 | 111 | 0.260 | 001 | 0.159 | 001 | 0.138 |
| 4 | 101 | 0.071 | 101 | 0.099 | 101 | 0.043 |
| 5 | 100 | 0.000 | 110 | 0.030 | 110 | 0.039 |
| 6 | 110 | 0.000 | 100 | 0.015 | 100 | 0.005 |

Note: Only six are shown for the Data Display domain as it exhausts the latent class population.

CDM-Based Distractor Analysis

Examining attribute mastery with student response to an item can provide useful information to educators. This approach can be applied to a CDM-based distractor analysis to assess which response category hindered an examinee from correctly answering an item. Two examples (items 3 and 17) are presented to demonstrate the utility of a CDM-based distractor analysis. These examples illustrate how the DINA model can be an effective method to further assess where students have failed to answer an item correctly, even when they mastered all required attributes.

Which fraction is equal to $\frac{2}{3}$?

- (A) $\frac{3}{4}$
- (B) $\frac{4}{9}$
- (C) $\frac{4}{6}$
- (D) $\frac{3}{2}$

Copyright © 2008 International Association for the Evaluation of Educational Achievement (IEA). All rights reserved.

FIGURE 7
Item 3 from TIMSS questionnaire.

Item 3 (M041069) targets the Fraction and Decimals topic, which requires the mastery of attributes 2, 4, and 5 to correctly solve the item. Attribute 2 represents recognizing multiples, computing with whole numbers using the four operations, and estimating computations; attribute 4 deals with solving problems involving proportions. Both attributes 2 and 4 are from the Whole Number topic. Attribute 5 is from the Fraction and Decimals topic, which represents recognizing, representing, and understanding fractions and decimals as parts of a whole and their equivalents. Figure 7 shows the actual item 3 taken from the TIMSS questionnaire.

This item requires the examinee to identify similar proportions or multiples among fractions. Students that mastered attributes 2, 4, and 5 should infer that the correct choice would be “C.” The option provided in “C” is equivalent to $\frac{2}{3}$ by dividing the numerator and the denominator by 2, which transforms the fraction to its reduced form.

Table 9 shows the conditional probabilities of students that have mastered the required attributes by region and by multiple choice options. The breakdown of these probabilities shows that there were students that mastered the required attributes yet selected the incorrect answer in option “A.” In both Minnesota and in the United States, more than 25% and more than 22% of examinees that mastered the required attributes selected the distractor (option A), respectively.

TABLE 9
DINA Item Parameter and Conditional Probability by Region for Item 3

| Item 3 | Massachusetts | | Minnesota | | United States | |
|-----------------------------|---------------|----------|--------------|----------|---------------|----------|
| | Not Mastered | Mastered | Not Mastered | Mastered | Not Mastered | Mastered |
| A | 0.250 | 0.134 | 0.228 | 0.257 | 0.255 | 0.226 |
| B | 0.036 | 0.045 | 0.053 | 0.040 | 0.051 | 0.041 |
| C | 0.536 | 0.731 | 0.404 | 0.541 | 0.250 | 0.587 |
| D | 0.179 | 0.090 | 0.316 | 0.162 | 0.444 | 0.147 |
| guessing (g) | 0.443 | | 0.235 | | 0.227 | |
| slip (s) | 0.172 | | 0.284 | | 0.307 | |
| discrimination (δ) | 0.385 | | 0.481 | | 0.466 | |

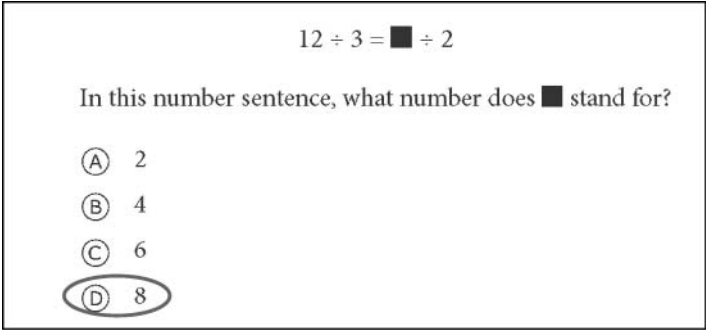
Note: Answer is C.

This tendency among the masters to select the distractor is also reflected in the high slip parameter estimate for both Minnesota and the US. Furthermore, among students that have not mastered the required attributes, Table 9 shows that option “D” was a popular choice.

