2013

Cognitive diagnostic analysis using hierarchically structured skills

Yu-Lan Su
University of Iowa

COGNITIVE DIAGNOSTIC ANALYSIS USING HIERARCHICALLY STRUCTURED SKILLS

by

Yu-Lan Su

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy degree
in Psychological and Quantitative Foundations
(Educational Psychology)
in the Graduate College of
The University of Iowa

May 2013

Thesis Supervisors: Associate Professor Won-Chan Lee
Assistant Professor Kyong Mi Choi

ABSTRACT

This dissertation proposes two modified cognitive diagnostic models (CDMs), the deterministic, inputs, noisy, “and” gate with hierarchy (DINA-H) model and the deterministic, inputs, noisy, “or” gate with hierarchy (DINO-H) model. Both models incorporate the hierarchical structures of the cognitive skills in the model estimation process, and can be used for situations where the attributes are ordered hierarchically. The Trends in International Mathematics and Science Study (TIMSS) 2003 data are analyzed to illustrate the proposed approaches. The simulation study evaluates the effectiveness of the proposed approaches under various conditions (e.g., various numbers of attributes, test lengths, sample sizes, and hierarchical structures). The simulation study attempts to address the model fits, items fit, and accuracy of item parameter recovery when the skills are in a specified hierarchy and varying estimation models are applied. The simulation analysis examines and compares the impacts of the misspecification of a skill hierarchy on various estimation models under their varying assumptions of dependent or independent attributes. The study is unique in incorporating a skill hierarchy with the conventional DINA and DINO models. It also reduces the number of possible latent classes and decreases the sample size requirements. 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. Its results demonstrate the proposed approaches to analyzing the hierarchically structured CDMs, illustrate the usage in applying cognitive diagnosis models to a large-scale assessment, and provide researchers and test users with practical guidelines.

Abstract Approved:

____

Thesis Supervisor

____

Title and Department

____

Date

____

Thesis Supervisor

____

Title and Department

____

Date

COGNITIVE DIAGNOSTIC ANALYSIS USING HIERARCHICALLY STRUCTURED SKILLS

by

Yu-Lan Su

A thesis submitted in partial fulfillment
of the requirements for the
Doctor of Philosophy degree
in Psychological and Quantitative Foundations
(Educational Psychology)
in the Graduate College of
The University of Iowa

May 2013

Thesis Supervisors: Associate Professor Won-Chan Lee
Assistant Professor Kyong Mi Choi

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

____

PH.D. THESIS

____

This is to certify that the Ph.D. thesis of

Yu-Lan Su

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Psychological and Quantitative Foundations
(Educational Psychology) at the May 2013 graduation.

Thesis Committee: ___________________________________
Won-Chan Lee, Thesis Supervisor

____
Kyong Mi Choi, Thesis Supervisor

____
Michael J. Kolen

____
Catherine J. Welch

____
Joyce L. Moore

To Ilene, Albert, and Jui-Sheng

ACKNOWLEDGMENTS

Words cannot express how I am deeply grateful to my dissertation committee chair, Dr. Won-Chan Lee, and co-chair, Dr. Kyong Mi Choi, for their guidance and encouragement. Advising a student on the doctoral dissertation is a tremendous time-consuming work. Both Dr. Lee and Dr. Choi showed me a great role-model of knowledgeable, supportive, caring, and patient advisor. I would not have been able to achieve the end of the Ph.D. process without them.

I would also like to take this opportunity to express my sincere appreciation to my committee members, Dr. Michael Kolen, Dr. Catherine Welch, and Dr. Joyce Moore. I thank them for providing me valuable feedback and encouragement. I sincerely thank my academic adviser, Dr. Joyce Moore, for leading and mentoring me throughout the program. I also want to thank Dr. Johnmarshall Reeve who was originally my advisor, and was always so supportive during my study at Iowa. He guided me to the motivation world and helped me gain deep insights into autonomy support and self-determination theory.

My life in the United States would not have been possible without the support from Measurement Research Department at ACT. I thank my supervisors, Dr. Deborah Harris and Dr. Xiaohong Gao, for their valuable advises, suggestions, great love, and guidance all the time. I would like to thank my mentors, Dr. Rongchun Zhu, Dr. Han-Wei Chen, Dr. Chunyan Liu, and Dr. Yuki Nozawa, for teaching, training, and developing my professional skills. Special thanks to the CDM study group, including Dr. Troy Chen, Dr. Yu Fang, Dr. Chunyan Liu, and Dr. Alan Huebner, who guided me into a new measurement world of CDM, helped me generate the ideas of this dissertation, kept me on track, and provided insightful comments. I also thank the department for sponsoring my assistantship for more than five years.

So many people in MRD helped me write this dissertation, develop a career, and raise babies in U.S. that it is hard to know where to start my acknowledgments. Many thanks to my dearest mentors and colleges at ACT, including Dr. Troy Chen, Dr. Chi-Yu Huang, Dr. Wei Tao, Dr. J.P. Kim, Dr. Tianli Li, Dr. Zhongmin Cui, Dr. Dongmei Li, Dr. Yi He, Dr. Yang Lu, Dr. Yi Cao, Juan Chen, and Xuan Wang. They were always there for me whenever I need help.

My families deserve the special recognition. I am indebted to my families' unconditional love and support. My heartfelt thanks also go to my dearest babies, Ilene and Albert, for making my life more meaningful and fun.

ABSTRACT

This dissertation proposes two modified cognitive diagnostic models (CDMs), the deterministic, inputs, noisy, “and” gate with hierarchy (DINA-H) model and the deterministic, inputs, noisy, “or” gate with hierarchy (DINO-H) model. Both models incorporate the hierarchical structures of the cognitive skills in the model estimation process, and can be used for situations where the attributes are ordered hierarchically. The Trends in International Mathematics and Science Study (TIMSS) 2003 data are analyzed to illustrate the proposed approaches. The simulation study evaluates the effectiveness of the proposed approaches under various conditions (e.g., various numbers of attributes, test lengths, sample sizes, and hierarchical structures). The simulation study attempts to address the model fits, items fit, and accuracy of item parameter recovery when the skills are in a specified hierarchy and varying estimation models are applied. The simulation analysis examines and compares the impacts of the misspecification of a skill hierarchy on various estimation models under their varying assumptions of dependent or independent attributes. The study is unique in incorporating a skill hierarchy with the conventional DINA and DINO models. It also reduces the number of possible latent classes and decreases the sample size requirements. 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. Its results demonstrate the proposed approaches to analyzing the hierarchically structured CDMs, illustrate the usage in applying cognitive diagnosis models to a large-scale assessment, and provide researchers and test users with practical guidelines.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AHM	Attribute hierarchy method
AIC	Akaike Information Criterion
AMSE	average mean-square error
ASB	average squared bias
AVAR	average variance
BIC	Bayesian Information Criterion
CDM	Cognitive diagnostic model
DINO	the deterministic, inputs, noisy, “and” gate (model)
DINO-H	the deterministic, inputs, noisy, “and” gate with hierarchy (model)
DINO	the deterministic, inputs, noisy, “or” gate (model)
DINO-H	the deterministic, inputs, noisy, “or” gate with hierarchy (model)
HO-DINO	the higher order deterministic, inputs, noisy, “and” gate (model)
IRT	Item response theory
MAIC	mean of Akaike Information Criterion
MBIC	mean of Bayesian Information Criterion
MDelta	mean of δ index
MIDI	mean of Item Discrimination Index
RUM	the reparameterized unified model

