Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 35

Hierarchical Cognitive Diagnostic Analysis for TIMSS 2003 Mathematics

Yu-Lan Su, Kyong Mi Choi, Won-Chan Lee, Taehoon Choi, & Melissa McAninch †

August 2013

---

†Yu-Lan Su is program development associate at ACT (email: yulan.su@act.org). Won-Chan Lee is Associate Professor and Co-director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@uiowa.edu). Kyong Mi Choi is Assistant Professor in Mathematics Education, Department of Teaching and Learning, N291 Lindquist Center, University of Iowa (email: kyongmi-choi@-uiowa.edu). Taehoon Choi and Melissa McAninch are research assistants at Department of Teaching and Learning, Mathematics Education, University of Iowa.

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

Attributes modified from the Common Core State Standards were adapted to construct the Q-matrices for two TIMSS 2003 Eighth Grade Mathematics booklets. A hierarchical structure of the mathematic attributes was built. The study used two modified cognitive diagnostic models, the deterministic, inputs, noisy, and gate model with hierarchical configuration and the deterministic, inputs, noisy, or gate model with hierarchical configuration, to analyze the data. Both approaches incorporated the hierarchical structures of the cognitive skills in the model estimation process, and were introduced for situations where the attributes were ordered hierarchically. This can facilitate reporting the mastery/non-mastery of skills with different levels of cognitive loadings. The purposes of the TIMSS data analysis are to construct the cognitive hierarchy of mathematic attributes, demonstrate the proposed approaches and the feasibility of retrofitting, compare the results of conventional DINA and DINO models to their hierarchical counterparts with different sample sizes, and promote the potential contributions of constructing skill hierarchies to teachers and students.

1 Introduction

Researchers have suggested that mathematical and scientific concepts (and other conceptual domains) are not independent knowledge segments, and there are learning sequences in the curriculum that fits learners' schema-constructing process (e.g., Clements & Sarama, 2004; Kuhn, 2001; Vosniadou & Brewer, 1992). Since the skills in mathematics are not independent of each other, it is crucial to use an estimation model that is consistent with the assumptions about relationships among attributes. Specifying attribute profiles incorrectly would affect the accuracy of estimates of item and attribute parameters. Considering the hierarchical nature of mathematics attributes, the conventional CDMs (cognitive diagnostic models) that do not assume this character, and the results of the calibration based on these models may be biased or less accurate. Hence, the relationships among attributes and the possible attribute profiles need to be identified correctly based on content specific theoretical background along with a careful look at the test blueprint before a CDM calibration is conducted and interpreted.

The nature of mathematics concepts is that they are not independent of each other (Battista, 2004). For example, number and operation, algebra, geometry, measurement, and probability are not independent domains. Educators have discussed the proper learning sequences in mathematics teaching and learning (Baroody, Cibulskis, Lai, & Li, 2004; Clements & Sarama, 2004). From mathematics educators' perspectives, mathematics concepts are hierarchically ordered. This needs to be reflected and considered in identifying the relationships among attributes, possible attribute profiles, and designing Q-matrices.

The illustration of applying CDMs to a large-scale assessment demonstrates the feasibility of retrofitting (i.e., analyzing an already existing data set) TIMSS 2003 data. Studies of international assessments, such as the TIMSS, allow for worldwide comparisons. Although the intention of TIMSS is not to provide individual level scores or comparisons, successful application of a CDM to a large scale assessment can be a promising way to provide informational feedback about examinees' mastery in varying levels of skills. While other studies have tried retrofitting the TIMSS data, no research has applied or studied the concept of hierarchically ordered skills. The current study provides information that allows future research to conduct international comparisons and identifications of how examinees do or do not master specific fine-grained fundamental, intermediate, and advanced skills. Such comparisons will provide educators and policymakers with information on student achievement across countries that will be helpful in evaluating curricular development and in developing education reform strategies.

2 Significance of Learning Sequences

Educators and researchers have long focused their attention on learning sequences, and advocated the importance of ordering instructions to build up

learning sequences. As early as 1922, Thorndike claimed that significant instructional time and effort was wasted because the associations between previous and later learning (the laws of learning) were neglected and not used to facilitate learning (Baroody et al., 2004). Thorndike recommended that educators recognize the relation of learning processes to principles of content prior to the initiation of learning or instruction. Gagne and Briggs (1974) developed a hierarchy of goals based on logical and empirical task analyses, which they applied to develop curricula for elementary education. In the mid twentieth century, information-processing theories used the input-process-output metaphor to describe learning processes (Baddeley, 1998).

Cognitive research has suggested that some preliminary knowledge can be defined as the foundation for other more sophisticated knowledge (e.g., Kuhn, 2001; Vosniadou & Brewer, 1992). The associations of knowledge skills are especially important for conceptual understanding and problem solving. Conceptual understanding implies that students have the ability to use knowledge, to apply it to related problems, and to make connections between related ideas (Bransford, Brown, & Cocking, 2000). This means that building conceptual understanding involves connecting newly introduced information to existing knowledge as the student builds an organized and integrated structure (Ausubel, 1968; Linn, Eylon, & Davis, 2004). Mathematics educators have clarified levels of development in students' understanding and constructing of mathematics concepts from early number and measurement ideas, to rational numbers and proportional reasoning, to algebra, geometry, calculus, and statistics (Lesh & Yoon, 2004). These levels of knowledge development are structured by researchers in ladder-like sequences, with each successive run closer to the most sophisticated level.

Empirically tested learning sequences should be fully articulated for curriculum developers to use as a ready-made artifact in developing coherent curricula. Researchers have called for the need for developing learning sequences to inform the development of coherent curricula over the span of K-12 science education (Krajcik, Shin, Stevens, & Short, 2010). Results from the TIMSS have shown that a coherent curriculum is the primary predictor of student achievement (Schmidt, Wang, & McKnight, 2005). If the curriculum is not built coherently to help learners make connections between ideas within and among disciplines or form a meaningful structure for integrating knowledge, students may lack foundational knowledge that can be applied to future learning and for solving problems that confront them in their lives (Krajcik et al., 2010; Schmidt et al., 2005).

2.1 Sequential Nature of Mathematical Concepts

Mathematics encompasses a wide variety of skills and concepts. These skills and concepts are related and often build on one another (Sternberg & Ben-Zeev, 1996). Some math skills obviously develop sequentially. For example, a child cannot begin to add numbers until he knows that those numbers represent quantities. Solving mathematical problems frequently involves separate

processes of induction, deduction, and mathematical conceptualization (Nesher & Kilpatrick, 1990). However, certain advanced skills do not seem to have a clear dependent relationship. For example, a student who often makes simple calculation errors may still be able to solve a calculus problem that requires sophisticated conceptual thinking.

Educators have tried to identify sets of expected milestones for a given age and grade as a means of assessing a child's progress, and of better understanding in which step students go wrong (Levine, Gordon, & Reed, 1987). NCTM (2000)'s Principles and Standards for School Mathematics also outlines recommendations for classroom mathematics instructions for both content matter and process based on different groups of students (i.e., K-2, 3-5, 6-8, and 9-12). The Standards expect all students to complete a core curriculum that has shifted its emphasis away from computation and routine problem practice toward reasoning, real-world problem solving, communication, and connections (NCTM, 2000).

A developmental progression embodies theoretical assumptions about mathematics; for example, a student needs to be able to build an image of a shape, match that image to the goal shape by superposition, and perform mental transformation in order to solve certain manipulative shape composition tasks (Clements, Wilson, & Sarama, 2004). Researchers have been devoted to finding evidence to support the assumptions. For example, the findings from Clements, Wilson et al. (2004) suggested that students demonstrate varying levels of thinking when given tasks involving the composition and decomposition of two-dimensional figures, and that the older students with previous experience in geometry tend to evince higher levels of thinking. Their results also showed that students moved through several distinct levels of thinking and competence in the domain of composition and decomposition of geometric figures.

The recognition of the sequential nature of mathematical concepts impacts the development of curriculum design and student learning. The attention on developing students' learning sequences in mathematics also impacts teacher education in mathematics. Researchers suggested that teachers in mathematics must be well-trained to demonstrate competencies in knowledge and skills in teaching mathematics, understanding of the sequential nature of mathematics, the mathematical structures inherent in the content strands, and the connections among mathematical concepts, procedures and their practical applications (Steeves & Tomey, 1998).

3 Attribute Hierarchies

Attribute hierarchies represent the interdependency among cognitive attributes. It refers to situations in which the mastery of a certain attribute is prerequisite to the mastery of another attribute. The attribute with the lower cognitive load is developed earlier than attributes with higher cognitive loads. Thus, the first attribute is located in the lowest layer of the hierarchy, and the second attribute is in the next highest layer of the same hierarchy. Four common types

of cognitive attribute hierarchies are linear, convergent, divergent, and unstructured (Gierl, Leighton, & Hunka, 2007; Leighton, Gierl, & Hunka, 2004; Rupp, Templin, & Henson, 2010). These four hierarchies are shown in Figures 1 to 4 taken from Gierl et al. (2007) and Leighton et al. (2004), using six attributes as an example. The linear attribute hierarchy requires all attributes to be ordered sequentially. If an examinee has mastered attribute 2, then he or she has also mastered attribute 1. Furthermore, an examinee who has mastered attribute 3 has also mastered attributes 1 and 2, and so on. The convergent attribute hierarchy specifies a situation in which a single attribute could be the prerequisite of multiple different attributes. It also includes situations where a single attribute could require the mastering of one or more of the multiple preceding attributes. In this case, an examinee mastering attributes 3 or 4 has also mastered attributes 1 and 2. An examinee mastering attribute 5 has mastered attribute 3, attribute 4, or both, and has also mastered attributes 1 and 2. This implies that an examinee could achieve a certain skill level through different paths with different mastered attributes. The divergent attribute hierarchy refers to different distinct tracks originating from the same single attribute. In a divergent attribute hierarchy, an examinee mastering attributes 2 or 4 has also mastered attribute 1. An examinee mastering attributes 5 or 6 has mastered attributes 1 and 4. The unstructured attribute hierarchy describes cases when a single attribute could be prerequisite to multiple attributes, and where those attributes have no direct relationship to each other. For example, an examinee mastering attributes 2, 3, 4, 5 or 6 means only that he or she has mastered attribute 1.

As with the Q-matrix, 0 means the attribute is not mastered, and 1 means the attribute is mastered. The number of possible attribute profiles is different for various attribute hierarchies. The more independent the attributes, the larger the number of possible attribute profiles. The higher the dependency among the attributes, the fewer the number of possible attribute profiles. An assessment could be a combination of various attribute hierarchies, and thus the possible number of attribute profiles is uniquely different for each assessment. Varying types of structures could appear for a certain type of hierarchy. When a test is developed based on attribute hierarchies, the number of possible attribute profiles reduces dramatically from 2^K. Hence, the complexity of estimating a CDM is decreased, and the sample size requirement is lowered.

Several CDMs have been applied to parameterize the latent attribute space to model the relationships among attributes and help improve the efficiency in estimating parameters. These approaches include log-linear (Maris, 1999; Xu & von Davier, 2008), unstructured tetrachoric correlation (Hartz, 2002), and structured tetrachoric correlation (de la Torre & Douglas, 2004; Templin, 2004). The log-linear models parameterize the latent class probabilities using a log-linear model that contains main effects and interaction effects (all possible combinations of attributes). The unstructured tetrachoric models represent the tetrachoric correlations of all attributes pairs directly, and reduce the complexity of model space. The structured tetrachoric models impose constraints on the tetrachoric correlation matrix to simplify the estimation process using prior

hypotheses about how strongly attributes are correlated. However, none of these approaches incorporate the hierarchical nature of cognitive skills and reduce the number of possible attribute profiles directly. The attribute hierarchy method (AHM) (Gierl, 2007; Gierl, Cui, & Zhou, 2009; Gierl et al., 2007; Leighton et al., 2004) is another cognitive diagnostic psychometric method designed to explicitly model the hierarchical dependencies among attributes underlying examinees' problem solving on test items. The AHM is based on the assumption that test items can be described by a set of hierarchically ordered skills, and that examinee responses can be classified into different attribute profiles based on the structured hierarchical models.

Researchers' attention to the impact of cognitive theory on test design has been very limited (Gierl & Zhou, 2008; Leighton et al., 2004). The assumption of skill dependency that AHM holds is consistent with findings from cognitive research (e.g., Kuhn, 2001; Vosniadou & Brewer, 1992) that suggests some preliminary knowledge can be defined as the foundation for other more sophisticated knowledge or skills with higher cognitive loadings. The concept of hierarchically ordered cognitive skills is clearly observed in mathematics learning. For example, a student learns to calculate using single digits before learning to calculate using multiple digits. If the hierarchical relationships among skills are specified, the number of permissible items is decreased and the possible attribute profiles can be reduced from the number of 2^K (Gierl et al., 2007; Leighton et al., 2004; Rupp et al., 2010). The AHM could be a useful technique to design a test blueprint based on the cognitive skill hierarchies. Since the AHM is more like an analytical method and a test developing guideline that focuses on estimating attribute profiles, it will be beneficial if a model based on AHM is developed to allow item parameters to be estimated directly. In addition, de la Torre and Karelitz (2009) tried to estimate item parameters based on a linear structure of five attributes with each item class unidimensionally represented as a cut point on the latent continuum. The focus of their study was to transform the IRT 2PL item parameters into the DINA model's slip and guessing parameters, and then examine how the congruence between the nature of the underlying latent trait (continuous or discrete) and fitted model affects DINA item parameter estimation and attribute classification under different diagnostic conditions, rather than focusing on the hierarchical structures of attributes and its estimation.

If an assessment is developed based on hierarchically structured cognitive skills, and the Q-matrix for each test is built up and coded based on those skills, analyzing the tests using CDMs would directly provide examinees, teachers, or parents with more valuable information about which fundamental, intermediate, or advanced skills the test-takers possess. Instructors can also take the feedback to reflect on their teaching procedures and curricular development. Moreover, for some test batteries that target various grade levels, conducting CDM calibrations incorporating the hierarchically structured cognitive skills would help estimate both item parameters and examinee attribute profiles based on different requirements about the mastery of various levels of skills.

While CDMs offer valuable information about specific skills, their usefulness is limited by the time and sample sizes required to test multiple skills simulta-

neously. To estimate examinees' ability parameters via a CDM, examinees' skill response patterns are classified into different attribute profiles (latent classes). For most CDMs without any relationship or constraints imposed on the latent classes, the maximal number of possible latent classes is 2^K, in which K is the number of attributes measured by the assessment. There are $2^K - 1$ parameters that need to be estimated by implementing CDMs with dichotomous latent attribute variables. As the number of attributes increase, the number of estimated parameters also increases, as does the required sample size and computing time needed to attain reliable results. Analyzing an assessment measuring many attributes is difficult due to the large sample size required to fit a CDM, and to obtain reliable parameter estimates, convergence, and computational efficiency.