This information can be useful for instructors in that specific responses chosen by students can be used for educational purposes. For example, among students that mastered the required skills, the instructor can focus on more specific materials to better train students to apply their skills. For students that lack mastery, the prevalence of choice “D” may infer teachers to instruct students on the proper use of multiples and proportions in fractions; they may also teach students that reciprocals of fractions are not always equivalent as they incorrectly selected.

Another example that demonstrates the distractor is item 17 (M031245). This item targets the Number Sentence topic in the Number domain. Figure 8 shows this item. As defined in the Q-matrix, this item requires the mastery of Attributes 2 and 7. In addition to Attribute 2, Attribute 7 deals with finding the missing number or operation and modeling simple situations involving unknowns in number sentence or expressions. Table 10 shows the DINA model parameter estimates and conditional probability of each option by region for Item 17. This is an item with very low guessing, but high slip, meaning students that mastered the two attributes required for this item still tended to slip and incorrectly solve this item.

When students slip, it would be desirable to assess which response category distracted students that have mastered both Attributes 2 and 7. There are many plausible reasons for a student to fall for a distractor. For students that mastered the required attributes, they may have answered “B” (4), if they failed to examine the problem carefully. The left-hand side of the problem is 12 divided by 3, which equals 4. Thus, the examinee is asked to find a number, when divided by 2 equals



Copyright © 2008 International Association for the Evaluation of Educational Achievement (IEA). All rights reserved.

FIGURE 8
Item 17 from TIMSS questionnaire.

4. However, a student that incorrectly solved this problem may not have multiplied 4 to 2 to get the final desired response “D” (8).

Among students that mastered the required attributes in Massachusetts, more than 43% selected the incorrect option “B.” It is clear that here, response category “B” was the distractor. Similarly, in Minnesota, more than a quarter of students that mastered the required attributes answered “B.” For the United States, more than a half selected “B” as well, and only over a third answered correctly. These results are again reflected in the high slip parameter estimates. For students that did not master the required attributes, their conditional probabilities for selecting the correct response “D” was low.

TABLE 10
DINA Item Parameter and Conditional Probability by Region for Item 17

| Item 17 | Massachusetts | | Minnesota | | United States | |
|--------------------|---------------|----------|--------------|----------|---------------|----------|
| | Not Mastered | Mastered | Not Mastered | Mastered | Not Mastered | Mastered |
| A | 0.039 | 0.014 | 0.073 | 0.080 | 0.087 | 0.029 |
| B | 0.706 | 0.432 | 0.582 | 0.253 | 0.628 | 0.506 |
| C | 0.157 | 0.054 | 0.200 | 0.067 | 0.165 | 0.074 |
| D | 0.098 | 0.500 | 0.146 | 0.600 | 0.120 | 0.391 |
| guessing (g) | 0.000 | | 0.000 | | 0.043 | |
| slip (s) | 0.485 | | 0.360 | | 0.598 | |
| discrimination (δ) | 0.515 | | 0.640 | | 0.359 | |

Note: Answer is D.

DISCUSSION

The CDM approach to analyze student attribute mastery provides a greater wealth of information than traditional methods of item analysis and measurement of student ability. The main findings from this study showed that the DINA model provides rich and fine-grained diagnostic information that can be directly applied to classroom instruction. By comparing proportion correct as a measure of student performance to proportion mastered, we were able to show that the former cannot fully assess whether a student had mastered all required skills to solve a particular item. These results indicate that even with low mastery, students were able to attain a high proportion for correct items; conversely, there were also items with higher proportion mastered, yet with lower proportion correct. These findings warrant further investigation into factors that influence the disparity between proportion correct and proportion mastered and provide motivational ground for naturally employing the CDM framework—differences between mastery and getting an item correct are captured by the guessing and slip parameters of the DINA model. The estimates from these parameters can be used to determine items that students slip and identify items where students need to learn additional materials so that they can fully master the skills required to answer correctly without guessing.

To examine the relationship between proportion correct and proportion mastered, the DINA model item parameters were estimated. As demonstrated in the results section, the information from DINA item parameters were easily interpretable and provided diagnostic information that linked the differences between an examinee's performance and their mastery of the skills. This information can be critical for instructors, because with this information, educators can hone in on specific areas that require further practice and mastery. As such, these direct applications can be beneficial to classroom instruction.

The DINA model provides a functional relationship between proportion correct and proportion mastered in that the two are linked as a function of the guessing and slip parameters:

$$p = (\eta_0)(g) + (\eta_1)(1 - s). \quad (2)$$

Equation (2) presents the general finding that proportion correct is not only derived from the relative sizes of the latent groups η_0 and η_1 , but also by the magnitude of the DINA parameters: guessing and slip. The disparity in the proportion correct and proportion mastered can then be attributed to these results. This also explains the unintuitive finding noted earlier that the benchmark participants—Massachusetts and Minnesota—which had fewer individuals expected to get an item right (i.e., low proportion mastered), can actually have a higher proportion of correct response.