CHAPTER I

INTRODUCTION

Cognitive Diagnostic Models (CDMs) are psychometric models developed to identify the examinees' ability to master fine-grained skills based on a pre-specified matrix. Cognitive diagnostic tests can be used to identify the skill combinations that the examinee is likely to either possess or not possess. The results provide specific informational feedback to the examinees so they can make inferences about their mastery of different cognitive skills. CDMs have been increasingly valued in the literature of measurement in recent years, although the popularity of their applied research remains lower than the more common Item Response Theory (IRT) models.

Unlike the IRT models that focus on item-level analysis and assign scores to examinees, CDMs focus more on providing examinees and teachers with informative and diagnostic feedback on the specific skills examinees need to improve. More specifically, they assign examinees attribute profiles indicating the skills that they have or have not mastered. In CDMs, examinees' attribute profiles (sometimes called the estimated skill patterns) are categorical latent variables (discrete variables). Thus, in the scoring process, CDMs classify examinee responses into various discrete latent classes (i.e., a set of combinations of 0s and 1s, where 0 means non-mastery and 1 means mastery of an individual skill). Almost all CDMs require specifying a $J \times K$ Q-matrix, in which K stands for the number of skills being measured by the test and J stands for the number of items in a test (Tatsuoka, 1995).

Applying CDMs has both pros and cons. CDMs are well-known for their advantages in providing informational feedback about examinees' ability to master multiple fine-grained skills. One common limitation in current CDMs is that they use 0/1 discrete variables to represent true person profiles, rather than a continuous variable. However, while scoring, it is possible to report the probability or percentage of mastering each attribute (e.g., report the probability estimates of mastery from EAP scoring results for each attribute for an examinee) to make up for the limitation at some point. A continuous score scale could be used for reporting purposes, instead of just 0/1. de la Torre and Karelitz (2009) examined the nature of the underlying latent trait (continuous or discrete) and varying estimation models, and found that low diagnostic items (i.e., high misclassification rate of attribute) were more influential on estimation inaccuracy than the noise introduced by fitting the wrong model and the nature of the latent trait.

Motivation for the Study

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 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.

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, Leighton, & Hunka, 2007; Leighton, Gierl, & Hunka, 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, Templin, & Henson, 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 simultaneously. 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, 2008b) 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 both the real data and simulated data.

The purpose for conducting the real data analysis is to illustrate the proposed approach of applying the DINA-H and DINO-H models, 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. The released Trends in International Mathematics and Science Study 2003 (i.e., TIMSS) math test was used in the study. The cognitive dimension was categorically defined as knowing facts and procedures, using concepts, solving routine problems, and reasoning. However, the four categories of cognitive domains defined in

TIMSS items are not deemed as hierarchically ordered. This is because the first three levels (i.e., knowing facts and procedures, using concepts, and solving routine problems) of cognitive domains are not considered prerequisite to their next level in mathematics learning. Hence, the study adopted the Common Core State Standards (CCSS) as the attributes to develop the Q-matrix and its hierarchy.

The purpose of conducting the simulation analysis is to evaluate and provide more detailed and supportive evidence about the effectiveness of the proposed models under various conditions (e.g., different number of attributes, different test lengths, sample sizes, and different structures of hierarchies). This would also allow for testing the reduction of the required sample sizes, model fits, the issues of convergence, and item parameter recoveries. Moreover, the simulation analysis is intended to explore the effect when the specified skill hierarchy in the data is inconsistent with the estimation model.

Significance of the Study

The study is unique in its distinctive feature of incorporating hierarchically structured skills into the conventional DINA/DINO models. 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). There is a need for a model whose model specification, the relationships among attributes, possible attribute profiles, and Q-matrix are consistent with the theoretical background and test blueprint. The current DINA and DINO models assume independent skills up to 2^K of attribute profiles, without considering the situations when skills are hierarchically related in a certain structure. 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-matrix. The conventional DINA/DINO models only work when the skills are independent. If the skills are hierarchically related and the conventional models are applied, the parameter estimates could be biased and less accurate. New models are developed to apply when the skills are hierarchically ordered.

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 simulation analysis would provide valuable information about the potential inaccuracy of parameter estimates due to misspecification of the relationships between attribute and possible attribute profiles. The simulation study examines the model fit and item parameter recovery when the data simulation models are different from the estimation models. Specifically, both the conventional DINA/DINO and new hierarchical models are applied when the attributes are independent or dependent in a specified hierarchical structure. The simulation study contributes to exploring the effect when model specification is consistent or inconsistent to the assumptions and characteristics of skills.

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.

The last contribution of the current study is to examine the performance of the proposed DINA-H and DINO-H models, and to provide information about model fit and item parameter recovery under varying conditions of different numbers of attributes, different test lengths, and sample sizes. The results of the study are expected to demonstrate the benefits, efficiencies, and feasibility of the proposed DINA-H and DINO-H models, and to facilitate the reduction of possible attribute profiles and the large sample size requirements in analyzing a CDM. The study also contributes to the analysis of tests that assess more attributes, and promote computational efficiency.

Research Objectives and Questions

This section includes two parts. The first part describes the objectives of the study, and the second part presents the research questions.

Objectives of the Inquiry

The study proposes to develop two modified CDMs, the DINA-H and the DINO-H models, which involve the hierarchical structure of cognitive skills. The purpose of these two models is to construct the hierarchies of cognitive skills in the model estimation process. This can facilitate reporting the mastery/non-mastery of skills with different levels of cognitive loadings. The intention of the study is to apply the proposed skill hierarchy approach in conjunction with the DINA and DINO models to analyze both the real data and the simulated data.