Most CDM application examples in the literature are limited to no more than eight attributes (Hartz, 2002; Maris, 1999; Rupp & Templin, 2008a) because of the long computing time for models with larger numbers of attributes and items. If the number of latent classes can be reduced from 2^K, the sample size needed to obtain stable parameter estimates from CDM calibrations will decrease. This will also result in faster computing time. One solution to decrease the number of latent classes is to impose hierarchical structures (Leighton et al., 2004) on skills. The resulting approach is able to assess and analyze more attributes by reducing the number of possible latent classes and the sample size requirement (de la Torre, 2008, 2009; de la Torre & Lee, 2010). Two methods to estimate attributes with hierarchical structures could be as de la Torre (2012) suggested:

> First, keeping the EM algorithm as is, but without any gain in efficiency, the prior value of attribute patterns not possible under the hierarchy can be set to 0, and second, for greater efficiency, but requiring minor modifications of the EM algorithm, attribute patterns not possible under the hierarchy can be dropped. (session 5: p.7)

To explore whether data from a test with the hierarchical orders fit CDMs, this study intended to consider hierarchically structured cognitive skills when determining attributes and identifying attribute profiles to reduce the number of possible latent classes and decrease sample size requirements, which is the more efficient method suggested by de la Torre (2012). The deterministic, inputs, noisy, and gate (DINA; Haertel, 1989, 1990; Junker & Sijtsma, 2001) model and the deterministic, inputs, noisy, or gate (DINO; Templin & Henson, 2006) model are employed in the study. The DINA model is increasingly valued for its interpretability and accessibility (de la Torre, 2009), for the invariance property of its parameters (de la Torre & Lee, 2010), and for good model-data fit (de la Torre & Douglas, 2008). Being one of the simplest CDMs with only two item parameters (slip and guessing), the DINA model is the foundation of other CDMs; it is easily estimated and has gained much attention in recent CDM studies (Huebner & Wang, 2011). Hence, the DINA model is a good choice to apply the proposed approach of imposing skill hierarchies on possible attribute profiles. Likewise, the DINO model is a statistically simpler CDM like the DINA

model, and is the counterpart of the DINA model based on slightly different assumptions about the possibility of answering an item correctly. Hence, these two models provide a good comparison in understanding the feasibility of analyzing the hierarchically structured test data using CDMs.

The proposed models with a skill hierarchy constraint on the possible attribute profiles are different from the conventional DINA and DINO models, and are referred to in this study as the DINA-H and DINO-H models. They are not different models from the conventional models in terms of mathematical representation. They differ only in the constraint on defining attribute profiles according to the skill hierarchy. The intent of this study is to apply the proposed skill hierarchy approach in conjunction with the DINA and DINO models to analyze the real data, to compare the results of conventional DINA and DINO models to their hierarchical counterparts with different sample sizes, and to promote the potential contributions of constructing skill hierarchies for teachers and students.

4 DINA Model

The DINA model, one of the most parsimonious CDMs that require only two interpretable item parameters, is the foundation of other models applied in cognitive diagnostic tests (Doignon & Falmagne, 1999; Tatsuoka, 1995, 2002). The DINA model is a non-compensatory, conjunctive CDM, and assumes that an examinee must know all the required attributes in order to answer an item correctly (Henson, Templin, & Willse, 2009). An examinee mastering only some of the required attributes for an item will have the same success probability as another examinee possessing none of the attributes. For each item, the examinee item respondents are scored into two latent classes: one class indicates answering the item correctly (scored 1), containing examinees who possess all attributes required for answering that item correctly; the other class indicates incorrectly answering the item (scored 0), containing examinees who lack at least one of the required attributes for answering that item correctly. This feature is true for any number of attributes specified in the Q-matrix (de la Torre, 2011). The complexity of the DINA model is not influenced by the number of attributes measured by a test because its parameters are estimated for each item but not for each attribute, unlike other non-compensatory conjunctive cognitive diagnostic models (e.g., the RUM) (Rupp & Templin, 2008b).

The DINA model has two item parameters, slip (s_j) and guess (g_j). The term slip refers to the probability of an examinee possessing all the required attributes but failing to answer the item correctly. The term guess refers to the probability of a correct response in the absence of one or more required attributes. However, the two item parameters also encompass other nuisances. Those nuisances confound the reasons why examinees who have not mastered some required attributes can answer an item correctly, and the reasons why examinees who have mastered all the required attributes can miss the correct response. Two examples of the common nuisances are the misspecifications in

the Q-matrix, and the usage of alternative strategies, as Junker and Sijtsma (2001) described when they first advocated the DINO model. Below are the mathematics presentations for the two item parameters:

$$s_j = P(X_{ij} = 0 \,|\, \eta_{ij} = 1), \tag{1}$$

$$g_j = P(X_{ij} = 1 \,|\, \eta_{ij} = 0), \tag{2}$$

and the item response function in the DINA model is defined as

$$P_j(\alpha_i) = P(X_{ij} = 1 \,|\, \alpha_i) = g_j^{1-\eta_{ij}}(1 - s_j)^{\eta_{ij}}, \tag{3}$$

where the η matrix refers to a matrix of binary indicators showing whether the examinee attribute profile pattern i has mastered all of the required skills for item j. The formula is defined as:

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}, \tag{4}$$

where α_{ik} refers to the binary mastery status of the k^{th} skill of the i^{th} skill pattern (1 denotes mastery of skill k, and 0 denotes non-mastery). And, as discussed in the previous section, q_{jk} here is the Q-matrix entries specifying whether the j^{th} item requires the k^{th} skill. The value of this deterministic latent response, η_{ij}, is zero if an examinee is missing at least one of the required attributes.

Analyzing a DINA model requires test content specialists to first construct a Q-matrix to specify which item measures the appropriate attributes, similar to implementing many other CDMs. However, many CDM analyses assume that the specification of a Q-matrix is correct (or true), without v erifying its suitability statistically. An incorrectly specified Q-matrix would mislead the results of the analysis. If the results show a model misfit because of an inappropriate Q-matrix, the misfit issue is hard to detect and solve (de la Torre, 2008). Hence, de la Torre (2008) proposed a sequential EM-based δ-method for validating the Q-matrices when implementing the DINA model. In his method, δ_j is defined as "the difference in the probabilities of correct responses between examinees in groups $\eta_j = 1$ and $\eta_j = 0$" (i.e., examinees with latent responses 1 and 0) (as cited in de la Torre, 2008, p. 344). δ_j serves as a discrimination index of item quality that accounts for both the slip and guessing parameters. Below is the computation formula for item j:

$$\delta_j = 1 - s_j - g_j. \tag{5}$$

The higher the guessing and/or slip parameters are, the lower the value of δ_j. This signifies that the less-discriminating items have high guessing and slip parameters, and have a smaller discrimination index value of δ_j. In contrast, an item that perfectly discriminates between examinees in groups $\eta_j = 1$ and $\eta_j = 0$ has a discrimination index of $\delta_j = 1$ because there is no guessing and slip. Therefore, the higher the value of δ_j is, the more discriminating the item is.

5 DINO Model

The DINO model is the disjunctive counterpart of the DINA model (Templin & Henson, 2006). Similar to the DINA model, the DINO model has two item parameters: s_j and g_j. In the DINO model, examinees are divided into two groups. The first group of examinees have at least one of the required skills specified in the Q-matrix ($\omega_{ij} = 1$), and the second group of examinees do not possess any skills specified in the Q-matrix ($\omega_{ij} = 0$). At least one Q-matrix skill must be mastered for a high probability of success in the DINO model. Hence, the slip parameter (s_j) indicates the probability that examinee i, who masters at least one of the required skills for item j, answers it incorrectly. The guessing parameter (g_j) refers to the probability of a correct response when the examinee possesses none of the required skills. In other words, the DINO model assumes that the probability of a correct response, given mastery on at least one skill, does not depend on the number and type of skills that are mastered. It allows for low levels on certain skills to be compensated for by high levels on other skills. The item parameters are defined as:

$$s_j = P(X_{ij} = 0 \,|\, \omega_{ij} = 1), \tag{6}$$

$$g_j = P(X_{ij} = 1 \,|\, \omega_{ij} = 0), \tag{7}$$

and the item response function in the DINO model is defined as

$$P_j(\omega_{ij}) = P(X_{ij} = 1 \,|\, \omega_{ij}) = g_j^{1-\omega_{ij}}(1 - s_j)^{\omega_{ij}} \tag{8}$$

where

$$\omega_{ij} = 1 - \prod_{k=1}^{K}(1 - \alpha_{ik})^{q_{jk}}. \tag{9}$$

Both the DINO and DINA models are simpler CDMs. They assign only two parameters per item and partition the latent space into exactly two sections (i.e., mastery and non-mastery). Other more complex models (such as rRUM, linear logistic, etc.) assign *K, K+1*, or more parameters per item, and partition the latent space into multiple sections. Nevertheless, the DINO model is more popular in the medical, clinical, and psychological fields, because such diagnoses in these fields are typically based on the presence of only some of the possible major symptoms. The absence of certain symptoms can be compensated for by the presence of others.

6 Methodology

6.1 Modified Methods

The study proposed two modified approaches: The DINA model with hierarchical configurations (DINA-H) and the DINO model with hierarchical configurations (DINO-H). Both models involve the hierarchical structures of the cognitive

skills in the estimation process and were introduced for situations where the attributes are ordered hierarchically. The DINA-H and DINO-H models have the same basic specifications as the conventional DINA and DINO models. The only difference is that the pre-specified possible attribute profiles under a certain skill hierarchy are adapted in the DINA-H and DINO-H models. In the conventional DINA and DINO models, the number of possible attribute profiles L is equal to 2^K (where K refers to the number of skills being measured). In the DINA-H and DINO-H models, L is equal to the number of all possible attribute profiles specified for each unique model. In the DINA and DINO models, the initial possible attribute profiles α is all the 2^K possible combinations of 0s and 1s, whereas α is set to be the possible attribute profiles specified for each unique DINA-H and DINO-H model. Examinees are classified into these specified possible attribute profiles during the estimation process. The number of parameters in the conventional DINA and DINO models is equal to $2J + 2^K - 1$ (where J refers to the number of items in a test). For the DINA-H and DINO-H models, the number of parameters is equal to $2J + L - 1$ where L represents the number of all possible attribute profiles specified for each unique hierarchical model.

Based on the attribute hierarchies, the number of attribute profiles could be found for each hierarchical model. The number of possible attribute profiles would be different for various attribute hierarchies. The more independent the attributes, the larger the number of possible attribute profiles. The higher the dependency among the attributes, the fewer the number of possible attribute profiles. Since the convergent and the divergent hierarchies could have varying structures, the numbers of possible attribute profiles would be different for various structures.

In the DINA and DINO models, the number of possible attribute profiles L equals 2^K, whereas L equals the maximum number of possible attribute profiles specified for each unique DINA-H and DINO-H models. In the DINA and DINO models, the initial possible attribute profiles α contain all 2^K possible combinations of 0s and 1s, whereas α are the possible attribute profiles specified for each unique DINA-H and DINO-H models. The major steps of EM computation described in de la Torre (2009) were followed to estimate examinee attribute profiles and item parameters. The criterion for convergence was set to be smaller than 0.001.

6.2 Description of Data

The data used in this study came from the TIMSS 2003 U.S. eighth grade mathematics test. This real data analysis is a retrofitting analysis, which means it is an analysis of an already existing assessment using other models (e.g., the DINA and the DINO models). The goal of the analysis of the real data is to benchmark the proposed models and to address research questions regarding whether the DINA-H and the DINO-H models provide reasonable parameter estimates, and whether they provide more stable calibration results than the conventional DINA and the conventional DINO models when there is a hierarchy in attributes.

TIMSS provides data on the mathematics and science curricular achievement of fourth and eighth grade students and on related contextual aspects such as mathematics and science curricula and classroom practices across countries, including the U.S. TIMSS is a sample-based assessment whose results can be generalized to a larger population. Its data were collected in a four-year cycle starting in 1995. TIMSS 2003 was the third comparison carried out by the International Association for the Evaluation of Educational Achievement (IEA), an international organization of national research institution and governmental research agency.

There were 49 countries that participated in TIMSS 2003: 48 participated at the eighth grade level and 26 at the fourth grade level (Martin, 2005). The TIMSS 2003 eighth grade assessment contained 383 items, 194 in mathematics and 189 in science (Neidorf & Garden, 2004). Each student took one booklet containing both mathematics and science items, which were only a subset of the items in the whole assessment item pool. The TIMSS scale was set at 500 and the standard deviation at 100 when it was developed in 1995. The average score over countries in 2003 is 467 with a standard deviation of 0.5. The assessment time for individual students was 72 minutes at fourth grade and 90 minutes at eighth grade. The released TIMSS math test included five domains in mathematics: Number and operation, algebra, geometry, measurement, and data analysis and probability. The items and data were available to the public, and could be downloaded from TIMSS 2003 International Data Explorer (http://nces.ed.gov/timss/idetimss/).

Two types of content domains, number-and-operation and algebra, were used in the study because the ability to do number-and-operation is the prerequisite for algebra, and also there were more released items available. The hierarchical ordering of mathematical skills in both number and algebra was found in other empirical studies. For example, Gierl and Leighton et al. (2009) used the think aloud method to identify the hierarchical structures and attributes for Basic Algebra on SAT, and identified five attributes, single ratio setup, conceptual geometric series, abstract geometric series, quadratic equation, and fraction transformation, for a basic Algebra item.

Booklets 1 and 2 from TIMSS 2003 were used in the study. One number-and-operation item and one algebra item were excluded in the analysis because they were too easy and only measure elementary-level attributes. There were 18 number-and-operation items, 11 algebra items, and 757 U.S. examinees for booklet 1 (B1). There were 21 number-and-operation items, 9 algebra items, and 740 U.S. examinees for booklet 2 (B2). About half of the items were released in 2003 and the others were released in 2007. Four number-and-operation items in booklet 1 were in constructed-response format, and five number-and-operation items in booklet 2 were in constructed-response format. One algebra item in booklet 2 was a constructed response item. Three of the constructed response items in number-and-operation in booklet 1, one of them in number-and-operation in booklet 2, and one in algebra in booklet 2 were multiple-scored items. However, in the current study, these items were rescored as 0/1 dichotomous items in the examinees' score matrix to conduct the CDM calibration.

For those examinees who got full score points of 2 were rescored as 1, and who got score point of 1 were rescored as 0. In addition to the small U.S. sample, a larger sample size including the benchmark participants for each booklet was also applied for the comparison analysis. The subsequently larger benchmarking sample of B1 is 1134, including the Basque Country of Spain (N=216), the U.S. state of Indiana (N=195), and the Canadian provinces of Ontario (N=357) and Quebec (N=366). The benchmarking sample of B2 is 1114, including the Basque Country of Spain (N=216), the U.S. state of Indiana (N=189), and the Canadian provinces of Ontario (N=346) and Quebec (N=363). Table 1 summarizes the difference between Booklet 1 and Booklet 2.

6.3 Q-Matrix

To analyze the real data using CDMs, the first step was to construct a Q-matrix that specified the skills necessary to solve each item. The current study adapted the attributes from the CCSS (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010), and Q-matrix from the consensus of two doctoral students majoring in secondary teaching and learning. These two content experts were former middle and high school math teachers. Independently, the two experts first answered each item, wrote down the strategies/process they used to solve each item, and then coded and matched the attributes for each item. The attributes used for coding were adapted from grades six to eight CCSS. A follow-up discussion time was scheduled to solve the coding inconsistencies between the experts, and reach an agreement. When they were not able to reach an agreement for an item through discussion, a professor in secondary school mathematics solved the conflict. Table 2 provides the attributes modified form the CCSS and their corresponding TIMSS items. To illustrate, Table 3 shows one item from each booklet with the attributes being measured. The Q-matrices of B1 and B2 are shown in Tables 4 and 5, respectively. The percentage of coders' overall agreement for constructing the Q-matrices is 88.89%.

The next step was for the two experts to arrange and organize the attributes into a hierarchical order which they thought reasonable based on the CCSS mathematics grade level arrangement. To determine the hierarchies among the attributes, the following order was followed to arrange those within-a-grade attributes: recognize/understand, use, compare, apply, and then solve real-world problems. The coders worked together and reached an agreement for the final decision of the hierarchical structure. Figure 1 shows the results of the hierarchical relationship among the attributes for the eighth grade TIMSS 2003 mathematics test. Note that attribute 10 in B1 and attribute 12 in B2 do not have any associated items, as shown in the gray circles in Figures 2 and 3, respectively. The final hierarchies for each booklet used to specify the maximum number of possible attribute profiles were different. The number of possible attribute profiles is decreased from $2^K = 2^{14} = 16384$ to 726 for B1 and 690 for B2.