Results showed that students from the US national sample had a significantly greater mastery of Attribute 13 than the benchmark participants, but had the lowest proportion correct. This can be attributed to the high guessing parameter estimates for the two benchmark states. This result also demonstrates the functional relationship expressed in equation (2). An implication for instructors and educational researchers is to ensure that students in Massachusetts and Minnesota fully understand and grasp the mastery of reading data from tables, pictographs, bar graphs, and pie charts to hinder them from guessing (Attribute 13). Furthermore, educators in Massachusetts should focus on materials that target Data Display, which is also confirmed by its large class size of students without the mastery of any attributes from this domain. For the US students as a whole, the large proportion mastered and proportion correct for Attribute 13 should be an indicator that instructors should focus on teaching students to correctly answer problems that require this attribute, because they have shown to slip and get the item wrong.

As shown through an example, CDMs also provide additional information that can be obtained for assessing how distractors behave. Traditionally, all incorrect response options may be considered distractors. However, under the DINA model used in this study, we can identify examinees that mastered all attributes required to solve the item, which can be used to calculate the conditional probability of selecting a response given their mastery. This can be used to examine the likelihood of a test taker to select an incorrect response, thereby providing educational researchers and instructors on additional materials that can be utilized to correct the test taker's mistake.

Although previous studies have noted that high estimates of DINA item parameters may be an indication of poor fit (de la Torre & Douglas, 2004), the information criteria statistics preferred the DINA model over IRT models. From the US data, items 1 and 6 have notably high estimates of the guessing parameter, 0.621 and 0.769, respectively. However, when we traced their item characteristics to the 3PL model used by TIMSS, we noted that the TIMSS-calibrated pseudo-guessing (lower asymptote) estimates were 0.305 and 0.204, respectively (Olson et al., 2009, p. 451). These estimates are high considering they were calibrated using an international sample derived from all participating countries. While interpretations of the IRT pseudo-guessing parameter and the DINA guessing parameter are somewhat different, they can provide some justification for the high DINA guessing estimates, because items with high guessing may be easy, which explains the high proportion of students that would get them right without mastering the attributes that were required for the items. Although such rationale can call for an evaluation as to whether these basic skills should be included in the Q-matrix, from a theoretical and mathematics education perspective, they should be specified, because these attributes were established prior to the test development and adhere to the curricular design of mathematics instruction. In other words, regardless of item difficulty or the number of students that are able to solve the item correctly

without mastering the required attributes, these attributes are important skills that should be included in the Q-matrix.

In addition, for items with high parameter estimates, the possibility of alternative or multiple strategies for solving an item that may better explain students' responses can be considered. For example, the deterministic input, noisy output, "or" gate model (DINO) introduced by Templin and Henson (2006) can be implemented. This model relaxes the requirements for solving an item; unlike the DINA model, which specifies the mastery of all required attributes to solve an item, the DINO model examines whether at least one of the required attributes is mastered. The inferences resulting from the DINO model can provide insights as to whether more attributes were specified than indicated in the Q-matrix. Furthermore, the multiple-strategy DINA model (de la Torre & Douglas, 2008), which extends the single-strategy DINA model to allow for multiple strategies, can be applied; this model would examine whether other methods of solving the items can better explain the results. As such, the application of other CDMs to these items can help explain the high estimates by considering methods that vary in structuring how the attributes specified in the Q-matrix are modeled.

One consideration that required further scrutiny in this study was the small sample sizes of the two regional entities. In recent years, many CDMs have been proposed and have demonstrated advantages of providing fine-grained information about examinees' mastery or non-mastery of attributes than what is available or inferred using IRT models. When a new model or an estimation technique is developed, it becomes necessary to evaluate the accuracy and the stability of their parameter estimates. However, in contrast to guidelines for optimal and/or minimal requirements in sample size and test length for IRT models, to our understanding, there is yet to be a study in the literature that provides such recommendations for CDMs. To address this issue thoroughly, a recovery study under varying practical testing conditions should be conducted.

Another note to consider over the sample size is the systematic restriction placed on the data-release scheme undertaken by TIMSS that limited the number of participants per booklet administered. Yet, given the shortcomings of the sample size, an analysis of the data under the CDM framework demonstrate its utility as the data provides a unique and rare opportunity to compare attribute mastery within a country that is based on an assessment developed for an international comparison in students' mathematics achievement. It is not often that regional entities within a country participate in the TIMSS—it was the first time that Massachusetts joined the testing and the second time for Minnesota (last testing was in 1995) at the fourth grade level, while following the same testing procedures as the other countries; besides these two regions, there were no other US states that participated in the TIMSS 2007 testing. To be able to compare these benchmark participants along with its international peers provides merit for conducting this study.