Adopting the CCSS as attributes to code the Q-matrix and skill hierarchies, the analysis of the TIMSS 2003 data demonstrates the proposed models and the feasibility of retrofitting. The purpose for conducting the real data analysis is to illustrate the proposed models, 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 to teachers and students.

The purpose of conducting the simulation analysis is to evaluate the effectiveness of the proposed models under various conditions (e.g., different number of attributes, different test lengths, and sample sizes). The objective includes examining item parameter recovery by analyzing and comparing various CDMs under varying assumptions of dependent or independent attributes. More specifically, the simulation study evaluates the item parameter recoveries and model fits when the specified skill hierarchy in the data is consistent or inconsistent with the estimation model, given a type of attribute profile patterns with dependent or independent skills, by using the conventional DINA and DINO models as well as the proposed DINA-H and DINO-H models.

Research Questions

This study is designed to address the following research questions:

1. How do the proposed DINA-H and DINO-H perform?
   - 1.1 Do the DINA-H and DINO-H models provide reasonable item parameter estimates?
   - 1.2 Do the DINA-H and DINO-H models provide stable calibration results with small sample size?
   - 1.3 Which CDM (the DINA-H or DINO-H model) performs better while applying a skill hierarchy?
2. When skills are ordered hierarchically, how do the conventional DINA and DINO models and the proposed DINA-H and DINO-H models compare?
   - 2.1 How do the conventional and new models compare in terms of parameter estimates?

2.2 How does the performance of the conventional and new models compare under varying conditions of different numbers of attributes, different test lengths, and different sample sizes?

3. What is the impact of misspecification of a skill hierarchy on the DINA(-H) and the DINO(-H) models?

3.1 Do the item parameters recover well when a specified skill hierarchy is inconsistent with an estimation model?

3.2 How do the models perform and compare under varying conditions with different numbers of attributes, test lengths, and sample sizes?

Research questions 1.1, 1.2, 1.3, and 2.1 are addressed by the real data analysis; research questions 2.1, 2.2, 3.1, and 3.2 are addressed by the simulation analysis. This study analyzes both the real and simulated data. The fit statistics and other evaluation criteria are discussed in Chapter III. Throughout the study, the CDMs are assumed to hold. Details about CDMs and their assumptions are provided in Chapter II, while the procedures for the analyses are described in Chapter III.

Overview of the Dissertation

The dissertation is composed of five chapters, including the current chapter. Following Chapter I, Chapter II contains an introduction of CDMs, learning sequence, and attribute hierarchy. Chapter III presents a description of the methodology used in this investigation. Chapter IV provides the results of the study. Chapter V summarizes the findings and considers the implications of the results for psychometricians and educators.

CHAPTER II

LITERATURE REVIEW

Chapter II consists of two main sections. The first section introduces general information about the cognitive diagnostic models, followed by more specific descriptions of several popular models. The second section discusses learning sequences and then introduces cognitive attribute hierarchies.

Cognitive Diagnostic Models

CDMs are developed to identify examinees' mastery or non-mastery of multiple fine-grained attributes based on a pre-specified matrix of attributes. Unlike Item Response Theory (IRT), which focuses on item-level analysis and assigning scores to examinees, CDMs emphasize providing examinees and teachers with informative and diagnostic feedback that allows examinees to know which specific skills they should improve.

Overview of Cognitive Diagnostic Models

CDMs have several synonyms that appear in the literature. These synonyms include cognitive diagnosis models (or cognitively diagnostic models) (de la Torre, 2009; Henson & Douglas, 2005; Huebner, 2010; Tatsuoka, 1995), diagnostic classification models (DCM) (Rupp & Templin, 2008a; Rupp et al., 2010), cognitive psychometric models (Rupp & Mislevy, 2007), multiple classification (latent class) models (Maris, 1999), latent response models (Maris, 1995), restricted latent class models (Haertel, 1989, 1990), structured IRT models (Rupp & Mislevy, 2007), and structured located latent class models (Xu & von Davier, 2006, 2008). These different terms highlight specific aspects of the model, such as the theoretical background, the statistical model, or the examinee respondent scoring. The most common term, CDM, is used consistently throughout the study.

In CDMs, examinee attribute profiles contain categorical latent variables (discrete variables). These variables are the center of the scoring process, in which CDMs classify examinee responses into discrete latent classes (i.e., a set of combinations of 0 and 1 values where 0 means non-mastery and 1 means mastery of an individual skill). Tests that implement CDMs consist of dichotomous items, polytomous items, or mixed formats. Recently, CDMs have been implemented in assessing both academic performance (i.e., achievement tests) and psychological properties (i.e., psychological disorder assessments). The former more often relies upon the term “skills” to describe the kind of cognitive skills an assessment is intended to measure, whereas the latter tends to use the term “attributes” to address the features of a certain symptom. These two terms (skills and attributes) are used interchangeably throughout this study, and are not meant to indicate different, discrete categories of analysis or not intended to be discriminatory. In other words, attributes are equivalent to basic knowledge, skills, or cognitive processes.

Assumptions

Various CDMs hold different assumptions about how all the skills measured by an individual assessment interact with each other. CDMs also consider how examinees’ skills influence their test performance. Some models, such as those implemented in assessing mathematics or reading comprehension skills, assume that an examinee must possess all the required skills for an item to answer that item correctly. For example, many studies adapted the DINA model to estimate the fraction and subtraction data collected by K. Tatsuoka (1990) and more recently by C. Tatsuoka (2002) (de la Torre, 2009; de la Torre & Douglas, 2004). Other models, like those often applied in medical and psychological diagnoses, assume that the absence of certain attributes (i.e., symptoms) can be compensated by the presence of other characteristics. For example, the DINO model was first advocated in analyzing gambling addictions (Templin & Henson, 2006).

Despite the assumptions individual CDMs might hold, classifying these models based on their assumptions occurs inconsistently within the literature. According to some introductory articles (DiBello, Roussos, & Stout, 2007; Roussos, Templin, & Henson, 2007; Huebner, 2010), CDMs fall into two categories: non-compensatory and compensatory. The term “non-compensatory” means that an attribute does not compensate for the deficiency of another attribute in order to correctly respond to an item. Under this scheme, the term “conjunctive” is used interchangeably with non-compensatory. The non-compensatory models have the statistical representation of a product term over attributes in their model equations (Henson, Templin, & Willse, 2009). Hence, a successful task requires mastery of all necessary skills for successful performance, and a lack of competency in any one of the required skills causes an unsuccessful performance on the task. The non-compensatory models include the DINA model (Haertel, 1989; Junker & Sijtsma, 2001), the non-compensatory reparameterized unified model (NC-RUM or RUM), and the reduced reparameterized unified model (rRUM) (DiBello et al., 2007; Hartz, 2002), the noisy inputs, deterministic, “and” gate (NIDA) model (Junker & Sijtsma, 2001), the HYBRID model (Gitomer & Yamamoto, 1991), the unified model (UM; DiBello, Stout, & Roussos, 1995), and the conjunctive multiple classification latent class model (the conjunctive MCLCM; Maris, 1999).