6.4 Conditions and Evaluation Indices

To understand whether the DINA-H and the DINO-H models worked in practice and provided reasonable parameter estimates, the U.S. sample was analyzed and compared via the DINA, DINA-H, DINO, and DINO-H models. To further investigate whether the DINA-H and the DINO-H models provided more stable calibration results than the conventional DINA and the conventional DINO models when sample size was smaller, the item fit indices for the large benchmark samples were analyzed and compared to the U.S. sample size via the same four models. There were eight conditions of grouping for each booklet. The study evaluated and compared model fit and item fit for each condition for the DINA, DINA-H, DINO, and DINO-H models.

Model Fit Indices. The model fit statistics used in this study included convergence, the AIC (Akaike, 1973, 1974), and the BIC (Schwarz, 1978). The δ index (de la Torre, 2008) and the item discrimination index (IDI; Robitzsch, Kiefer, George, & Uenlue, 2011) were used as the item fit criteria.

First of all, convergence was monitored and recorded for each condition. The estimated parameter difference between two iterations was set to be smaller than 0.001 as the criterion for convergence. Second, the AIC is defined as:

$$AIC = -2ln(likelihood) + 2p, \tag{10}$$

where ln(Likelihood) is the log-likelihood of the data under the model (see Equations 12 and 13) and p is the number of parameters in the model. For the conventional DINA and DINO models, $P = 2J + 2^K - 1$. For the DINA-H and DINO-H models, $P = 2J + L - 1$ where L is equal to the maximum number of possible attribute profiles specified for each unique model. For the observed data X and the attribute profiles (α):

$$Likelihood(X) = \prod_{i=1}^{I} Likelihood(X_i) = \prod_{i=1}^{I} \sum_{l=1}^{L} Likelihood(X_i \mid \alpha_l) p(\alpha_l). \tag{11}$$

Likelihood (X_i) is the marginalized likelihood of the response vector of examinee i, and $p(\alpha_\mathbf{l})$ is the prior probability of the attribute profile vector $\alpha_\mathbf{l}$.

$$ln(X) = log \prod_{i=1}^{I} Likelihood(X_i) = \sum_{i=1}^{I} logLikelihood(X_i). \tag{12}$$

For a given dataset, the larger the log-likelihood, the better the model fit; the smaller the AIC value, the better the model fit (Xu & von Davier, 2008). Third, the BIC is defined as:

$$\text{BIC} = -2\ln(likelihood) + p\ln(N), \tag{13}$$

where N is the sample size. Again, the smaller the BIC value, the better the model fit. The AIC and BIC for each condition are reported in the results section.

Item Fit Indices. The item fit indices included the δ index and the IDI. The δ index is the sequential EM-based δ-method, and serves as a discrimination index of item quality that accounts for both the slip and guessing parameters. δ_j is defined as the difference in the probabilities of correct responses between examinees in groups $\eta_j = 1$ and $\eta_j = 0$ (i.e., examinees with latent responses 1 and 0) (as cited in de la Torre, 2008, p.344) in the DINA and DINA-H models, and in groups $\omega_j = 1$ and $\omega_j = 0$ in the DINO and DINO-H models. The higher the value of δ_j, the lower the guessing and/or slip parameters are, which means the more discriminating the item is. The computational formula for δ_j for item j was as shown in Equation 5.

An additional item discrimination index applied in the study was the IDI, which provides the diagnostic accuracy for each item j. A higher IDI value means that an item has higher diagnostic accuracy with low guessing and slip. IDI is defined as:

$$\text{IDI}_j = 1 - \frac{g_j}{1 - s_j} \tag{14}$$

The mean and standard deviation of δ_j and IDI for each condition are reported and evaluated. In addition, correlation, mean, and standard deviation of both item parameters for each condition are reported.

7 Results

This section presents the calibration results from the real data analysis of the DINA-H and DINO-H models. The results for the DINA-H and DINO-H models were compared to the results for the DINA and DINO models, respectively.

7.1 DINA and DINA-H

The following paragraphs provide the results of the model fit, item fit, and item parameter estimates using the DINA and DINA-H models based on two TIMSS 2003 booklets with different sample sizes.

Model Fit. The results of model fit for both the smaller U.S. and the larger benchmark samples of both booklets show that the values of both AIC and BIC for the DINA-H model are smaller than those of the conventional DINA model because the numbers of parameters (i.e., possible attribute profiles) are largely decreased in the hierarchical models. For a given dataset, the smaller the AIC or BIC value, the better the model fit. As shown in Table 6, the differences were computed by subtracting the DINA-H condition values from those of the DINA. The positive values in the differences of AIC and BIC, thus, indicate that the DINA-H model performs better than the DINA model for both the smaller U.S. and the larger benchmark samples of both booklets.

Using 0.001 as the criteria for convergence, all the conditions took fewer than 60 cycles of iterations to reach convergence, except for the conditions of using DINA(-H) to estimate B1 benchmark data which took more than 100 cycles to converge. For additional information, the computation of the model fit indices

is illustrated as follows. The log-likelihood results for the B1 U.S. sample under the DINA and DINA-H models are -11410 and -11518 (from Equation 13), respectively. The AIC result for the DINA model is $-2\ln(\text{Likelihood}) + 2p = (-2) \times (-11410) + 2(2 \times 29 + 2^{14} - 1) = 55702$. The AIC result for the DINA-H model is $(-2) \times (-11518) + 2 \times (2 \times 29 + 726 - 1) = 24602$. Hence, the magnitudes of the model fit indices are highly sensitive to the numbers of possible attribute profiles in the model.

Item Fit. The results of item fit indices, δ and IDI, for TIMSS B1 and B2 data under the DINA and DINA-H models, are shown in Tables 7 to 10, respectively. The higher the item fit indices, the δ and IDI, the better the item fit. The differences between the DINA and the DINA-H models were computed by subtracting the DINA condition values from those of the DINA-H. The positive values in the differences of δ and IDI indicate that the DINA-H model performs better than the conventional DINA model, while the negative values in the differences indicate that the DINA model performs better than the DINA-H model. For the δ index for TIMSS B1, about 28% of the items perform better under the DINA-H model for the U.S. sample, and about 38% for the benchmark sample (see Table 7). For the IDI index, about 21% of the items have higher values under the DINA-H model for the U.S. sample, and about 31% for the benchmark sample (see Table 8). For TIMSS B2, about 20% of the items have higher d values under the DINA-H model for the U.S. sample, and about 20% for the benchmark sample (see Table 9). In terms of the IDI index, about 20% of the items produce better results under the DINA-H model for the U.S. sample, and about 27% for the benchmark sample (see Table 10). Generally speaking, in terms of item fit, items perform better in the conventional DINA model for both small and large sample sizes.

The differences between the small U.S. and large benchmark samples under the DINA and the DINA-H models were computed by subtracting the U.S. sample condition values from those of the benchmark sample condition. Results are shown in the last column of Tables 7 to 10. The positive values in the differences indicate that the model performs better under a larger sample condition than the smaller sample condition (see the highlighted cells in the tables), while the negative values in the differences indicate that the model performs better under the small sample condition than the large sample condition. For the δ index results for TIMSS B1, about 55% of the items perform better under the DINA model for the large sample, and about 31% under the DINA-H model (see Table 7). In terms of the IDI results, about 45% of the items perform better under the DINA model for the large sample, and about 21% under the DINA-H model (see Table 8). For TIMSS B2, about 23% of the items produce better results in the δ index under the DINA model for the large sample, and about 27% under the DINA-H model (see Table 9). About 20% of the items show better IDI results for TIMSS B2 under the DINA model for the large sample, and about 23% under the DINA-H model (see Table 10).

For B1, it shows that the DINA model is a better model if using larger sample sizes and the DINA-H model is more appropriate to apply under a small sample condition. However, the results for B2 are inconsistent with those found in B1.

In B2, the DINA-H model is not necessarily superior to the DINA model under a small sample condition. This may be due to the small difference in sample sizes between the U.S. and the benchmark data or the sample dependent calibration results in CDMs, and will need more analyses using different datasets to provide more evidence.

Item Parameter Estimates. The correlations of both the slip and guessing parameter estimates between the DINA and the DINA-H models are very high (i.e., all larger than 0.95) for the smaller U.S. sample for both booklets, as shown in Table 11. For the larger benchmark sample, the high correlational results are only found in B2. The correlations between the two models are slightly lower for the B1 data. The correlations of both item parameter estimates between the smaller U.S. and the larger benchmark samples are very high (i.e., all larger than 0.90) for the DINA-H model for both booklets. For the DINA model, the correlations between two sample sizes are also high for the B2 data; however, the results are less similar for the B1 data. The correlations between models are higher than the correlations between sample sizes conditions.

The results of item parameter estimates, guessing and slip, for TIMSS B1 and B2 data under the DINA and DINA-H models are shown in Tables 12 to 15, respectively. The means of the guessing and slip parameter estimates for both U.S. and benchmark data under the DINA-H model are slightly higher than those in the DINA model for both TIMSS booklets. The standard deviations of the guessing and slip parameter estimates for both U.S. and benchmark data under the DINA-H Model are slightly lower than those in the DINA model for both TIMSS booklets, except for the results of the U.S. sample in B1. Generally speaking, in terms of parameter estimates, items perform similarly under the conventional DINA and the DINA-H models for both small and large sample sizes. The means of the differences of parameter estimates between the two models are less than 0.07 for B1 and less than 0.03 for B2. The mean of the differences of parameter estimates between the small and large sample sizes are also small for both booklets.

7.2 DINO and DINO-H

This section shows the results of the model fit, item fit, and item parameter estimates from the real data analysis using the DINO and DINO-H models calibrating two TIMSS 2003 booklets with different sample sizes.

Model Fit. For both the U.S. and the benchmark samples of both booklets, the results of model fit for the DINO-H model are better than those of the conventional DINO model because the numbers of parameters are largely decreased in the conventional hierarchical models. Similar to Table 6, the values of differences in Table 16 were computed by subtracting the DINO-H conditions values from those of the DINO conditions. The positive values in the differences of AIC and BIC indicate that the DINO-H model performs better than the conventional DINO model for both the smaller U.S. and the larger benchmark samples of both booklets. This result is consistent with what is found under the DINA and DINA-H models. All the DINO(-H) conditions converged

with the maximum number of cycles equal to 73, using the same 0.001 criteria.

Item Fit. Tables 17 to 20 list the results of item fit indices, δ and IDI, for both B1 and B2 data under the DINO and DINO-H models, respectively. As for the DINA model tables, the positive values highlighted in the tables showing the differences of δ and IDI indicate that the DINO-H model performs better than the conventional DINO model. For B1, the results of δ index show that about 28% of the items perform better under the DINO-H model for the smaller U.S. sample, and about 21% of them perform better for the larger benchmark sample (see Table 17). The results of IDI index show that about 17% of the items perform better under the DINO-H model for the U.S. sample, and about 14% of them perform better for the benchmark sample (see Table 18). For B2, about 13% of the items have higher δ index results under the DINO-H model for the smaller U.S. sample, and about 23% of the items for the benchmark sample (see Table 19). For the IDI index, fewer items (about 17%) perform better under the DINO-H model for the U.S. sample than for the benchmark sample (about 23% of the items) (see Table 20). In terms of the results of item fit, items perform better under the conventional DINO model than the DINO-H model for both small and large sample sizes. DINO-H model works better than the conventional model for the smaller sample size for B1, while this is not so for B2.

As shown in Tables 17 to 20, the differences between the small U.S. and large benchmark samples under the DINO and the DINO-H models were computed by subtracting the U.S. sample condition values from those of the benchmark sample condition. The positive differences shown in the highlighted cells indicate that the model performs better under a larger sample condition than the smaller sample condition. For B1, the δ index results show that about 55% of the items perform better under the DINO model for the large sample, and about 21% of the items perform better under the DINO-H model for the large sample (see Table 17). The IDI results show that about 55% of the items under the DINO model perform better for the large sample, and about 10% of the items under the DINO-H model perform better for the large sample (see Table 18). For B2, about 33% of the items under the DINO model show higher δ index results for the large sample, and about 20% of the items under DINO-H model show higher δ index results for the large sample (see Table 19). More items (about 27%) have larger IDI index results under the DINO model for the large sample than they are under the DINO-H model (about 17%) (see Table 20). For both booklets, the results show that the DINO model is a better model if using larger sample sizes and the DINO-H model is more appropriate to apply under a small sample condition. This finding is consistent with the results of B1 data for the DINA and DINA-H models.

Item Parameter Estimates. The results of correlations of item parameter estimates between different models and different sample sizes for the DINO and DINO-H Models are listed in Table 21. The correlations between the DINO and DINO-H models for the smaller U.S. sample are very high and above 0.96 for both booklets. The correlations between the two models for the larger benchmark sample are slightly lower than the corresponding values for the smaller U.S.

sample, with the lowest correlation appearing for the guessing parameter estimates of B1. The correlations of item parameter estimates between the smaller U.S. and the larger benchmark samples for the DINO-H model are relatively high and all above 0.93 for both booklets. The correlations between the two samples for the DINO model are also high and close to the corresponding values for the DINO-H model, except for the lowest correlation (0.817) appearing for the guessing parameter estimate of B1. The DINO-H model item parameter estimates are similar for different sample sizes, but they are less similar for the DINO model.

Tables 22 to 25 present the results of item parameter estimates for TIMSS B1 and B2 data under the DINO and DINO-H models. The means of item parameter estimates for the DINO model are slightly lower than those for the DINO-H model for both samples sizes and for both booklets. The standard deviations of item parameter estimates for the two models are similar for both samples sizes and for both booklets. Comparing the results from two sample sizes for each model in both booklets, the parameter estimates are smaller for the small U.S. sample than those for the larger benchmark sample for both booklets, except for the guessing parameter for the DINO model in B1.

7.3 DINA(-H) vs. DINO(-H)

The calibration results from analyzing two TIMSS 2003 mathematics booklets with the DINA and DINA-H models were compared to the results analyzed via the DINO and DINO-H models.

Model Fit. Both results from the DINA and DINO models show that the hierarchical models have better model fit than their corresponding conventional models. The differences of model fit results between the DINA(-H) and DINO(-H) models for both the U.S. and the benchmark samples for both booklets are shown in Table 26. The differences were computed by subtracting the DINA(-H) condition values from those of the DINO(-H) condition. The positive values in the differences of AIC and BIC, thus, indicate that the DINA(-H) model performs better than the DINO(-H) model for both the smaller U.S. and the larger benchmark samples of both booklets. Comparing two booklets, the differences between the differences of the DINA/ DINO and DINA-H/DINO-H models are larger in B1 than in B2. Comparing the results in Table 6 to Table 16, the differences of the model fit indices between the conventional and the hierarchical models are larger in the pair of DINA and DINA-H comparison for both samples for both booklets, except for the results from the benchmark sample in B2, in which the DINO and DINO-H comparison shows the larger difference. This implies that the DINA model outperforms the DINO model when applying a skill hierarchy.

Item Fit. The differences of item fit results between the DINA(-H) and DINO(-H) models for both the U.S. and the benchmark samples for both booklets are shown in Tables 27 to 30. Similar to the model fit results, the negative values in the differences of δ and IDI between the DINA(-H) and DINO(-H) models mean that items in the DINA(-H) model perform better than the DINO(-H)

model. For the δ index results of the smaller U.S. sample in TIMSS B1, about 62% of the items perform better under the DINA model than the DINO model and about 59% of the items perform better under the DINA-H model than the DINO-H model (see Table 27). For the larger benchmark sample, about 41% of the items show higher δ index results under the DINA model than the DINO model, and about 62% of the items show better results under the DINA-H model than the DINO-H model. For the IDI index results of the U.S. sample in TIMSS B1, about 66% of the items perform better under the DINA model than the DINO model, and about 72% of the items perform better under the DINA-H model than the DINO-H model (see Table 28). For the larger benchmark sample, about 38% of the items show higher IDI index under the DINA model than the DINO model, and about 76% of the items show better results under the DINA-H model than the DINO-H model.