The advantages of the method demonstrated in this study should be explored further in other analyses; attribute mastery can be mapped to demographic information on students, schools, teachers, and administrators to explain factors leading to success, while attempting to provide pedagogical implications for the education stakeholders through their data. It is such ability to trigger attribute-specific information to instructors that can be used directly in classrooms, which signifies the importance of the CDM approach in the educational setting.

REFERENCES

- American Federation of Teachers (AFT). (1999). What TIMSS tells us about mathematics achievement, curriculum, and instruction. In *Educational Issues Policy Brief* (Vol. 10, pp. 2–10). Washington, DC: AFT Educational Issues Department.
- Birenbaum, M., Tatsuoaka, C., & Yamada, T. (2004). Diagnostic assessment in TIMSS-R: Between-countries and within-country comparisons of eighth graders' mathematics performance. *Studies in Educational Evaluation*, 30(2), 151–173.
- Burke, M. J., & Henson, R. (2008). *LCDM user's manual*. Greensboro, NC: University of North Carolina at Greensboro.
- de la Torre, J. (in press). The generalized DINA model framework. *Psychometrika*.
- de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45, 343–362.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34, 115–130.
- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333–353.
- de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, 73, 595–624.
- de la Torre, J., & Lee, Y. S. (2008, March). *Relationships between cognitive diagnosis, CTT and IRT indices: An empirical investigation*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.
- Dogan, E., & Tatsuoaka, K. K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R. *Educational Studies in Mathematics*, 68(3), 263–272.
- Doignon, J. P., & Falmagne, J. C. (1999). *Knowledge spaces*. New York: Springer-Verlag.
- Doornik, J. A. (2002). *Object-Oriented Matrix Programming Using Ox* (Version 3.1) [Computer software]. London: Timberlake Consultants Press.
- Ferrini-Mundy, J., & Schmidt, W. H. (2005). International comparative studies in mathematics education: Opportunities for collaboration and challenges for researchers. *Journal for Research in Mathematics Education*, 36, 164–175.
- Foy, P., & Olson, J. F. (2009). *TIMSS 2007 user guide for the international database*. Chestnut Hill, MA: IEA.
- Ginsburg, A., Cooke, G., Leinwant, S., Noell, J., & Pollock, E. (2005). *Reassessing U.S. international mathematics performance: New findings from the 2003 TIMSS and PISA*. Washington, DC: American Institutes for Research.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 333–352.

- Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74, 191–210.
- Holliday, W. G., & Holliday, B. W. (2003). Why using international comparative math and science achievement data from TIMSS is not helpful. *The Educational Forum*, 67, 250–258.
- Hutchison, D., & Schagen, I. (2007). Comparisons between PISA and TIMSS—Are we the man with two watches? In T. Loveless (Ed.), *Lessons Learned: What international assessments tell us about math achievement* (pp. 227–262). Washington, DC: Brookings Institute Press.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258–272.
- Lee, Y.-S., Choi, K. M., & Park, Y. S. (2009, April). *Cognitive diagnosis modeling application to TIMSS: A comparison between the U.S. and Korea via CTT, IRT, and DINA*. Presented at the Annual Meeting of the American Education Research Association, San Diego, CA.
- Lemke, M., Sen, A., Pahlke, E., Partelow, L., Miller, D., Williams, T., et al. (2004). *International outcomes of learning in mathematics literacy and problem solving: PISA 2003 results from the U.S. perspective*. (NCES 2005-003). Washington, DC: US Department of Education, National Center for Education Statistics.
- Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 33, 379–416.
- Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: IEA.
- Olson, J. F., Martin, M. O., & Mullis, I. V. S. (2009). *TIMSS 2007 technical report*. Chestnut Hill, MA: IEA.
- Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 51, 337–350.
- Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345–354.
- Tatsuoka, K. K., Corter, J., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, 41(4), 901–926.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287–305.
- Um, E., Dogan, E., Im, S., Tatsuoka, K., & Corter, J. E. (2003, April). *Comparing eighth grade diagnostic test results for Korean, Czech, and American students*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (RR-05-16). Princeton, NJ: Educational Testing Service.
- Wang, J. (2001). TIMSS primary and middle school data: Some technical concerns. *Educational Researcher*, 30, 17–21.