The term “compensatory” means that in order to perform a task successfully, a high enough level of competence in one attribute can compensate for a deficiency or low level of competence in another attribute, through the interaction of skills required by that task. Under this scheme, the term “disjunctive” is used interchangeably with compensatory. The probability of a positive response for an item is high when test takers master (or possess) at least one of the required attributes. Some of the compensatory models include the DINO model (i.e., the counterpart of the DINA model) (Templin & Henson, 2006), the compensatory reparameterized unified model (C-RUM), the noisy input deterministic ”or” (NIDO) model (i.e., the counterpart of the NIDA model) (Templin, Henson, & Douglas, 2007), the General Diagnostic Model (GDM; von Davier, 2005), and the compensatory multiple classification latent class model (the compensatory MCLCM; Maris, 1999).

Certain literature reviews (Henson et al., 2009; Rupp et al., 2010) classify CDMs in a different manner. According to these reviews, different CDMs fall into the non-compensatory and/or compensatory categories. However, each of these categories contains both conjunctive and disjunctive models. Under this scheme, conjunctive is not an alternative term for non-compensatory, and disjunctive is not an alternative term for compensatory. Instead, the non-compensatory models are characterized by “the conditional relationship between any attribute and the item response depending on mastery or non-mastery of the remaining attributes” (Henson et al., 2009, p. 192). Compensatory models are characterized by the “conditional relationship between any attribute and the item response not depending on mastery or non-mastery of the remaining required attributes" (Henson et al., 2009, p.192). Furthermore, the term "conjunctive" refers to the presence of all attributes leading to a positive response, from a deterministic perspective (Rupp & Templin, 2008a). That means a missing skill cannot be made up by the mastery of other skills. An examinee needs to possess all the required skills for an item in order to answer that item correctly. Disjunctive models define the probability of a correct response such that mastering a subset of the attributes is sufficient to have a high probability of a correct response. According to this system, certain CDMs are classified in different categories as compared to the previous paragraph. For example, the DINA model (e.g., de la Torre & Douglas, 2004; Junker & Sijtsma, 2001) is viewed under both systems as a non-compensatory model that uses a conjunctive condensation function (Maris, 1999). However, the DINO model is classified as a non-compensatory model with a disjunctive condensation function, instead of a compensatory disjunctive model.

CDMs also hold assumptions about the conditional independence of attribute profiles and independence among examinees. Conditional independence means that item responses are independent conditionally on the latent class of the examinee, which is similar to local independence in IRT (Rupp et al., 2010). For those latent classes without any relationship or constraints imposed, the maximum number of possible latent classes is 2^K where K is the number of the attributes measured by an assessment. There are 2^K - 1 parameters, which need to be estimated by implementing most CDMs with dichotomous latent attribute variables. Therefore, the number of parameters will be greater if more attributes are measured.

To improve the efficiency of implementing CDMs, a variety of methods have been developed to reduce the number of parameters that need to be estimated. One method that reduces the complexity of the model space indirectly via parameters is a log-linear model (Henson et al., 2009). This method estimates the main effects and the interaction effects for each possible combination of attributes. In other methods, the relationships among attributes are assumed to be unstructured tetrachoric, structured tetrachoric, or hierarchical ordered in different CDMs. The unstructured tetrachoric models represent the tetrachoric correlations of all attribute pairs directly, without placing any constraints on the patterns of the tetrachoric correlations (Hartz, 2002). The difficulty in estimating multiple tetrachoric correlations and threshold parameters occurs when the structured tetrachoric models assume a priori about the tetrachoric correlations among the attributes. Two examples are the higher order DINA model in de la Torre and Douglas (2004), and the generalized normal-ogive model in Templin (2004). Another method has been proposed that applies hierarchies among cognitive skills in test developing process, and hence, reduces the number of possible attribute profiles (e.g., the attribute hierarchy method in Gierl et al. (2007)). To conduct CDM estimations, several approaches have been applied to model the relationship among attributes, and different attitudes toward the assumptions lead to the selection of a specific CDM.

Estimation

Estimating parameters using CDMs includes estimating the item parameters (or attribute parameters) and attribute profiles (i.e., respondent profiles or respondent parameters). Item parameters and attribute profiles may be estimated simultaneously by joint maximum likelihood estimation or marginalized maximum likelihood estimation using the expectation-maximization (EM) algorithms, such as the DINA model estimation described in de la Torre (2009), and the estimation of the generalized diagnostic model in von Davier (2005). Alternatively, item parameters and attribute profiles may be estimated simultaneously using a Markov Chain Monte Carlo (MCMC) approach, such as in the higher order DINA model estimation by de la Torre and Douglas (2004). Both approaches have their own advantages and disadvantages (Rupp et al., 2010). Initializing the EM algorithm requires setting up the starting values of the item parameters that must be estimated. However, choosing different starting values could result in a slight change in item parameter estimates. Moreover, computing the EM algorithm becomes more intensive as the number of latent classes increases. Sometimes imposing more constraints on the relationships among attributes makes the computation more complex.

Conversely, conducting MCMC estimation does not require pre-calculation of the maximum likelihood estimators of the item parameters. MCMC estimation occasionally results in convergent issues, and the results are not reproducible because of its random simulations, whereas EM results can be replicated. Employing a CDM to analyze data, another common approach can be applied in which the marginal maximum likelihood (MML) algorithm is used by first assuming a prior population distribution of latent classes, estimating item parameters under this assumption, and then estimating the individual respondent parameters (Templin, 2004).

To score and classify examinee respondents, there are three common approaches in assigning an examinee into a latent class: the maximum likelihood estimation (MLE), the maximum a posteriori (MAP) estimate of the posterior distribution, and an expected a posteriori (EAP) estimate for each attribute (Huebner & Wang, 2011). For MLE classification, the likelihood is computed at each attribute profile, and the examinee is assigned an estimated attribute profile that maximizes the likelihood. Sometimes when the distribution of attribute profiles is expected, the prior probabilities are obtained. At this point, MAP classification can be applied by computing the posterior probability using Bayes' theorem. EAP provides probability estimates for each attribute, whereas MLE and MAP do not (Huebner & Wang, 2011). However, computing MLE and MAP are more statistically straightforward. MLE and MAP find the mode of the posterior distribution, whereas EAP finds the average (Embretson & Reise, 2000). EAP calculates the probabilities of mastery for each attribute for an examinee and sets up a cutoff probability value (usually at 0.5) to classify the attribute into mastery or non-mastery (Huebner & Wang, 2011). Hence, this cutoff can be altered based on different research purposes.