For the δ index results of the U.S. sample in TIMSS B2, about 67% of the items perform better under the DINA model than the DINO model, and about 70% of the items perform better under the DINA-H model than the DINO-H model (see Table 29). For the benchmark sample, about 57% of the items perform higher δ index results under the DINA model than the DINO model, and about 70% of the items perform better under the DINA-H model than the DINO-H model. For the IDI index results of the smaller U.S. sample in TIMSS B2, about 77% of the items perform better under the DINA model than the DINO model and about 80% of the items perform better under the DINA-H model than the DINO-H model (see Table 30). For the larger benchmark sample, about 63% of the items perform better under the DINA model than the DINO model and about 63% of the items perform better under the DINA-H model than the DINO-H model. Generally speaking, items in the DINA(-H) model show better item fit than in the DINO(-H) model.

7.4 Summary of the Results

The purposes of the study are to apply the hierarchical models of cognitive skills when using two cognitive diagnostic models, DINA and DINO, to analyze the retrofitted TIMSS 2003 eighth grade mathematics data. Attributes modified from the Common Core State Standards were adapted to construct the Q-matrices for two TIMSS 2003 Eighth Grade Mathematics booklets. A hierarchical structure of the mathematic attributes was built as well. The study evaluated the model fit (MAIC and MBIC), item fit (Delta ($\bar{\delta}_j$) and IDI), and item parameter estimates of slip and guessing for each condition. In general, the DINA-H and DINO-H models show better model fit than the conventional DINA and DINO models when skills are hierarchically ordered. The study suggests that the DINA-H/DINO-H models, instead of the conventional DINA/DINO models, should be considered when skills are hierarchically ordered.

The DINA-H and DINO-H models produce better model fit results than the conventional DINA and DINO models for both the smaller U.S. and the larger benchmark samples of both booklets. This is so because the numbers of parameters in the hierarchical models are smaller than those in the conventional

models. The item fit results are inconsistent with the model fit results. Items display better item fit in the conventional DINA and DINO models than in the DINA-H and DINO-H models for both small and large sample sizes. However, the values of item fit indices decrease (i.e., worse fit) when applying the conventional models to the smaller sample size condition, whereas the results are either very similar or sometimes become better when applying the DINA-H and DINO-H models to the smaller sample size condition. It implies that the conventional models are more sensitive to the small sample sizes, while the DINA-H and DINO-H models perform consistently across different sample sizes. The DINA and DINO models are better models if using a larger sample size, and the DINA-H and DINO-H models are superior and more appropriate to use for a small sample size. This finding supports the assumption that decreasing the number of possible attribute profiles will decrease the sample size requirement for conducting CDM calibrations.

Comparing the performances of the DINA and DINO models when applying a skill hierarchy, the results of analyzing two TIMSS 2003 mathematics datasets show that the DINA-H model outperforms the DINO-H model. The DINO model is more often to be used in medical and psychological assessments; however, the DINA model which assumes that the skills could not be compensated for each other is preferred in educational assessment. This may be the reason why the DINA model fits the TIMSS data better than the DINO model. The real data analysis shows that the DINA-H/DINO-H models outperform the conventional DINA/DINO models in the model fit results, but not in the item fit results. The hierarchical models perform consistently across various sample sizes, while the conventional models are more sensitive to and perform poorly for small sample sizes.

Limitations. First of all, the development and the misspecification of the Q-matrix and hierarchy in the real data analysis is one concern, although two independent coders separately coded the Q-matrix and construct the skill hierarchy based on the CCSS. There is still a possibility that other alternate hierarchical structures are available because teachers may use different instructions and students may use varying learning strategies and various problem-solving strategies in answering an item. Empirical and theoretical evidence needs to be provided to justify the distinct hierarchies for a test before conducting real data analysis and evaluating the fit. In addition, the misspecification of a Q-matrix would introduce bias and the resultant outcome of analysis would be questionable. Pilot studies could be helpful in validating the Q-matrix.

Sometimes inconsistent findings appear between the DINA(-H) and the DINO(-H) models, between the guessing and slip parameter estimates, between model fit and item fit indices, and between the two TIMSS booklets. This may be due to the differences in the nature of the two models and in the two fit indices. In the item fit results of the real data analysis, the DINA-H model is shown to be a better model than the conventional DINA model when the sample size is smaller in one booklet; however, this finding is not fully supported by the results based on the other booklet data. The somewhat dissimilar results between the two booklets data may be due to the differences in the items and attributes

of the two booklets. The DINO-H model is shown to be a more appropriate model with smaller sample size based on both booklets. This may be due to the small sample size difference between the U.S. and the benchmark data or sample dependent calibration results in CDMs, and will need more analyses using different datasets to provide conclusive evidence. In addition, the real data analysis is a retrofitting analysis. TIMSS study was not originally developed and intended to be analyze via CDMs.

Future Research Questions. Since a certain type of hierarchy models could have various types of structures, if a skill hierarchy based on a test with real data is available, the proposed approach can be applied to analyze the real data and examine its feasibility. The other content domains in TIMSS (i.e., geometry, measurement, and data analysis and probability) could play roles in forming different hierarchical structures. The fit of other structures from different content domains can be further examined.

The proposed approaches facilitate reporting the mastery/non-mastery of skills with different levels of cognitive loadings in future studies. If an assessment is developed based on hierarchically structured cognitive skills, and the Q-matrix for each test is built up and coded based on these skills, analyzing the tests using the proposed approaches would directly provide examinees, teachers, or parents with valuable information about levels and relationships among the skills. For example, based on the attributes from the CCSS and the hierarchical structure, test developers can build up blue-print, develop items and construct tests that are closely tie to the curriculum and map to the cognitive hierarchical structure. It will also facilitate developing parallel forms in terms of attribute levels. The reporting of the mastery and non-mastery of skills with different levels of loadings provides direct feedback of what parts are not acquired by the examinees and need more attention and time during the learning process. Instructors can also take the feedback to reflect on their teaching procedures and curricular development. Moreover, for some test batteries that target various grade levels, conducting CDM calibrations incorporating the hierarchically structured cognitive skills would help estimate both item parameters and examinee attribute profiles based on different requirements about the mastery of various levels of skills. In future studies, ideas about how students from different countries vary in reaching mastery levels of expected content knowledge and skills will provide opportunities to reform and to improve students performance by applying findings of this study to curriculum development, teacher education, and other kinds of support in education.

8 Conclusions

When cognitive skills are ordered hierarchically, leading to a smaller number of attribute profiles than the full independent attribute profiles, an appropriate model should incorporate the hierarchy in the estimation process. The DINA-H and DINO-H models are introduced to fulfill the goal of providing models whose model specifications, the relationships among attributes, possible attribute pro-

files, and Q-matrices are consistent with the theoretical background. Through the analysis conducted in the study and the evaluation indices, in general, the DINA-H and DINO-H models are deemed to be a better option with better model fit when calibrating items with hierarchically structured attributes and with smaller sample sizes.

The TIMSS data analysis shows the illustration of applying CDMs to a large-scale assessment, which demonstrates the feasibility of retrofitting. This successful application can be a promising way to provide informational feedback about examinees' mastery in varying levels of hierarchically ordered cognitive skills. This can help inform instructors to reflect on their teaching procedures and curricular development. This study contributes to education practices by incorporating skill hierarchies with assessments. The contributions include providing detailed informational feedback on students' learning progresses on varying hierarchical levels, and also promoting teacher enhancement of instructional procedures to match student development in the future. Specifically, by using the proposed models, the examinees' estimated attribute profiles can be obtained and then compared to the pre-specified attribute profiles. Using this feedback, teachers can determine whether their teaching sequence matches students' learning sequences, and whether their instructional procedures need to be modified.The study is unique in its incorporation of hierarchically structured skills into the estimation process of the conventional DINA/DINO models, by proposing the new DINA-H and DINO-H models. To sum up, the results of the study demonstrate the benefits, efficiencies, and feasibility of the proposed DINA-H and DINO-H approaches, which facilitate the reduction of possible attribute profiles in analyzing a CDM.

9 References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov & F. Csaki (Eds.), *Proceedings of the second international symposium on information theory* (pp. 267–281). Budapest: Akad. Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. doi: 10.1109/TAC.1974.1100705

Ausubel, D. P. (1968). *Educational Psychology: A Cognitive View*. New York: Holt, Rinehart and Winston, Inc.

Baddeley, A. D. (1998). *Human memory: Theory and practice*. Boston: Allyn and Bacon.

Baroody, A. J., Cibulskis, M., Lai, M-L., & Li, X. (2004). Comments on the use of learning trajectories in curriculum development and research. *Mathematical Thinking and Learning, 6*, 227-260. doi:10.1207/s15327833mtl0602_8

Battista, M. T. (2004). Applying cognition-based assessment to elementary school students' development of understanding of area and volume measurement. *Mathematical Thinking and Learning, 6*, 185-204. doi:10.1207/s15327833mtl0602_6

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How People Learn: Brain, Mind, Experience, and School*. Washington, DC: National Research Council.

Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning, 6*, 81-89. doi: 10.1207/s15327833mtl0602_1

Clements, D. H., Wilson, D. C., & Sarama, J. (2004). Young children's composition of geometric figures: A learning trajectories. *Mathematical Thinking and Learning*, *6*, 163-184. doi:10.1207/s15327833mtl0602_5

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. doi:10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115-130. doi: 10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199. doi: 10.1007/S11336-011-9207-7

de la Torre, J. (2012). *Cognitive Diagnosis Modeling: A General Framework Approach.* Session 5: Estimation of CDMs. Training session provided at the annual meeting of the National Council of Measurement Research. Vancouver, Canada.

de la Torre, J., & Douglas, J. A. (2004). Higher order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353. doi: 10.1007/BF02295640

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*, 595-624. doi: 10.1007/S11336-008-9063-2

de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement, 46,* 450–469. doi: 10.1111/j.1745-3984.2009.00092.x

de la Torre, J., & Lee, Y. S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*, *47*, 115-127. doi:10.1111/j.1745-3984.2009.00102.x

Doignon, J. P., & Falmagne, J. C. (1999). *Knowledge spaces.* New York, NY: Springer-Verlag.

Gagné, R. M., & Briggs, L. J. (1974). *Principles of instructional design.* New York: Holt. Rinehart & Winston.

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule space model and attribute hierarchy method. *Journal of Educational Measurement, 44,* 325–340. doi:10.1111/j.1745-3984.2007.00042.x

Gierl, M. J., Cui, Y., & Zhou, J. (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement, 46*, 293-313. doi: 10.1111/j.1745-3984.2009.00082.x

Gierl, M. J., Leighton, J. P., Wang, C., Zhou, J., Gokiert, R., & Tan, A. (2009). Validating Cognitive Models of Task Performance in Algebra on the SAT. (College Board Research Report No. 2009-3). New York, NY: The College Board.

Gierl, M. J., Leighton, J. P., & Hunka, S. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 242–274). Cambridge, UK: Cambridge University Press.

Gierl, M. J., & Zhou, J. (2008). Computer adaptive-attribute testing: A new approach to cognitive diagnostic assessment. *Journal of Psychology*, *216*, 29–39. doi: 10.1027/0044-3409.216.1.29

Haertel, E. H. (1989). Using restricted latent class models to map skill structure of achievement items. *Journal of Educational Measurement, 26,* 301–321. doi:10.1111/j.1745-3984.1989.tb00336.x

Haertel, E. H. (1990). Continuous and discrete latent structure models of item response data. *Psychometrika, 55,* 477–494. doi:10.1007/BF02294762

Hartz, S. M. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practicality.* Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign, IL.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191-210. doi: 10.1007/s11336-008-9089-5

Huebner, A., & Wang, C. (2011). A Note on Comparing Examinee Classification Methods for Cognitive Diagnosis Models. *Educational and Psychological Measurement, 71*, 407-419. doi: 10.1177/0013164410388832

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. doi: 10.1177/01466210122032064

Krajcik. J., Shin, N., Stevens, S. Y., & Short, H. (2010). *Using Learning Progressions to Inform the Design of Coherent Science Curriculum Materials.* Presented at the annual meeting of the American Educational Research Association. San Diego, CA.

Kuhn, D. (2001). Why development does (and does not) occur: Evidence from the domain of inductive reasoning. In J. L. McClelland & R. Siegler (Eds.), *Mechanisms of cognitive development: Behavioral and neural perspectives* (pp. 221–249). Hillsdale, NJ: Erlbaum.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement*, *41*, 205–236. doi: 10.1111/j.1745-984.2004.tb01163.x

Lesh, R., & Yoon, C. (2004). Evolving communities of mind: in which development involves several interesting and simultaneously developing strands. *Mathematical Thinking and Learning, 6*, 205-226. doi: 10.1207/

s15327833mtl0602_7

Levine, M. D., Gordon, B. N., & Reed, M. S. (1987). *Developmental variation and learning disorders.* Cambridge, MA, US: Educators Publishing Service.

Linn, M. C., Eylon, B.-S., & Davis, E. A. (2004). The knowledge integration perspective on learning. In: M. C. Linn, E. A. Davis, & P. Bell (Eds.), *Internet Environments for Science Education* (pp. 29–46). Mahwah, NJ: Lawrence Erlbaum Associates.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 178-212. doi: 10.1007/BF02294535

Martin, M. O. (Eds.). (2005). *TIMSS 2003 User Guide for the International Database.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center. Boston College.

National Council of Teachers of Mathematics (NCTM; 2000). *Principles and standards for school mathematics.* Reston, VA: NCTM.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics.* National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C.

Neidorf, T. S., & Garden, R. (2004). Developing the TIMSS 2003 mathematics and science assessment and scoring guides. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski, (Eds). *TIMSS 2003 Technical Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* (pp. 23-65). Chestnut Hill, MA: TIMSS & PIRLS International Study Center. Boston College.

Nesher, P., & Kilpatrick, J. (Eds.). (1990). *Mathematics and cognition: A research synthesis by the International Group for the Psychology of Mathematics Education.* ICMI Study Series. Cambridge: Cambridge University Press.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue. A. (2011). *CDM: Cognitive diagnosis modeling.* (Retrieved from http://cran.r-project.org/web /packages/CDM/index.html on 11/29/2011).

Rupp, A. A., & Templin, J. L. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68,* 78-96. doi: 10.1177/ 0013164407301545

Rupp, A. A., & Templin, J. L. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219-262. doi: 10.1080/15366360802490866

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic assessment: Theory, methods, and applications.* New York: Guilford.

Schmidt, W. H., Wang, H. C., & McKnight, C. C. (2005). Curriculum coherence: An examination of US mathematics and science content standards from an international perspective. *Journal of Curriculum Studies, 37*, 525-559. doi:10.1080/0022027042000294682

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. doi: 10.1214/aos/1176344136

Steeves, K. J., & Tomey, H. A. (1998). Personal written communications to the editors.

Sternberg, R. J., & Ben-Zeev, T. (1996). *The nature of mathematical thinking.* Mahwah, NJ: Lawrence Erlbaum associates, Inc.

Su, Y.-L. (2013). *Cognitive diagnostic analysis using hierarchically structured skills.* Unpublished doctoral dissertation. The University of Iowa, IA.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale NJ: Erlbaum.