To evaluate the model fit for the CDM estimation, the common fit statistics used are the Akaike Information Criterion (AIC; Akaike, 1973, 1974), the Bayesian Information Criterion (BIC; Schwarz, 1978), and the Bayes factor (Kass, 1993; Kass & Raftery, 1995). The BIC and Bayes factor were considered the same meaningfully, and BIC was easier in its computation (de la Torre & Douglas, 2008). For a given dataset, the larger the log-likelihood, the better the model fit; the smaller the AIC, BIC, and Bayes factor values, the better the model fit (Xu & von Davier, 2008).

Q-Matrix

Specifying a $J \times K$ Q-matrix is required for implementing almost all CDMs (Tatsuoka, 1995). Using *K* for the number of skills being measured, and *J* for the number of items in a test, the Q-matrix can be illustrated as follows:

$$Q = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & \cdots & 1 \end{pmatrix}$$ where *J* is the number of rows and *K* is the number of columns,

and $$q_{jk} = \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ item requires the } k^{\text{th}} \text{ skill} \\ 0 & \text{otherwise} \end{cases}.$$

There are 2^K -1 possible rows for a Q-matrix because there is no item measuring no attributes (i.e., the case of all zeros doesn't exist).

For item *j*, attribute *k*, and examinee *i*, $\boldsymbol{\alpha}_i = \{ \alpha_{ik} \}$ represents the examinee's binary skills vector, *k*=1, . . . , *K*, where 1 on the *k*th element denotes mastery of skill *k*, and 0 denotes non-mastery.

$$\alpha_{ik} = \begin{cases} 1 & \text{if the examinee has mastered the } k^{\text{th}} \text{ skill} \\ 0 & \text{otherwise} \end{cases}$$

when there are 2^K possible examinee attribute profiles. For example, with $K = 2$ skills, there are $2^K = 4$ possible skill patterns of mastery and non-mastery: {0,0}, {1,0}, {0,1}, and {1,1}.

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 et al., 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, 2008a).

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

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

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 verifying 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.

An extension of the DINA model is the higher-order DINA (HO-DINA) model (de la Torre & Douglas, 2004). The HO-DINA model is a higher order unidimensional latent trait model, and has the same basic specifications as the DINA. The main difference between these two models is that the HO-DINA model assumes that the independency of the mastery of one attribute to other attributes is conditional on a higher order latent ability variable θ. The basis of the higher order latent trait approach is to parsimoniously model the joint distribution of the attributes. The model is very similar to the 2 PL IRT model:

$$P(\alpha_{ik} \mid \theta_i, \lambda) = \frac{\exp(1.7\lambda_{1k}(\theta_i - \lambda_{ok}))}{1 + \exp(1.7\lambda_{1k}(\theta_i - \lambda_{ok}))} \tag{6}$$

where λ_{1k} refers to the slop parameter for attribute *k*, and λ_{ok} refers to the difficulty parameter.

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:

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

$$g_j = P(X_{ij} = 1 | \omega_{ij} = 0) \ , \tag{8}$$

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

$$P_j(\omega_{ij}) = P(X_{ij} = 1 | \omega_{ij}) = (1 - s_j)^{\omega_{ij}} g_j^{(1-\omega_{ij})} \tag{9}$$

where $$\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}} . \tag{10}$$

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.

rRUM

The reduced Reparameterized Unified Model (rRUM; DiBello et al., 2007) is a non-compensatory conjunctive CDM. It assumes that an examinee must know all the required attributes in order to answer an item correctly (Henson et al., 2009). The rRUM is a more complicated model, as it allows for different probabilities of item response depending on what required attributes an examinee has mastered. It discriminates between different examinee response classes, and allows items with the same Q-matrix coding to have different response probabilities (Hartz, 2002; Templin, Henson, Templin, & Roussos, 2008). The rRUM also addresses the issue in the DINA model that all examinees who lack at least one required attribute have the same probability of answering an item correctly (Henson et al., 2009).

The rRUM model includes two different parameters, π_j^* and r_{jk}^* : π_j^* refers to the probability of correctly applying all the required skills for the jth item, and r_{jk}^* refers to the penalty for not mastering the kth skill to item j. For every single non-mastered attribute, the probability of answering an item correctly is reduced by the penalty, r_{jk}^* , that discriminates non-masters to masters. The smaller r_{jk}^{*} is, the higher the level of discrimination. The mathematical presentations for the two parameters are:

$$\pi_j^{*} = \prod_{k=1}^{K} \pi_{jk}^{\eta_{jk}} \ , \tag{11}$$

$$\text{and} \quad r_{jk}^{*} = \frac{r_{jk}}{\pi_{jk}} \ , \tag{12}$$

where π_{jk} refers to the probability of correctly applying the mastered attribute *k* to item *j* and means the probability of not slipping at the attribute level, and r_{jk} stands for the probability of guessing for attribute *k* to item *j*.

In addition, π_j^{*} is meaningfully the same as one minus the probability of slipping at the attribute level in the DINA model (Rupp et al., 2010), assuming conditional independence:

$$\pi_j^{*} = \prod_{k=1}^{K} \pi_{jk}^{\eta_{jk}} = \prod_{k=1}^{K} (1 - s_{jk})^{q_{jk}} \ . \tag{13}$$

The parameter r_{jk} is the probability of guessing at the attribute level, and is defined as a ratio of slipping and guessing probabilities:

$$r_{jk}^{*} = \frac{r_{jk}}{\pi_{jk}} = \frac{g_{jk}}{1 - s_{jk}} \ . \tag{14}$$

The probability of correctly answering item *j* for examinee *i* with $\boldsymbol{\alpha}_{\mathrm{i}}$is

$$P_j(\boldsymbol{\alpha}_i) = P(X_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \pi_j^{*} \prod_{k=1}^{K} r_{jk}^{*q_{jk}(1-\alpha_{jk})} \quad , \tag{15}$$

where α_{jk} refers to the binary mastery status of the k^{th} skill of the j^{th} item, in which 1 denotes mastery of skill k of item j and 0 denotes non-mastery, and η matrix represents binary indicators signifying whether the j^{th} item requires mastery of the k^{th} skill.

The rRUM is a simplified version of the Reparameterized Unified Model (RUM, DiBello et al., 2007). The RUM includes one more continuous latent variable, and is sometimes called the full RUM. It is a cognitive diagnosis model developed by Hartz (2002), and is designed to "model the probability of a correct response as a function of both attribute specific item characteristics and attribute specific examinee characteristics" (Henson, Templin, & Douglas, 2007, p. 561). Practically, the simplified rRUM is used more often than the full RUM, because the "continuous residual ability parameter designed to capture the influence of skills is not explicit in the CDMs" (Henson et al., 2007, p. 561). In addition, the higher-order rRUM, implemented by Templin (2004), used a similar concept of the HO-DINA model to construct a slightly different unidimensional higher order approach. The HO-rRUM provides a one-factor model for the tetrachoric correlation of each pair of attributes. Its general robustness and the associated estimation procedure were validated by Templin et al. (2008).

Learning Sequence and Attribute Hierarchy

This part contains three sections. The first section briefly introduces learning theories that signify the importance of identifying learning sequences. The second section describes the definition of attribute hierarchies. The third section introduces the attribute hierarchy method.

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. Gagné 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). In the late twentieth century, constructivism became popular and is now commonly applied to develop mathematics curricula and sequences of instructions. Constructivists believe learners actively experience and construct their new knowledge. Previous research has shown that students' development of conceptualizations and reasoning can be classified into different levels of sophistication (Battista & Clements, 1996; Battista & Larson, 1994). Teachers need to know what cognitive processes and conceptualizations that students must acquire to make progress in constructing meaning of the new mathematic idea (Battista, 2004).

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.

There have been studies about constructing instructional or learning processes in various subject areas. For example, one recent significant study, called the Hypothetical Learning Trajectories (HLTs, Clements & Sarama, 2004), emphasizes the goals for meaningful student learning, tasks geared to achieve this learning, and hypotheses about the process of student learning. HLTs, based on constructivism, provide richer information of learners' requisite knowledge, development, and difficulties than previous efforts to define learning sequences. In addition, in science education, learning progressions are research-based descriptions of how students build their knowledge, and gain more expertise within and across a discipline over a broad span of time (Duschl, Schweingruber, & Shouse, 2007; Smith, Wiser, Anderson, & Krajcik, 2006). Learning progressions describe a potential learning path that can provide a guide for a coherent curriculum and as such inform the design of coherent curriculum materials.

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).

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 grade-by-grade recommendations for classroom mathematics instruction for both content matter and process. 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.

Researchers' efforts to identify learning sequences in mathematics have also attracted the attention of educators and policy makers. For example, in the area of algebra it has been claimed in the Final Report from the National Mathematics Advisory Panel (2008) of the U.S. Department of Education:

> The coherence and sequential nature of mathematics dictate the foundational skills that are necessary for the learning of algebra. The most important foundational skill not presently developed appears to be proficiency with fractions (including decimals, percent, and negative fractions). The teaching of fractions must be acknowledged as critically important and improved before an increase in student achievement in algebra can be expected. (p. 18).

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). For example, the licensure regulations for mathematics specialists for elementary and middle education by the Virginia Department of Education requires understanding of the sequential nature of mathematics and the mathematical structures inherent in the content strands (http://www.doe.virginia.gov/VDOE/Compliance/ TeacherED/nulicvr.pdf). Teachers' knowledge of the sequential nature of mathematical concepts and capability of applying this knowledge to their instruction will benefit students in learning mathematics.

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 et al., 2007; Leighton et al., 2004; Rupp et al., 2010). These four hierarchies are shown in Figure 2-1 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.

Tables 2-1 through 2-4 from Rupp et al. (2010), modified as indicated, show the matrices of attribute profiles associated with each kind of attribute hierarchies, also using the previous example of six attributes; Table 2-2 was modified by adding two missing profiles in their tables. 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.

Attribute Hierarchy Method

The Attribute Hierarchy Method (AHM) (Gierl, 2007; Gierl, Cui et al., 2009; Gierl, Leighton et al., 2009; Gierl et al., 2007; Gierl & Zhou, 2008) is a psychometric method which applies the structured hierarchical models in developing item and classifying examinee profiles into different attribute patterns. By specifying the hierarchical relationships among attributes, the number of permissible items can be reduced. Without specifying the hierarchical relationships among attributes, there are 2^K - 1 possible rows (items) for a Q-matrix, except for no item measuring no attributes. Imposing the constraints of the attribute hierarchies can produce a reduced Q-matrix. This reduced Q-matrix is a J by K matrix where J is the reduced number of items and K is the number of attributes. The reduced Q-matrix is used to develop items that measure each specific attribute combination specified in the hierarchy, and represents only those items that fit the dependencies defined in the attribute hierarchy. Similarly, by specifying the hierarchical relationships among attributes, the number of possible attribute profiles can be reduced. Without specifying the hierarchical relationships among attributes, there are 2^K possible attribute profiles. The reduced α-matrix consists only of the reduced number of attribute profiles, and is applied in the estimation process. This AHM method is therefore useful to design a test blueprint based on the cognitive attribute hierarchies.

In the statistical pattern classification process, the examinees' observed attribute profiles are compared to their expected attribute profiles under the assumption that the cognitive model is true. In the IRT-based AHM approach (Gierl et al., 2007; Leighton et al., 2004), the expected item characteristic curve can be calculated for each item using the 2 PL IRT model. The *a*- and *b*- item parameters for each item can be estimated based on the expected item response patterns, under the assumption that examinees' responses are consistent with the attribute hierarchy. In the non-IRT based classification method (Gierl & Zhou, 2008), the AHM focuses on classifying examinees' attribute profile patterns and estimating attribute probabilities and item probabilities, unlike the IRT and other CDM models (e.g., DINA, DINO, RUM, etc.) that can estimate item parameters. Furthermore, a person-fit statistic, the Hierarchy Consistency Index (HCI; Cui & Leighton, 2009) has been developed to examine whether examinees' actual item response patterns match the expected response patterns based on the hierarchical relationship among attributes measured by test items.

AHM is a relatively new measurement model, and its related research has grown in recent years. Currently, this approach has not obtained item parameter estimates based on the observed respondents from the real data, and the estimated attribute probabilities are not group invariant (i.e., they are different for different samples). Practically, it is difficult to establish an AHM-based item pool using what the IRT-based approach does conventionally. It is also difficult to conduct linking studies, which are commonly needed for testing programs in developing new forms.