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. doi: 10.1111/1467-9876.00272

Templin, J. L. (2004). *Generalized linear mixed proficiency models for cognitive diagnosis.* Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign, IL.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305. doi: 10.1037/1082-989X.11.3.287

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, *24*, 535–585. doi: 10.1016/0010-0285(92)90018-W

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model to NAEP data* (Research Report RR-08-27). Princeton, NJ: Educational Testing Service.

Table 1. The Comparison between Booklet 1 and Booklet 2

	Booklet 1	Booklet 2
Number-and-operation items	18	21
Algebra items	11	9
Total number of items	29	30
U.S. sample	757	740
Benchmark sample	1134	1114
Number of attributes	14	14
Attributes unused	10^{th}	12^{th}
Number of possible attribute profiles	726	690

Table 2. Attributes Modified from the CCSS and the Corresponding Items in TIMSS 2003 Eighth Grade Mathematics

Attribute	Booklet 1 Item	Booklet 2 Item
1. Understand concepts of a ratio and a unit rate and use language appropriately.	1, 5, 24	11, 13, 18, 23
2. Use ratio and rate reasoning to solve real-world and mathematical problems	3, 7, 14, 21, 23, 25, 28	6, 11, 13, 16, 22, 23, 27, 30
3. Compute fluently with multi-digit numbers and find common factors and multiples.	17, 19, 22	9, 19, 25, 26
4. Apply and extend previous understandings of numbers to the system of rational numbers.	8, 9, 16	1, 8, 17
5. Apply and extend previous understandings of arithmetic to algebraic expressions.	6, 11, 12, 15, 29	3, 4, 7, 15, 29
6. Reason about and solve one-variable equations and inequalities.	2, 4, 10, 13, 20, 27	2, 5, 15, 16, 18, 22, 24, 28, 29
7. Recognize and represent proportional relationships between quantities.	3, 4, 25, 28	10, 18, 23, 25, 27
8. Use proportional relationships to solve multi-step ratio and percent problems.	21, 28	11, 13, 30
9. Apply and extend previous understandings of operations with fractions to add, subtract, multiply, and divide rational numbers.	6, 8, 9, 17	1, 9, 20, 24
10. Solve real-world and mathematical problems involving the four operations with rational numbers.		20, 21
11. Solve real-life and mathematical problems using numerical and algebraic expressions and equations.	10, 27, 28	2
12. Know and apply the properties of integer exponents to generate equivalent numerical expressions.	26	
13. Compare two fractions with different numerators and different denominators; Understand a fraction a/b with $a > 1$ as a sum of fractions $1/b$.	1, 17, 18	9
14. Solve multi-step word problems posed with whole numbers and having whole-number answers using the four operations, including problems in which remainders must be interpreted. Represent these problems using equations with a letter standing for the unknown quantity; Generate a number or shape pattern that follows a given rule. Identify apparent features of the pattern that were not explicit in the rule itself.	10, 14, 27	2, 6, 12, 14
15. Use equivalent fraction as a strategy to add and subtract fractions.	1, 17, 22	9, 20, 21, 26

Table 3. Sample Items from TIMSS 2003 Mathematics Test with the Attributes

Booklet	Item ID	Content	Item	Attributes
1	M012004	Number	Alice can run 4 laps around a track in the same time that Carol can run 3 laps. When Carol has run 12 laps, how many laps has Alice run?	1. Use ratio and rate reasoning to solve real-world and mathematical problems 7. Recognize and represent proportional relationships between quantities.
2	M022253	Algebra	If $4(x+5) = 80$, then $x =$?	5. Apply and extend previous understandings of arithmetic to algebraic expressions. 6. Reason about and solve one-variable equations and inequalities.

Table 4. Q-Matrix of Booklet 1 for the Eighth Grade TIMSS 2003 Mathematics Test

	Item\Attribute	1	2	3	4	5	6	7	8	9	11	12	13	14	15	Sum
1	M012001	1	0	0	0	0	0	0	0	0	0	0	1	0	1	3
2	M012002	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
3	M012004	0	1	0	0	0	0	1	0	0	0	0	0	0	0	2
4	M012040	0	0	0	0	0	1	1	0	0	0	0	0	0	0	2
5	M012041	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
6	M012042	0	0	0	0	1	0	0	0	1	0	0	0	0	0	2
7	M032570	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
8	M032643	0	0	0	1	0	0	0	0	1	0	0	0	0	0	2
9	M012016	0	0	0	1	0	0	0	0	1	0	0	0	0	0	2
10	M012017	0	0	0	0	0	1	0	0	0	1	0	0	1	0	3
11	M022251	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
12	M022185	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
13	M022191	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
14	M022194	0	1	0	0	0	0	0	0	0	0	0	0	1	0	2
15	M022196	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
16	M022198	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
17	M022199	0	0	1	0	0	0	0	0	1	0	0	1	0	1	4
18	M022043	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
19	M022046	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
20	M022050	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
21	M022057	0	1	0	0	0	0	0	1	0	0	0	0	0	0	2
22	M022066	0	0	1	0	0	0	0	0	0	0	0	0	0	1	2
23	M022232	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
24	M022234B	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
25	M032142	0	1	0	0	0	0	1	0	0	0	0	0	0	0	2
26	M032198	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
27	M032640	0	0	0	0	0	1	0	0	0	1	0	0	1	0	3
28	M032755	0	1	0	0	0	0	1	1	0	1	0	0	0	0	4
29	M032163	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
	Sum	3	7	3	3	5	6	4	2	4	3	1	3	3	3	

Table 5. Q-Matrix of Booklet 2 for the Eighth Grade TIMSS 2003 Mathematics Test

	Item\Attribute	1	2	3	4	5	6	7	8	9	10	11	13	14	15	Sum
1	M012016	0	0	0	1	0	0	0	0	1	0	0	0	0	0	2
2	M012017	0	0	0	0	0	1	0	0	0	0	1	0	1	0	3
3	M022251	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
4	M022185	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
5	M022191	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
6	M022194	0	1	0	0	0	0	0	0	0	0	0	0	1	0	2
7	M022196	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
8	M022198	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
9	M022199	0	0	1	0	0	0	0	0	1	0	0	1	0	1	4
10	M012025	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
11	M012027	1	1	0	0	0	0	0	1	0	0	0	0	0	0	3
12	M012029	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
13	M022139	1	1	0	0	0	0	0	1	0	0	0	0	0	0	3
14	M022144	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
15	M022253	0	0	0	0	1	1	0	0	0	0	0	0	0	0	2
16	M022156	0	1	0	0	0	1	0	0	0	0	0	0	0	0	2
17	M022104	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
18	M022106	1	0	0	0	0	1	1	0	0	0	0	0	0	0	3
19	M022110	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
20	M032307	0	0	0	0	0	0	0	0	1	1	0	0	0	1	3
21	M032523	0	0	0	0	0	0	0	0	0	1	0	0	0	1	2
22	M032701	0	1	0	0	0	1	0	0	0	0	0	0	0	0	2
23	M032704	1	1	0	0	0	0	1	0	0	0	0	0	0	0	3
24	M032525	0	0	0	0	0	1	0	0	1	0	0	0	0	0	2
25	M032381	0	0	1	0	0	0	1	0	0	0	0	0	0	0	2
26	M032416	0	0	1	0	0	0	0	0	0	0	0	0	0	1	2
27	M032160	0	1	0	0	0	0	1	0	0	0	0	0	0	0	2
28	M032540	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
29	M032698	0	0	0	0	1	1	0	0	0	0	0	0	0	0	2
30	M032529	0	1	0	0	0	0	0	1	0	0	0	0	0	0	2
	Sum	4	8	4	3	5	9	5	3	4	2	1	1	4	4	

Table 6. Results of Model Fit Indices for TIMSS Data under the DINA and DINA-H Models

		Booklet 1		Booklet 2	
Model Fit		AIC	BIC	AIC	BIC
U.S. Sample	DINA	55702	131814	56946	132693
	DINA-H	24602	28226	25755	29205
	Difference	31101	103587	31191	103488
Benchmark Sample	DINA	67861	150602	70821	153295
	DINA-H	36953	40894	39679	43435
	Difference	30908	109708	31143	109859

Table 7. Results of Item Fit Index- δ for TIMSS B1 Data under the DINA and DINA-H Models

	U.S. Sample			Benchmark Sample			Benchmark - U.S.	
Item	DINA	DINA-H	Difference	DINA	DINA-H	Difference	DINA	DINA-H
1	0.697	0.649	-0.048	0.551	0.564	0.013	-0.146	-0.085
2	-0.121	-0.108	0.014	-0.128	-0.093	0.035	-0.007	0.014
3	0.652	0.594	-0.058	0.686	0.497	-0.189	0.034	-0.097
4	0.323	0.273	-0.051	0.249	0.240	-0.009	-0.075	-0.033
5	0.436	0.324	-0.111	0.465	0.318	-0.147	0.029	-0.006
6	-0.278	-0.296	-0.018	-0.339	-0.319	0.020	-0.061	-0.023
7	0.443	0.404	-0.039	0.750	0.418	-0.333	0.307	0.013
8	-0.113	-0.111	0.001	-0.051	-0.086	-0.035	0.062	0.025
9	0.406	0.408	0.002	0.381	0.320	-0.061	-0.025	-0.088
10	0.482	0.457	-0.025	0.135	0.368	0.233	-0.347	-0.089
11	0.216	0.189	-0.027	0.121	0.140	0.019	-0.095	-0.050
12	0.424	0.396	-0.028	0.218	0.332	0.114	-0.206	-0.063
13	0.546	0.409	-0.137	0.607	0.428	-0.179	0.061	0.019
14	0.593	0.470	-0.123	0.744	0.454	-0.290	0.151	-0.016
15	0.679	0.642	-0.037	0.908	0.571	-0.337	0.229	-0.071
16	0.863	0.564	-0.299	1.000	0.616	-0.384	0.137	0.052
17	0.523	0.495	-0.028	0.585	0.451	-0.134	0.061	-0.044
18	0.539	0.310	-0.229	0.947	0.628	-0.319	0.409	0.319
19	0.586	0.465	-0.122	0.958	0.516	-0.442	0.372	0.051
20	0.394	0.418	0.023	0.263	0.296	0.033	-0.132	-0.122
21	0.461	0.370	-0.091	0.548	0.400	-0.148	0.087	0.030
22	0.777	0.764	-0.012	0.643	0.638	-0.005	-0.133	-0.127
23	0.395	0.416	0.021	0.197	0.244	0.047	-0.198	-0.172
24	0.276	0.306	0.030	0.290	0.311	0.020	0.014	0.004
25	0.278	0.285	0.007	0.096	0.120	0.024	-0.181	-0.165
26	0.995	0.975	-0.020	1.000	0.958	-0.042	0.005	-0.017
27	0.334	0.345	0.011	0.928	0.087	-0.841	0.594	-0.258
28	0.484	0.458	-0.027	0.914	0.269	-0.644	0.429	-0.188
29	0.343	0.337	-0.006	0.063	0.129	0.066	-0.280	-0.208
Mean	0.436	0.387	-0.049	0.473	0.338	-0.135	0.038	-0.048
SD	0.275	0.252	0.075	0.378	0.258	0.236	0.224	0.107
Min	-0.278	-0.296	-0.299	-0.339	-0.319	-0.841	-0.347	-0.258
Max	0.995	0.975	0.030	1.000	0.958	0.233	0.594	0.319

Table 8. Results of Item Fit Index-IDI for TIMSS B1 Data under the DINA and DINA-H Models

	U.S. Sample			Benchmark Sample			Benchmark - U.S.	
Item	DINA	DINA-H	Difference	DINA	DINA-H	Difference	DINA	DINA-H
1	0.705	0.657	-0.047	0.566	0.587	0.021	-0.139	-0.070
2	-4.730	-5.054	-0.324	-14.689	-6.546	8.143	-9.959	-1.491
3	0.767	0.714	-0.053	0.799	0.633	-0.166	0.033	-0.080
4	0.323	0.273	-0.051	0.251	0.242	-0.008	-0.073	-0.031
5	0.459	0.353	-0.106	0.483	0.345	-0.138	0.024	-0.008
6	-21.590	-14.246	7.344	-8.912	-4.269	4.643	12.678	9.977
7	0.485	0.437	-0.048	0.805	0.436	-0.369	0.320	-0.001
8	-1.590	-1.551	0.039	-0.458	-0.996	-0.538	1.132	0.555
9	0.480	0.482	0.001	0.428	0.381	-0.047	-0.052	-0.101
10	0.528	0.507	-0.020	0.167	0.394	0.226	-0.360	-0.114
11	0.587	0.546	-0.041	0.354	0.393	0.040	-0.234	-0.153
12	0.599	0.583	-0.016	0.375	0.517	0.142	-0.224	-0.067
13	0.723	0.559	-0.164	0.667	0.489	-0.178	-0.056	-0.070
14	0.647	0.549	-0.098	0.744	0.532	-0.212	0.097	-0.017
15	0.710	0.702	-0.007	0.909	0.640	-0.269	0.199	-0.062
16	0.998	0.694	-0.304	1.000	0.659	-0.341	0.002	-0.035
17	0.725	0.711	-0.014	0.719	0.672	-0.047	-0.006	-0.038
18	0.580	0.346	-0.234	1.000	0.719	-0.281	0.420	0.372
19	0.627	0.503	-0.123	1.000	0.561	-0.439	0.373	0.058
20	0.724	0.699	-0.024	0.592	0.606	0.014	-0.131	-0.093
21	0.481	0.395	-0.085	0.548	0.400	-0.148	0.067	0.005
22	0.832	0.803	-0.029	0.721	0.716	-0.006	-0.110	-0.087
23	0.947	0.908	-0.039	0.966	0.885	-0.081	0.019	-0.023
24	0.983	0.968	-0.015	0.953	0.931	-0.021	-0.030	-0.036
25	0.429	0.432	0.003	0.184	0.222	0.038	-0.245	-0.210
26	1.000	0.996	-0.004	1.000	1.000	0.000	0.000	0.003
27	0.982	0.983	0.001	0.931	0.469	-0.463	-0.051	-0.515
28	0.995	0.985	-0.011	0.914	0.731	-0.183	-0.082	-0.254
29	0.516	0.519	0.003	0.139	0.265	0.126	-0.377	-0.254
Mean	-0.348	-0.157	0.191	-0.236	0.090	0.326	0.112	0.247
SD	4.229	2.937	1.379	3.313	1.591	1.754	3.058	1.899
Min	-21.590	-14.246	-0.324	-14.689	-6.546	-0.538	-9.959	-1.491
Max	1.000	0.996	7.344	1.000	1.000	8.143	12.678	9.977

Table 9. Results of Item Fit Index- δ for TIMSS B2 Data under the DINA and DINA-H Models

	U.S. Sample			Benchmark Sample			Benchmark - U.S.	
Item	DINA	DINA-H	Difference	DINA	DINA-H	Difference	DINA	DINA-H
1	0.415	0.394	-0.022	0.335	0.320	-0.015	-0.081	-0.074
2	0.683	0.643	-0.041	0.492	0.375	-0.117	-0.191	-0.268
3	0.121	0.099	-0.021	0.074	0.049	-0.026	-0.046	-0.051
4	0.343	0.307	-0.036	0.313	0.296	-0.017	-0.030	-0.011
5	0.366	0.329	-0.036	0.330	0.290	-0.040	-0.036	-0.040
6	0.363	0.364	0.001	0.429	0.425	-0.004	0.066	0.061
7	0.679	0.596	-0.083	0.509	0.443	-0.066	-0.170	-0.153
8	0.308	0.337	0.028	0.307	0.336	0.029	-0.002	-0.001
9	0.815	0.478	-0.337	0.743	0.497	-0.246	-0.072	0.020
10	0.391	0.303	-0.087	0.257	0.194	-0.063	-0.133	-0.109
11	0.472	0.488	0.016	0.493	0.545	0.052	0.021	0.057
12	0.562	0.527	-0.035	0.430	0.409	-0.021	-0.132	-0.118
13	-0.031	-0.030	0.002	-0.027	-0.018	0.008	0.005	0.011
14	0.546	0.533	-0.013	0.491	0.444	-0.047	-0.055	-0.089
15	0.576	0.557	-0.019	0.517	0.492	-0.025	-0.059	-0.065
16	0.808	0.660	-0.147	0.690	0.594	-0.096	-0.117	-0.066
17	0.440	0.385	-0.056	0.504	0.412	-0.092	0.064	0.027
18	0.440	0.371	-0.069	0.573	0.449	-0.124	0.133	0.078
19	0.269	0.261	-0.008	0.107	0.283	0.176	-0.162	0.022
20	0.703	0.669	-0.034	0.813	0.746	-0.067	0.110	0.077
21	0.411	0.361	-0.050	0.397	0.360	-0.036	-0.015	-0.001
22	0.166	0.168	0.002	0.123	0.118	-0.006	-0.043	-0.051
23	0.485	0.497	0.012	0.318	0.386	0.068	-0.167	-0.112
24	0.606	0.547	-0.059	0.335	0.311	-0.024	-0.271	-0.236
25	0.735	0.640	-0.095	0.542	0.293	-0.249	-0.192	-0.347
26	0.480	0.462	-0.018	0.476	0.217	-0.259	-0.004	-0.246
27	0.230	0.210	-0.019	0.300	0.159	-0.141	0.070	-0.052
28	0.230	0.194	-0.036	0.148	0.129	-0.019	-0.082	-0.065
29	0.471	0.452	-0.018	0.018	0.018	0.000	-0.453	-0.434
30	0.619	0.350	-0.270	0.417	0.158	-0.259	-0.202	-0.192
Mean	0.457	0.405	-0.052	0.382	0.324	-0.058	-0.075	-0.081
SD	0.204	0.173	0.078	0.204	0.175	0.099	0.122	0.125
Min	-0.031	-0.030	-0.337	-0.027	-0.018	-0.259	-0.453	-0.434
Max	0.815	0.669	0.028	0.813	0.746	0.176	0.133	0.078

Table 10. Results of Item Fit Index-IDI for TIMSS B2 Data under the DINA and DINA-H Models

	U.S. Sample			Benchmark Sample			Benchmark - U.S.	
Item	DINA	DINA-H	Difference	DINA	DINA-H	Difference	DINA	DINA-H
1	0.560	0.532	-0.028	0.453	0.434	-0.018	-0.107	-0.098
2	0.686	0.654	-0.032	0.498	0.420	-0.079	-0.188	-0.235
3	0.386	0.326	-0.060	0.219	0.149	-0.070	-0.167	-0.177
4	0.522	0.474	-0.048	0.476	0.461	-0.015	-0.046	-0.013
5	0.557	0.488	-0.070	0.401	0.348	-0.053	-0.156	-0.140
6	0.473	0.477	0.005	0.490	0.512	0.022	0.018	0.035
7	0.744	0.671	-0.073	0.550	0.501	-0.050	-0.193	-0.170
8	0.493	0.509	0.016	0.459	0.481	0.022	-0.034	-0.028
9	0.822	0.724	-0.098	0.754	0.693	-0.061	-0.068	-0.031
10	0.414	0.323	-0.091	0.305	0.226	-0.079	-0.109	-0.097
11	0.504	0.519	0.015	0.518	0.567	0.049	0.014	0.048
12	0.640	0.605	-0.035	0.522	0.528	0.007	-0.118	-0.077
13	-0.385	-0.360	0.025	-0.396	-0.248	0.147	-0.011	0.112
14	0.623	0.607	-0.016	0.587	0.577	-0.011	-0.036	-0.031
15	0.710	0.698	-0.013	0.640	0.635	-0.005	-0.070	-0.063
16	0.952	0.831	-0.121	0.841	0.746	-0.095	-0.111	-0.085
17	0.448	0.391	-0.057	0.509	0.419	-0.090	0.061	0.029
18	0.776	0.732	-0.044	0.898	0.820	-0.078	0.122	0.088
19	0.280	0.270	-0.009	0.127	0.312	0.185	-0.153	0.041
20	0.944	0.936	-0.008	0.938	0.889	-0.049	-0.006	-0.047
21	0.701	0.631	-0.070	0.811	0.714	-0.097	0.110	0.083
22	0.171	0.172	0.001	0.124	0.118	-0.006	-0.047	-0.054
23	0.502	0.535	0.033	0.325	0.406	0.081	-0.177	-0.129
24	0.658	0.599	-0.060	0.409	0.380	-0.029	-0.249	-0.218
25	0.825	0.725	-0.100	0.675	0.404	-0.270	-0.150	-0.321
26	0.994	0.975	-0.019	0.987	0.543	-0.444	-0.007	-0.432
27	0.721	0.682	-0.039	0.926	0.573	-0.353	0.204	-0.110
28	0.320	0.267	-0.052	0.220	0.191	-0.029	-0.099	-0.077
29	0.786	0.771	-0.016	0.043	0.043	0.000	-0.743	-0.728
30	0.896	0.562	-0.334	0.723	0.324	-0.399	-0.172	-0.238
Mean	0.591	0.544	-0.047	0.501	0.439	-0.062	-0.090	-0.105
SD	0.276	0.256	0.067	0.305	0.240	0.140	0.161	0.170
Min	-0.385	-0.360	-0.334	-0.396	-0.248	-0.444	-0.743	-0.728
Max	0.994	0.975	0.033	0.987	0.889	0.185	0.204	0.112

Table 11. Correlations of Item Parameter Estimates between Different Models and Sample Sizes for the DINA and DINA-H Models

Between DINA and DINA-H			
		Smaller U.S. Sample	Larger Benchmark Sample
B1	Guessing	0.951	0.794
	Slip	0.997	0.827
B2	Guessing	0.974	0.952
	Slip	0.962	0.967
Between Small and Large Samples			
		DINA	DINA-H
B1	Guessing	0.748	0.911
	Slip	0.853	0.972
B2	Guessing	0.919	0.901
	Slip	0.947	0.943

Table 12. Results of Guessing Parameter Estimates for TIMSS B1 Data under the DINA and DINA-H Models

	U.S. Sample		Benchmark Sample	
Item	DINA	DINA-H	DINA	DINA-H
1	0.292	0.338	0.423	0.396
2	0.147	0.129	0.137	0.107
3	0.199	0.239	0.172	0.288
4	0.677	0.727	0.744	0.751
5	0.514	0.595	0.497	0.603
6	0.291	0.317	0.377	0.394
7	0.471	0.522	0.182	0.540
8	0.184	0.183	0.162	0.173
9	0.440	0.439	0.510	0.520
10	0.432	0.444	0.672	0.567
11	0.152	0.157	0.221	0.216
12	0.284	0.283	0.363	0.311
13	0.209	0.323	0.303	0.448
14	0.324	0.386	0.256	0.400
15	0.278	0.272	0.091	0.321
16	0.002	0.249	0.000	0.319
17	0.199	0.201	0.228	0.220
18	0.389	0.585	0.000	0.246
19	0.349	0.458	0.000	0.403
20	0.151	0.180	0.181	0.192
21	0.498	0.566	0.452	0.600
22	0.157	0.188	0.248	0.254
23	0.022	0.042	0.007	0.032
24	0.005	0.010	0.014	0.023
25	0.370	0.374	0.427	0.421
26	0.000	0.003	0.000	0.000
27	0.006	0.006	0.068	0.099
28	0.002	0.007	0.086	0.099
29	0.322	0.312	0.392	0.358
Mean	0.254	0.294	0.249	0.321
SD	0.179	0.195	0.206	0.192
Min	0.000	0.003	0.000	0.000
Max	0.677	0.727	0.744	0.751

Table 13. Results of Slip Parameter Estimates for TIMSS B1 Data under the DINA and DINA-H Models

	U.S. Sample		Benchmark Sample	
Item	DINA	DINA-H	DINA	DINA-H
1	0.011	0.013	0.026	0.040
2	0.974	0.979	0.991	0.986
3	0.149	0.167	0.142	0.214
4	0.000	0.000	0.007	0.009
5	0.051	0.080	0.038	0.079
6	0.987	0.979	0.962	0.925
7	0.086	0.074	0.068	0.042
8	0.929	0.928	0.889	0.913
9	0.154	0.152	0.109	0.159
10	0.086	0.099	0.193	0.065
11	0.632	0.653	0.659	0.645
12	0.293	0.322	0.419	0.357
13	0.244	0.267	0.090	0.124
14	0.083	0.144	0.000	0.146
15	0.043	0.086	0.001	0.107
16	0.135	0.187	0.000	0.065
17	0.278	0.304	0.187	0.329
18	0.072	0.106	0.053	0.126
19	0.064	0.077	0.042	0.081
20	0.455	0.403	0.556	0.513
21	0.041	0.064	0.000	0.000
22	0.066	0.048	0.108	0.109
23	0.583	0.542	0.796	0.724
24	0.719	0.683	0.695	0.667
25	0.352	0.341	0.476	0.459
26	0.005	0.021	0.000	0.042
27	0.660	0.649	0.003	0.814
28	0.513	0.535	0.000	0.632
29	0.335	0.351	0.545	0.512
Mean	0.310	0.319	0.278	0.341
SD	0.312	0.302	0.336	0.321
Min	0.000	0.000	0.000	0.000
Max	0.987	0.979	0.991	0.986

Table 14. Results of Guessing Parameter Estimates for TIMSS B2 Data under the DINA and DINA-H Models

	U.S. Sample		Benchmark Sample	
Item	DINA	DINA-H	DINA	DINA-H
1	0.326	0.346	0.404	0.417
2	0.313	0.340	0.495	0.519
3	0.192	0.205	0.265	0.278
4	0.314	0.340	0.345	0.346
5	0.290	0.346	0.493	0.544
6	0.405	0.398	0.446	0.404
7	0.234	0.292	0.416	0.442
8	0.317	0.325	0.362	0.362
9	0.176	0.182	0.242	0.221
10	0.552	0.636	0.586	0.664
11	0.464	0.453	0.459	0.417
12	0.316	0.344	0.394	0.365
13	0.113	0.113	0.095	0.093
14	0.330	0.345	0.345	0.326
15	0.235	0.242	0.291	0.283
16	0.041	0.135	0.130	0.203
17	0.543	0.600	0.486	0.570
18	0.127	0.136	0.065	0.099
19	0.693	0.705	0.737	0.626
20	0.042	0.045	0.054	0.093
21	0.175	0.211	0.092	0.145
22	0.806	0.809	0.869	0.877
23	0.481	0.433	0.661	0.564
24	0.315	0.367	0.485	0.508
25	0.156	0.242	0.261	0.431
26	0.003	0.012	0.006	0.182
27	0.089	0.098	0.024	0.118
28	0.489	0.531	0.524	0.547
29	0.128	0.134	0.397	0.396
30	0.072	0.273	0.160	0.330
Mean	0.291	0.321	0.353	0.379
SD	0.197	0.193	0.216	0.190
Min	0.003	0.012	0.006	0.093
Max	0.806	0.809	0.869	0.877

Table 15. Results of Slip Parameter Estimates for TIMSS B2 Data under the DINA and DINA-H Models

	U.S. Sample		Benchmark Sample	
Item	DINA	DINA-H	DINA	DINA-H
1	0.258	0.260	0.261	0.263
2	0.004	0.018	0.013	0.106
3	0.688	0.695	0.661	0.673
4	0.343	0.353	0.342	0.358
5	0.344	0.324	0.178	0.167
6	0.232	0.238	0.125	0.171
7	0.086	0.112	0.075	0.115
8	0.375	0.338	0.332	0.302
9	0.009	0.340	0.015	0.282
10	0.057	0.061	0.156	0.142
11	0.064	0.059	0.047	0.038
12	0.121	0.129	0.175	0.226
13	0.919	0.917	0.932	0.926
14	0.124	0.122	0.163	0.230
15	0.188	0.201	0.192	0.224
16	0.152	0.205	0.179	0.203
17	0.017	0.015	0.010	0.018
18	0.433	0.493	0.362	0.452
19	0.038	0.034	0.156	0.091
20	0.256	0.286	0.134	0.161
21	0.413	0.428	0.511	0.495
22	0.028	0.023	0.008	0.006
23	0.034	0.070	0.021	0.050
24	0.079	0.085	0.179	0.181
25	0.110	0.118	0.196	0.276
26	0.517	0.526	0.518	0.601
27	0.682	0.692	0.676	0.723
28	0.281	0.275	0.328	0.323
29	0.401	0.413	0.585	0.586
30	0.308	0.377	0.423	0.512
Mean	0.252	0.274	0.265	0.297
SD	0.229	0.226	0.232	0.229
Min	0.004	0.015	0.008	0.006
Max	0.919	0.917	0.932	0.926

Table 16. Results of Model Fit Indices for TIMSS Data under the DINO and DINO-H Models

	.	Booklet 1		Booklet 2	
Model Fit		AIC	BIC	AIC	BIC
U.S. Sample	DINO	55806	131917	57048	132795
	DINO-H	24745	28370	25868	29319
	Difference	31061	103548	31179	103476
Benchmark Sample	DINO	67897	150638	70889	153363
	DINO-H	37094	41035	39771	43528
	Difference	30802	109603	31211	109927

Table 17. Results of Item Fit Index- δ for TIMSS B1 Data under the DINO and DINO-H Models

	U.S. Sample			Benchmark Sample			Benchmark - U.S.	
Item	DINO	DINO-H	Difference	DINO	DINO-H	Difference	DINO	DINO-H
1	0.738	0.618	-0.120	0.934	0.625	-0.309	0.196	0.007
2	-0.131	-0.125	0.006	-0.098	-0.090	0.008	0.033	0.035
3	0.690	0.549	-0.142	0.729	0.442	-0.287	0.038	-0.107
4	0.338	0.293	-0.044	0.376	0.248	-0.128	0.038	-0.045
5	0.388	0.320	-0.068	0.407	0.313	-0.093	0.019	-0.006
6	-0.324	-0.262	0.062	-0.339	-0.201	0.138	-0.016	0.061
7	0.404	0.395	-0.009	0.379	0.396	0.017	-0.026	0.001
8	-0.055	-0.098	-0.043	-0.004	-0.056	-0.052	0.051	0.042
9	0.565	0.411	-0.154	0.386	0.248	-0.138	-0.178	-0.163
10	0.545	0.437	-0.108	0.821	0.266	-0.555	0.276	-0.171
11	0.210	0.199	-0.012	0.165	0.133	-0.032	-0.045	-0.066
12	0.420	0.402	-0.018	0.339	0.334	-0.005	-0.081	-0.068
13	0.491	0.437	-0.055	0.497	0.406	-0.091	0.006	-0.030
14	0.567	0.462	-0.105	0.761	0.473	-0.288	0.194	0.012
15	0.679	0.600	-0.079	0.697	0.502	-0.195	0.018	-0.098
16	0.716	0.555	-0.161	0.990	0.520	-0.471	0.274	-0.035
17	0.355	0.375	0.020	0.434	0.304	-0.131	0.079	-0.072
18	0.335	0.224	-0.111	0.349	0.146	-0.203	0.014	-0.078
19	0.451	0.424	-0.027	0.436	0.344	-0.091	-0.015	-0.079
20	0.454	0.447	-0.006	0.331	0.324	-0.008	-0.122	-0.123
21	0.413	0.267	-0.146	0.925	0.236	-0.689	0.512	-0.032
22	0.831	0.715	-0.116	0.807	0.614	-0.193	-0.024	-0.101
23	0.503	0.440	-0.063	0.270	0.249	-0.022	-0.233	-0.191
24	0.303	0.339	0.036	0.313	0.317	0.004	0.011	-0.022
25	0.256	0.289	0.033	0.055	0.119	0.064	-0.200	-0.170
26	0.996	0.754	-0.242	0.998	0.226	-0.773	0.002	-0.528
27	0.226	0.231	0.005	0.092	0.060	-0.033	-0.134	-0.171
28	0.184	0.234	0.051	0.157	0.118	-0.039	-0.026	-0.116
29	0.333	0.344	0.011	0.138	0.140	0.002	-0.195	-0.204
Mean	0.410	0.354	-0.055	0.426	0.267	-0.158	0.016	-0.087
SD	0.279	0.228	0.075	0.344	0.197	0.220	0.159	0.112
Min	-0.324	-0.262	-0.242	-0.339	-0.201	-0.773	-0.233	-0.528
Max	0.996	0.754	0.062	0.998	0.625	0.138	0.512	0.061