Table 2-1. Linear Attribute Hierarchy Profiles of Six Attributes

	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
Profile 1	0	0	0	0	0	0
Profile 2	1	0	0	0	0	0
Profile 3	1	1	0	0	0	0
Profile 4	1	1	1	0	0	0
Profile 5	1	1	1	1	0	0
Profile 6	1	1	1	1	1	0
Profile 7	1	1	1	1	1	1

Table 2-2. Convergent Attribute Hierarchy Profiles of Six Attributes

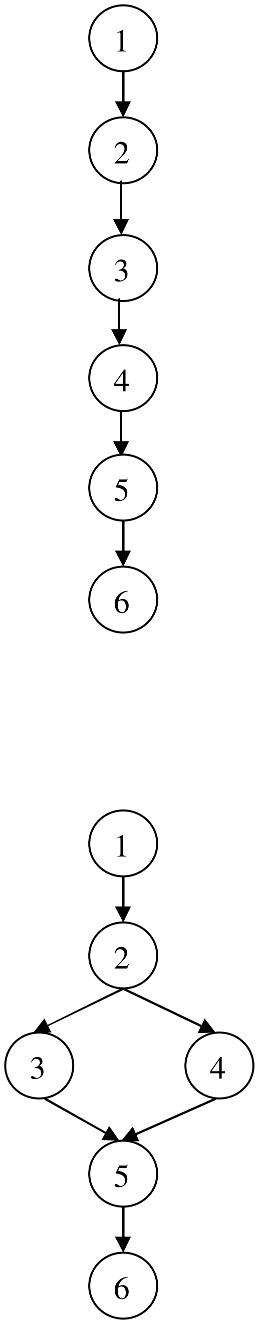
	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
Profile 1	0	0	0	0	0	0
Profile 2	1	0	0	0	0	0
Profile 3	1	1	0	0	0	0
Profile 4	1	1	1	0	0	0
Profile 5	1	1	0	1	0	0
Profile 6	1	1	1	1	0	0
Profile 7	1	1	0	1	1	0
Profile 8	1	1	1	0	1	0
Profile 9	1	1	1	1	1	0
Profile 10	1	1	0	1	1	1
Profile 11	1	1	1	0	1	1
Profile 12	1	1	1	1	1	1

Table 2-3. Divergent Attribute Hierarchy Profiles of Six Attributes

	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
Profile 1	0	0	0	0	0	0
Profile 2	1	0	0	0	0	0
Profile 3	1	1	0	0	0	0
Profile 4	1	1	1	0	0	0
Profile 5	1	0	0	1	0	0
Profile 6	1	0	0	1	1	0
Profile 7	1	0	0	1	0	1
Profile 8	1	0	0	1	1	1
Profile 9	1	1	0	1	0	0
Profile 10	1	1	0	1	1	0
Profile 11	1	1	0	1	0	1
Profile 12	1	1	0	1	1	1
Profile 13	1	1	1	1	0	0
Profile 14	1	1	1	1	1	0
Profile 15	1	1	1	1	0	1
Profile 16	1	1	1	1	1	1

Table 2-4. Unstructured Attribute Hierarchy Profiles of Six Attributes

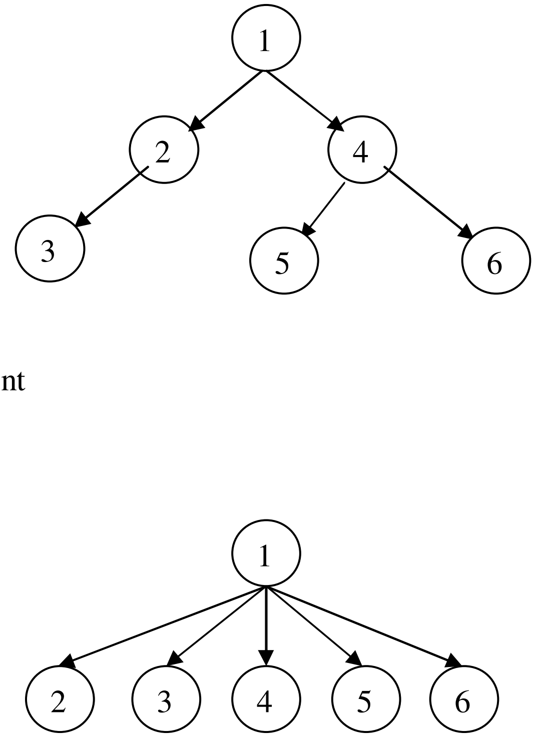
	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute 6
Profile 1	0	0	0	0	0	0
Profile 2	1	0	0	0	0	0
Profile 3	1	1	0	0	0	0
Profile 4	1	0	1	0	0	0
Profile 5	1	0	0	1	0	0
Profile 6	1	0	0	0	1	0
Profile 7	1	0	0	0	0	1
Profile 8	1	1	1	0	0	0
Profile 9	1	1	0	1	0	0
Profile 10	1	1	0	0	1	0
Profile 11	1	1	0	0	0	1
Profile 12	1	0	1	1	0	0
Profile 13	1	0	1	0	1	0
Profile 14	1	0	1	0	0	1
Profile 15	1	0	0	1	1	0
Profile 16	1	0	0	1	0	1
Profile 17	1	0	0	0	1	1
Profile 18	1	1	1	1	0	0
Profile 19	1	1	1	0	1	0
Profile 20	1	1	1	0	0	1
Profile 21	1	1	0	1	1	0
Profile 22	1	1	0	1	0	1
Profile 23	1	1	0	0	1	1
Profile 24	1	0	1	1	1	0
Profile 25	1	0	1	0	1	1
Profile 26	1	0	1	1	0	1
Profile 27	1	0	0	1	1	1
Profile 28	1	1	1	1	1	0
Profile 29	1	1	1	0	1	1
Profile 30	1	1	0	1	1	1
Profile 31	1	0	1	1	1	1
Profile 32	1	1	1	1	0	1
Profile 33	1	1	1	1	1	1



a. Linear

b. Convergent

Figure 2-1. Attribute Hierarchies of Six Attributes



c. Divergent

d. Unstructured

Figure 2-1.––continued

CHAPTER III

METHODOLOGY

This chapter describes methodology used in the current dissertation. It contains four main sections. The first section revisits the research questions and presents the analyses to address these questions. The second section introduces the proposed models. The third section describes the real data analysis. The fourth section consists of the simulation design and the detailed procedures.

Research Purposes and Questions Revisited

The purposes of the study are to apply the hierarchical models of the cognitive skills when using two cognitive diagnostic models - DINA and DINO- to analyze both simulated data and the retrofitting of TIMSS 2003 mathematics data. To achieve these purposes, three research questions are to be answered. This section describes how each research question is addressed.

Research Question 1

Research Question 1 examines the proposed DINA-H and DINO-H models.