Table 18. Results of Item Fit Index-IDI for TIMSS B1 Data under the DINO and DINO-H Models

	U.S. Sample			Benchmark Sample			Benchmark - U.S.	
Item	DINO	DINO-H	Difference	DINO	DINO-H	Difference	DINO	DINO-H
1	0.792	0.681	-0.111	1.000	0.670	-0.330	0.208	-0.011
3	0.818	0.678	-0.140	0.883	0.591	-0.292	0.066	-0.087
4	0.343	0.297	-0.046	0.385	0.254	-0.130	0.042	-0.043
5	0.406	0.341	-0.065	0.426	0.336	-0.089	0.020	-0.004
6	-11.841	-6.722	5.119	-1.742	-1.134	0.608	10.100	5.589
7	0.412	0.417	0.005	0.381	0.411	0.030	-0.031	-0.006
8	-0.467	-1.158	-0.691	-0.028	-0.489	-0.462	0.439	0.668
9	0.652	0.494	-0.158	0.554	0.326	-0.229	-0.098	-0.168
10	0.634	0.531	-0.103	0.992	0.331	-0.661	0.358	-0.200
11	0.564	0.543	-0.021	0.453	0.368	-0.085	-0.111	-0.175
12	0.582	0.565	-0.017	0.530	0.502	-0.027	-0.053	-0.063
13	0.607	0.553	-0.053	0.541	0.457	-0.084	-0.066	-0.097
14	0.632	0.552	-0.081	0.823	0.572	-0.252	0.191	0.020
15	0.691	0.640	-0.051	0.744	0.562	-0.182	0.053	-0.078
16	0.751	0.638	-0.113	0.990	0.567	-0.423	0.240	-0.071
17	0.688	0.680	-0.008	0.994	0.614	-0.380	0.306	-0.066
18	0.335	0.244	-0.091	0.354	0.167	-0.187	0.019	-0.077
19	0.451	0.431	-0.019	0.436	0.359	-0.076	-0.015	-0.072
20	0.702	0.682	-0.020	0.649	0.620	-0.029	-0.053	-0.062
21	0.464	0.320	-0.145	0.991	0.283	-0.708	0.527	-0.037
22	0.874	0.797	-0.077	0.880	0.693	-0.187	0.006	-0.105
23	0.876	0.878	0.002	0.826	0.863	0.037	-0.050	-0.015
24	0.956	0.949	-0.007	0.956	0.907	-0.049	0.000	-0.042
25	0.409	0.435	0.025	0.111	0.221	0.110	-0.298	-0.214
26	1.000	0.799	-0.201	1.000	0.476	-0.524	0.000	-0.322
27	1.000	0.971	-0.029	0.652	0.392	-0.260	-0.348	-0.580
28	1.000	0.981	-0.019	1.000	0.587	-0.413	0.000	-0.395
29	0.495	0.507	0.012	0.281	0.279	-0.002	-0.214	-0.228
Mean	0.172	0.276	0.104	0.574	0.385	-0.188	0.401	0.109
SD	2.373	1.425	0.992	0.543	0.394	0.261	1.911	1.092
Min	-11.841	-6.722	-0.691	-1.742	-1.134	-0.708	-0.348	-0.580
Max	1.000	0.981	5.119	1.000	0.907	0.608	10.100	5.589

Note. Item 2 was removed because its IDI of the DINO model is -65444066333947.9 and is -19610.25 of the DINO-H model for the U.S. sample.

Table 19. Results of Item Fit Index- δ for TIMSS B2 Data under the DINO and DINO-H Models

	U.S. Sample			Benchmark Sample			Benchmark - U.S.	
Item	DINO	DINO-H	Difference	DINO	DINO-H	Difference	DINO	DINO-H
1	0.381	0.344	-0.038	0.293	0.288	-0.005	-0.088	-0.056
2	0.708	0.355	-0.353	0.755	0.244	-0.511	0.047	-0.111
3	0.096	0.099	0.003	0.073	0.055	-0.018	-0.023	-0.044
4	0.345	0.322	-0.023	0.287	0.281	-0.006	-0.058	-0.041
5	0.367	0.342	-0.025	0.275	0.296	0.022	-0.092	-0.046
6	0.339	0.338	-0.001	0.447	0.392	-0.055	0.108	0.054
7	0.647	0.556	-0.091	0.513	0.404	-0.110	-0.134	-0.152
8	0.361	0.339	-0.022	0.301	0.291	-0.010	-0.060	-0.048
9	0.658	0.582	-0.076	0.542	0.493	-0.049	-0.116	-0.089
10	0.298	0.293	-0.004	0.203	0.213	0.010	-0.095	-0.080
11	0.552	0.437	-0.115	0.489	0.390	-0.100	-0.063	-0.048
12	0.568	0.516	-0.053	0.382	0.382	0.000	-0.186	-0.134
13	-0.036	-0.025	0.011	0.001	-0.001	-0.002	0.038	0.024
14	0.505	0.494	-0.010	0.455	0.408	-0.047	-0.050	-0.086
15	0.564	0.535	-0.029	0.526	0.486	-0.040	-0.038	-0.049
16	0.649	0.622	-0.028	0.653	0.583	-0.069	0.003	-0.038
17	0.373	0.338	-0.035	0.414	0.328	-0.086	0.041	-0.010
18	0.355	0.333	-0.021	0.454	0.427	-0.027	0.100	0.094
19	0.129	0.141	0.012	0.054	0.174	0.119	-0.075	0.033
20	0.671	0.598	-0.073	0.759	0.664	-0.095	0.088	0.066
21	0.468	0.436	-0.032	0.438	0.467	0.029	-0.030	0.031
22	0.199	0.185	-0.015	0.131	0.126	-0.005	-0.069	-0.059
23	0.553	0.486	-0.067	0.377	0.333	-0.044	-0.177	-0.153
24	0.533	0.510	-0.023	0.338	0.304	-0.034	-0.195	-0.206
25	0.705	0.619	-0.086	0.567	0.300	-0.267	-0.137	-0.319
26	0.487	0.328	-0.160	0.500	0.189	-0.311	0.013	-0.138
27	0.210	0.217	0.007	0.331	0.123	-0.209	0.121	-0.094
28	0.221	0.219	-0.002	0.120	0.134	0.014	-0.100	-0.085
29	0.452	0.446	-0.006	0.014	0.025	0.011	-0.438	-0.421
30	0.481	0.293	-0.189	0.529	0.111	-0.418	0.048	-0.181
Mean	0.428	0.377	-0.051	0.374	0.297	-0.077	-0.054	-0.080
SD	0.192	0.161	0.075	0.204	0.163	0.137	0.113	0.108
Min	-0.036	-0.025	-0.353	0.001	-0.001	-0.511	-0.438	-0.421
Max	0.708	0.622	0.012	0.759	0.664	0.119	0.121	0.094

Table 20. Results of Item Fit Index-IDI for TIMSS B2 Data under the DINO and DINO-H Models

	U.S. Sample			Benchmark Sample			Benchmark - U.S.	
Item	DINO	DINO-H	Difference	DINO	DINO-H	Difference	DINO	DINO-H
1	0.537	0.487	-0.050	0.426	0.406	-0.020	-0.111	-0.082
2	0.983	0.515	-0.468	0.984	0.327	-0.657	0.001	-0.188
3	0.305	0.318	0.014	0.211	0.165	-0.047	-0.093	-0.154
4	0.491	0.477	-0.014	0.428	0.428	-0.001	-0.063	-0.049
5	0.498	0.474	-0.024	0.326	0.344	0.018	-0.172	-0.130
6	0.447	0.454	0.007	0.521	0.489	-0.033	0.075	0.035
7	0.657	0.612	-0.045	0.530	0.450	-0.080	-0.127	-0.163
8	0.525	0.492	-0.033	0.435	0.409	-0.025	-0.090	-0.082
9	1.000	0.975	-0.025	0.999	0.953	-0.046	-0.001	-0.022
10	0.315	0.306	-0.008	0.218	0.234	0.015	-0.096	-0.073
11	0.606	0.503	-0.103	0.595	0.470	-0.125	-0.012	-0.033
12	0.596	0.581	-0.015	0.465	0.482	0.018	-0.131	-0.098
13	-0.438	-0.287	0.150	0.017	-0.011	-0.027	0.454	0.277
14	0.547	0.561	0.014	0.540	0.518	-0.021	-0.007	-0.042
15	0.695	0.663	-0.032	0.691	0.628	-0.063	-0.004	-0.035
16	0.816	0.799	-0.017	0.801	0.755	-0.046	-0.015	-0.044
17	0.376	0.343	-0.033	0.416	0.334	-0.082	0.040	-0.009
18	0.772	0.719	-0.054	0.902	0.835	-0.067	0.130	0.116
19	0.131	0.144	0.013	0.064	0.193	0.130	-0.067	0.050
20	0.966	0.935	-0.031	0.966	0.886	-0.080	0.000	-0.049
21	0.700	0.684	-0.016	0.769	0.775	0.007	0.069	0.091
22	0.200	0.187	-0.013	0.131	0.126	-0.005	-0.069	-0.061
23	0.595	0.517	-0.078	0.390	0.341	-0.049	-0.205	-0.176
24	0.588	0.569	-0.019	0.409	0.372	-0.037	-0.179	-0.197
25	0.786	0.721	-0.065	0.620	0.440	-0.180	-0.166	-0.281
26	0.852	0.779	-0.073	0.794	0.516	-0.278	-0.058	-0.263
27	0.681	0.680	-0.002	0.740	0.439	-0.301	0.059	-0.241
28	0.289	0.287	-0.002	0.177	0.194	0.017	-0.112	-0.093
29	0.761	0.744	-0.016	0.035	0.060	0.025	-0.725	-0.685
30	0.684	0.523	-0.161	0.691	0.248	-0.443	0.007	-0.275
Mean	0.565	0.525	-0.040	0.510	0.427	-0.083	-0.056	-0.099
SD	0.292	0.251	0.095	0.287	0.238	0.155	0.177	0.165
Min	-0.438	-0.287	-0.468	0.017	-0.011	-0.657	-0.725	-0.685
Max	1.000	0.975	0.150	0.999	0.953	0.130	0.454	0.277

Table 21. Correlations of Item Parameter Estimates between Different Models and Sample Sizes for the DINO and DINO-H Models

Between DINO and DINO-H			
		Smaller U.S. Sample	Larger Benchmark Sample
B1	Guessing	0.967	0.697
	Slip	0.995	0.950
B2	Guessing	0.960	0.917
	Slip	0.983	0.920
Between Small and Large Samples			
		DINO	DINO-H
B1	Guessing	0.817	0.975
	Slip	0.967	0.936
B2	Guessing	0.937	0.930
	Slip	0.935	0.948

Table 22. Results of Guessing Parameter Estimates for TIMSS B1 Data under the DINO and DINO-H Models

	U.S. Sample		Benchmark Sample	
Item	DINO	DINO-H	DINO	DINO-H
1	0.194	0.289	0.000	0.308
2	0.131	0.125	0.109	0.100
3	0.154	0.260	0.096	0.306
4	0.647	0.694	0.601	0.727
5	0.569	0.619	0.548	0.618
6	0.351	0.301	0.534	0.378
7	0.577	0.553	0.615	0.568
8	0.173	0.183	0.151	0.170
9	0.301	0.421	0.310	0.514
10	0.315	0.386	0.007	0.537
11	0.162	0.167	0.199	0.228
12	0.301	0.309	0.301	0.331
13	0.318	0.352	0.421	0.483
14	0.330	0.376	0.163	0.355
15	0.304	0.338	0.240	0.392
16	0.238	0.315	0.010	0.396
17	0.161	0.176	0.003	0.191
18	0.665	0.693	0.637	0.729
19	0.549	0.558	0.564	0.614
20	0.192	0.209	0.179	0.198
21	0.477	0.569	0.008	0.597
22	0.120	0.182	0.111	0.272
23	0.071	0.061	0.057	0.039
24	0.014	0.018	0.014	0.033
25	0.369	0.375	0.444	0.419
26	0.000	0.190	0.000	0.248
27	0.000	0.007	0.049	0.093
28	0.000	0.004	0.000	0.083
29	0.339	0.334	0.351	0.361
Mean	0.277	0.313	0.232	0.355
SD	0.194	0.196	0.223	0.202
Min	0.000	0.004	0.000	0.033
Max	0.665	0.694	0.637	0.729

Table 23. Results of Slip Parameter Estimates for TIMSS B1 Data under the DINO and DINO-H Models

	U.S. Sample		Benchmark Sample	
Item	DINO	DINO-H	DINO	DINO-H
1	0.068	0.092	0.066	0.067
2	1.000	1.000	0.990	0.990
3	0.156	0.191	0.175	0.252
4	0.016	0.013	0.023	0.025
5	0.043	0.061	0.045	0.068
6	0.973	0.961	0.805	0.823
7	0.018	0.052	0.006	0.036
8	0.882	0.915	0.853	0.886
9	0.134	0.168	0.303	0.238
10	0.140	0.177	0.172	0.197
11	0.627	0.634	0.636	0.639
12	0.279	0.289	0.360	0.335
13	0.190	0.211	0.081	0.110
14	0.103	0.162	0.076	0.172
15	0.017	0.062	0.063	0.106
16	0.046	0.130	0.000	0.084
17	0.484	0.448	0.563	0.506
18	0.000	0.083	0.014	0.125
19	0.000	0.018	0.000	0.042
20	0.354	0.344	0.490	0.478
21	0.110	0.163	0.067	0.167
22	0.049	0.103	0.082	0.114
23	0.426	0.499	0.673	0.712
24	0.683	0.642	0.672	0.651
25	0.375	0.336	0.501	0.462
26	0.004	0.056	0.002	0.526
27	0.774	0.762	0.858	0.847
28	0.816	0.761	0.843	0.799
29	0.328	0.322	0.511	0.499
Mean	0.314	0.333	0.342	0.378
SD	0.327	0.307	0.333	0.306
Min	0.000	0.013	0.000	0.025
Max	1.000	1.000	0.990	0.990

Table 24. Results of Guessing Parameter Estimates for TIMSS B2 Data under the DINO and DINO-H Models

	U.S. Sample		Benchmark Sample	
Item	DINO	DINO-H	DINO	DINO-H
1	0.329	0.361	0.395	0.422
2	0.012	0.334	0.012	0.501
3	0.218	0.212	0.272	0.278
4	0.358	0.353	0.383	0.376
5	0.369	0.380	0.568	0.565
6	0.420	0.407	0.411	0.410
7	0.337	0.352	0.455	0.494
8	0.327	0.350	0.392	0.420
9	0.000	0.015	0.000	0.024
10	0.648	0.664	0.727	0.701
11	0.358	0.432	0.333	0.439
12	0.385	0.372	0.440	0.410
13	0.120	0.113	0.085	0.087
14	0.418	0.387	0.388	0.379
15	0.248	0.272	0.236	0.288
16	0.146	0.156	0.162	0.189
17	0.619	0.648	0.582	0.655
18	0.105	0.131	0.049	0.085
19	0.856	0.839	0.797	0.725
20	0.024	0.042	0.027	0.086
21	0.200	0.201	0.132	0.135
22	0.796	0.802	0.868	0.869
23	0.376	0.455	0.588	0.642
24	0.374	0.387	0.489	0.514
25	0.191	0.239	0.347	0.382
26	0.085	0.093	0.130	0.178
27	0.098	0.102	0.116	0.157
28	0.542	0.543	0.559	0.556
29	0.142	0.153	0.397	0.393
30	0.223	0.267	0.237	0.337
Mean	0.311	0.335	0.353	0.390
SD	0.219	0.209	0.234	0.213
Min	0.000	0.015	0.000	0.024
Max	0.856	0.839	0.868	0.869

Table 25. Results of Slip Parameter Estimates for TIMSS B2 Data under the DINO and DINO-H Models

	U.S. Sample		Benchmark Sample	
Item	DINO	DINO-H	DINO	DINO-H
1	0.290	0.295	0.312	0.290
2	0.280	0.311	0.233	0.255
3	0.686	0.689	0.655	0.667
4	0.297	0.325	0.331	0.343
5	0.264	0.279	0.157	0.139
6	0.240	0.255	0.142	0.198
7	0.016	0.092	0.032	0.102
8	0.312	0.311	0.306	0.289
9	0.342	0.403	0.457	0.483
10	0.054	0.043	0.070	0.086
11	0.090	0.131	0.177	0.171
12	0.046	0.112	0.178	0.208
13	0.917	0.912	0.913	0.914
14	0.077	0.118	0.157	0.213
15	0.188	0.193	0.238	0.226
16	0.204	0.222	0.185	0.228
17	0.008	0.014	0.004	0.017
18	0.541	0.536	0.496	0.488
19	0.016	0.021	0.149	0.102
20	0.305	0.360	0.214	0.251
21	0.332	0.363	0.431	0.398
22	0.005	0.013	0.002	0.006
23	0.070	0.059	0.035	0.025
24	0.093	0.103	0.173	0.181
25	0.104	0.142	0.086	0.318
26	0.428	0.579	0.370	0.633
27	0.692	0.680	0.553	0.720
28	0.237	0.238	0.320	0.310
29	0.406	0.401	0.588	0.582
30	0.296	0.441	0.234	0.552
Mean	0.261	0.288	0.273	0.313
SD	0.223	0.223	0.212	0.226
Min	0.005	0.013	0.002	0.006
Max	0.917	0.912	0.913	0.914

Table 26. Differences of Model Fit Results between the DINA(-H) and DINO(-H) Models for TIMSS Data

		Booklet 1		Booklet 2	
Model Fit		AIC	BIC	AIC	BIC
U.S. Sample	DINO - DINA	104	104	102	102
	DINO-H - DINA-H	143	143	114	114
Benchmark Sample	DINO - DINA	36	36	68	68
	DINO-H - DINA-H	141	141	93	93

Table 27. Differences of Item Fit Index- δ between the DINA(-H) and DINO(-H) Models for TIMSS B1 Data

	U.S. Sample		Benchmark Sample	
Item	DINO-DINA	DINO-H - DINA-H	DINO-DINA	DINO-H - DINA-H
1	0.041	-0.031	0.383	0.061
2	-0.010	-0.017	0.030	0.003
3	0.038	-0.046	0.043	-0.055
4	0.014	0.020	0.127	0.008
5	-0.048	-0.005	-0.058	-0.005
6	-0.045	0.035	0.000	0.118
7	-0.038	-0.009	-0.372	-0.022
8	0.058	0.013	0.047	0.030
9	0.158	0.003	0.005	-0.072
10	0.063	-0.020	0.686	-0.102
11	-0.006	0.009	0.044	-0.007
12	-0.003	0.006	0.121	0.002
13	-0.055	0.027	-0.110	-0.022
14	-0.026	-0.008	0.017	0.019
15	-0.001	-0.042	-0.211	-0.069
16	-0.147	-0.009	-0.010	-0.097
17	-0.168	-0.120	-0.150	-0.147
18	-0.204	-0.086	-0.598	-0.482
19	-0.136	-0.041	-0.522	-0.171
20	0.059	0.029	0.069	0.028
21	-0.048	-0.103	0.377	-0.165
22	0.054	-0.050	0.164	-0.024
23	0.108	0.024	0.073	0.005
24	0.027	0.033	0.023	0.006
25	-0.022	0.004	-0.041	-0.001
26	0.001	-0.222	-0.002	-0.733
27	-0.108	-0.114	-0.836	-0.027
28	-0.301	-0.223	-0.756	-0.151
29	-0.010	0.007	0.074	0.011
Mean	-0.026	-0.032	-0.048	-0.071
SD	0.096	0.068	0.322	0.167
Min	-0.301	-0.223	-0.836	-0.733
Max	0.158	0.035	0.686	0.118

Table 28. Differences of Item Fit Index-IDI between the DINA(-H) and DINO(-H) Models for TIMSS B1 Data

	U.S. Sample		Benchmark Sample	
Item	DINO-DINA	DINO-H - DINA-H	DINO-DINA	DINO-H - DINA-H
1	0.087	0.024	0.434	0.083
2			5.288	-2.148
3	0.051	-0.035	0.084	-0.042
4	0.020	0.024	0.134	0.012
5	-0.053	-0.012	-0.058	-0.009
6	9.748	7.524	7.170	3.135
7	-0.073	-0.020	-0.424	-0.025
8	1.123	0.393	0.430	0.507
9	0.172	0.012	0.126	-0.055
10	0.107	0.024	0.825	-0.062
11	-0.023	-0.003	0.100	-0.025
12	-0.017	-0.018	0.155	-0.014
13	-0.116	-0.005	-0.126	-0.032
14	-0.015	0.002	0.079	0.040
15	-0.019	-0.063	-0.165	-0.078
16	-0.247	-0.056	-0.010	-0.092
17	-0.037	-0.030	0.275	-0.058
18	-0.245	-0.102	-0.646	-0.552
19	-0.176	-0.072	-0.564	-0.202
20	-0.022	-0.017	0.057	0.014
21	-0.017	-0.076	0.443	-0.118
22	0.042	-0.006	0.158	-0.023
23	-0.071	-0.030	-0.141	-0.022
24	-0.027	-0.019	0.003	-0.024
25	-0.019	0.003	-0.073	-0.001
26	0.000	-0.198	0.000	-0.524
27	0.018	-0.012	-0.280	-0.077
28	0.005	-0.003	0.086	-0.144
29	-0.021	-0.012	0.142	0.014
Mean	0.363	0.258	0.466	-0.018
SD	1.854	1.427	1.643	0.741
Min	-0.247	-0.198	-0.646	-2.148
Max	9.748	7.524	7.170	3.135

Note. Item 2 for the U.S. sample was removed because its extreme IDI values for the DINO(-H) model.

Table 29. Differences of Item Fit Index- δbetween the DINA(-H) and DINO(-H) Models for TIMSS B2 Data

	U.S. Sample		Benchmark Sample	
Item	DINO-DINA	DINO-H - DINA-H	DINO-DINA	DINO-H - DINA-H
1	-0.034	-0.050	-0.042	-0.032
2	0.025	-0.288	0.263	-0.131
3	-0.025	0.000	-0.002	0.006
4	0.003	0.016	-0.026	-0.015
5	0.001	0.012	-0.055	0.006
6	-0.024	-0.026	0.018	-0.033
7	-0.032	-0.040	0.004	-0.039
8	0.053	0.002	-0.005	-0.045
9	-0.157	0.104	-0.201	-0.005
10	-0.093	-0.010	-0.054	0.019
11	0.080	-0.051	-0.004	-0.155
12	0.006	-0.011	-0.048	-0.027
13	-0.005	0.005	0.028	0.018
14	-0.041	-0.039	-0.036	-0.036
15	-0.012	-0.022	0.009	-0.007
16	-0.159	-0.039	-0.038	-0.011
17	-0.067	-0.047	-0.090	-0.084
18	-0.085	-0.037	-0.118	-0.022
19	-0.140	-0.120	-0.053	-0.110
20	-0.031	-0.071	-0.054	-0.082
21	0.057	0.075	0.041	0.107
22	0.033	0.016	0.008	0.008
23	0.068	-0.011	0.059	-0.053
24	-0.073	-0.037	0.003	-0.007
25	-0.030	-0.021	0.025	0.007
26	0.007	-0.135	0.024	-0.028
27	-0.020	0.007	0.032	-0.036
28	-0.009	0.025	-0.028	0.005
29	-0.019	-0.006	-0.003	0.007
30	-0.138	-0.057	0.112	-0.047
Mean	-0.029	-0.029	-0.008	-0.027
SD	0.063	0.068	0.077	0.050
Min	-0.159	-0.288	-0.201	-0.155
Max	0.080	0.104	0.263	0.107

Table 30. Differences of Item Fit Index-IDI between the DINA(-H) and DINO(-H) Models for TIMSS B2 Data

	U.S. Sample		Benchmark Sample	
Item	DINO-DINA	DINO-H - DINA-H	DINO-DINA	DINO-H - DINA-H
1	-0.023	-0.045	-0.027	-0.029
2	0.297	-0.139	0.486	-0.092
3	-0.082	-0.008	-0.008	0.015
4	-0.031	0.003	-0.047	-0.033
5	-0.059	-0.014	-0.075	-0.004
6	-0.026	-0.024	0.031	-0.024
7	-0.086	-0.059	-0.020	-0.051
8	0.032	-0.017	-0.024	-0.072
9	0.177	0.251	0.245	0.260
10	-0.100	-0.017	-0.087	0.007
11	0.102	-0.016	0.077	-0.097
12	-0.044	-0.024	-0.057	-0.046
13	-0.053	0.073	0.412	0.238
14	-0.076	-0.046	-0.047	-0.058
15	-0.015	-0.035	0.051	-0.007
16	-0.136	-0.031	-0.040	0.009
17	-0.072	-0.048	-0.093	-0.086
18	-0.003	-0.013	0.004	0.015
19	-0.149	-0.127	-0.063	-0.118
20	0.022	-0.002	0.028	-0.003
21	-0.001	0.053	-0.042	0.062
22	0.029	0.015	0.007	0.008
23	0.093	-0.018	0.065	-0.065
24	-0.070	-0.030	0.000	-0.008
25	-0.039	-0.004	-0.054	0.035
26	-0.143	-0.197	-0.193	-0.028
27	-0.040	-0.003	-0.185	-0.134
28	-0.030	0.020	-0.043	0.003
29	-0.026	-0.027	-0.008	0.016
30	-0.212	-0.039	-0.032	-0.076
Mean	-0.037	-0.015	0.009	-0.012
SD	0.079	0.070	0.143	0.085
Min	-0.212	-0.197	-0.193	-0.134
Max	0.177	0.251	0.486	0.260

Figure 1: Linear Hierarchy

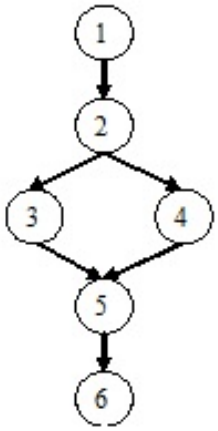

Figure 2: Convergent Hierarchy

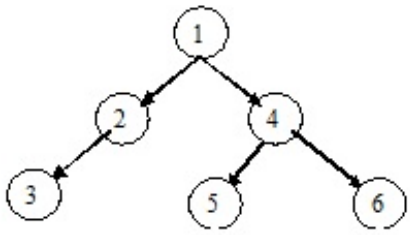

Figure 3: Divergent Hierarchy

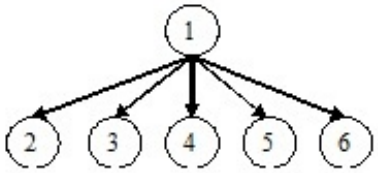

Figure 4: Unstructured Hierarchy

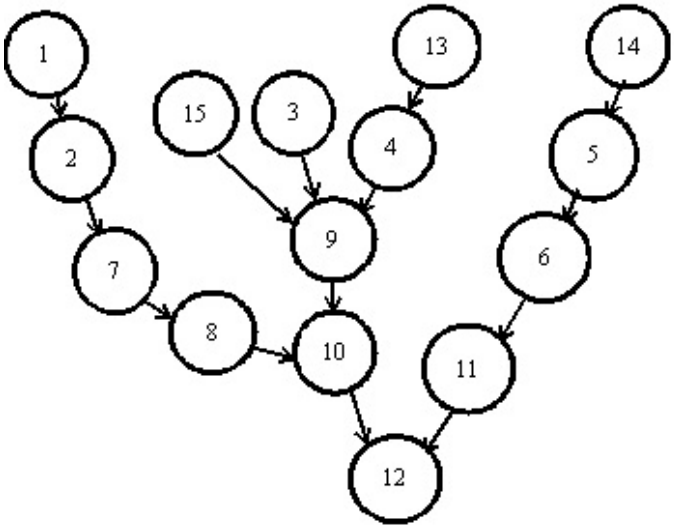

Figure 5: Hierarchical relationship among the attributes for the eighth grade TIMSS 2003 mathematics test

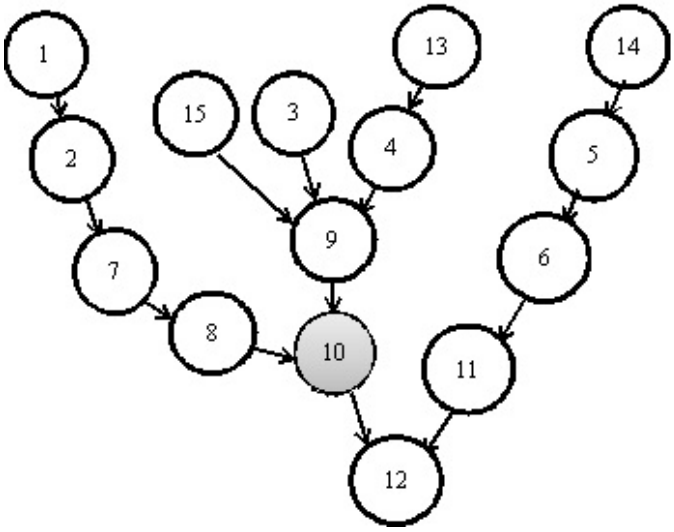

Figure 6: Hierarchical relationship among the attributes for booklet 1

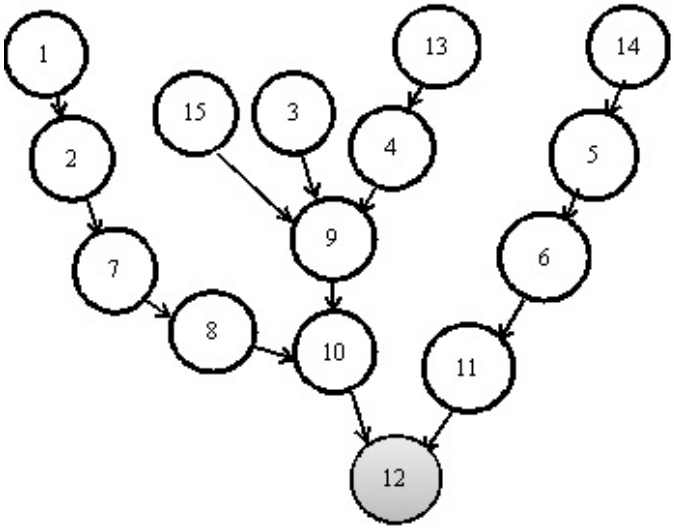

Figure 7: Hierarchical relationships among the attributes for booklet 2