1. *How do the proposed DINA-H and DINO-H perform?*
    1.1 *Do the DINA-H and DINO-H models provide reasonable item parameter estimates?*

To address Research Question 1.1, the calibration results from the real data analysis using the DINA-H and DINO-H models were evaluated and compared to the results from the DINA and DINO models. The following statistics were compared: The AIC and BIC for the model fit, and the δ and IDI for the item fit.

1.2 Do the DINA-H and DINO-H models provide stable calibration results with small sample size?

The calibration results from the real data analysis using the DINA/DINO models and the DINA-H/DINO-H models with the US sample size were compared to the results with a larger sample size. The following statistics were compared: The AIC and BIC for the model fit, and the δ and IDI for the item fit.

1.3 Which CDM (the DINA-H or DINO-H model) performs better while applying a skill hierarchy?

The calibration results from analyzing two datasets of the two TIMSS 2003 mathematics booklets and the simulation study with the DINA and DINA-H models were compared to the results analyzed via the DINO and DINO-H models. To address Research Question 1.3, the AIC and BIC for the model fit, the δ and IDI for the item fit, and correlations of item parameter estimates between different models and sample sizes were compared.

Research Question 2

Research Question 2 concerns the issue when the assumptions about the relationships among attributes are inconsistent with the estimation models in the simulation study.

2. *When skills are ordered hierarchically, how do the conventional DINA and DINO models and the proposed DINA-H and DINO-H models compare?*

2.1 How do the conventional and new models compare in terms of parameter estimates?

The attributes adapted in developing the Q-matrix for the TIMSS 2003 mathematics test were constructed into a hierarchical structure by two content experts in mathematics learning. The calibration results from the real data analysis and the main effect of model consistency from the simulation analysis using the DINA-H and DINO-H models were evaluated and compared to the results from the DINA and DINO models. In addition, the results from the simulation analysis for data simulated from the hierarchically ordered skills were analyzed and compared via the new DINA-H and DINO-H models and the conventional DINA and DINO models. The following statistics were compared: The AIC and BIC for the model fit, the δ and IDI for the item fit, and the ASB, AVAR, and AMSE of item parameter estimates for each condition.

2.2 How do the performances of the conventional and new models compare under varying conditions of different numbers of attributes, different test lengths, and sample sizes?

In the simulation analysis, data simulated from different hierarchically ordered skills were analyzed via the new DINA-H and DINO-H models and the conventional DINA and DINO models. The calibration results from the new DINA-H and DINO-H models were compared to the conventional DINA and DINO models for the main effects of different numbers of attributes, test lengths, and sample sizes, and the interaction effects of test length by attribute, sample size by attribute, and sample size by test length, and the three-way interaction effect of sample size by test length by attribute. The ASB, AVAR, and AMSE of item parameter estimates for each condition were compared.

Research Question 3

Research Question 3 concerns the misspecification of a skill hierarchy on the results of model estimations.

3. *What is the impact of misspecification of a skill hierarchy on the DINA(-H) and the DINO(-H) models?*

3.1 Do the item parameters recover well, when a specified skill hierarchy is inconsistent with an estimation model?

The calibration results from simulation analysis using varying estimation models consistent or inconsistent with the specifications on the skill hierarchies were compared to each other. The following statistics were compared under each condition: The AIC and BIC for the model fit, and the ASB, AVAR, and AMSE of item parameter estimates for each condition.

3.2 How do the models perform and compare under varying conditions with different numbers of attributes, test lengths, and sample sizes?

The calibration results from simulation analysis using varying estimation models consistent or inconsistent with the specifications on the skill hierarchies were compared to each other for the main effects of different numbers of attributes, test lengths, and sample sizes, and the interaction effects of test length by attribute, sample size by attribute, and sample size by test length, and the three-way interaction effect of sample size by test length by attribute.

In addition, the calibration results from analyzing the DINA-H/DINO-H models with different skill hierarchies for different sample sizes were compared to each other, and also to the DINA/DINO model. More specifically, the bias indices for the DINA-H/DINO-H models with higher dependent skill hierarchy were compared to the DINA-H/DINO-H models with the lower dependent skill hierarchy for various sample sizes in the simulation analysis. The ASB, AVAR, and AMSE of item parameter estimates for each condition were evaluated.

Proposed DINA-H and DINO-H Models

The study proposed two modified CDMs: The DINA with hierarchy (DINA-H) model and the DINO with hierarchy (DINO-H) model. 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 proposed DINA-H and DINO-H models were applied with real data and simulation analyses and were compared with the conventional DINA and DINO models.

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.

Four common types of cognitive attribute hierarchies are: linear, convergent, divergent, and unstructured. The skill hierarchies and the attribute profiles for the condition of six attributes were described in Figure 2-1 and Tables 2-1 to 2-4 in Chapter II. Figure 3-1 shows the attribute hierarchies for the condition of eight attributes. The condition of eight attributes includes two more attributes at the end of the hierarchy than the condition of six attributes. Like the condition of six attributes, the linear attribute hierarchy (see Table 3-1) of eight attributes requires all attributes to be ordered sequentially. In a linear hierarchy, for an examinee to have mastered attribute 8, he or she must have also mastered attributes 1 through 7. The convergent attribute hierarchy (see Table 3-2) specifies a situation in which a single attribute could be the prerequisite of multiple different attributes and situations in which a single attribute could require mastering one or more of the multiple preceding attributes. The relationships among attributes 2 to 5 are the same as specified in the condition of six attributes. In the condition of eight attributes, attribute 5 is the prerequisite of attributes 6 and 7, and mastering either attribute 6 or 7 could lead to mastering attribute 8. The divergent attribute hierarchy (see Table 3-3) refers to different distinct paths originating from the same single attribute. The relationships among attributes 1 and 4 to 6 are the same as specified in the condition of six attributes. In the condition of eight attributes, attributes 7 and 8 appear parallel at the end of attribute 3. That means when an examinee has mastered attributes 7 or 8, he or she has also mastered attributes 1 to 3. The unstructured attribute hierarchy (see Table 3-4) describes cases when a single attribute could be prerequisite to multiple different attributes which have no direct relationships to each other. Similar to the condition of six attributes, attribute 1 is the common prerequisite to attributes 2 to 8 in the condition of eight attributes. However, a certain type of a hierarchy model could have various types of structures, except for the liner hierarchy model. The specified conditions in Figure 3-1 are ones among many possible structures under a hierarchy. This is especially so for the convergent and the divergent hierarchies.

Based on the attribute hierarchies, the number of attribute profiles was found for each hierarchical model. The attribute profiles for the condition of six attributes were listed in Chapter II. Tables 3-1 to 3-4 list the attribute profiles of each attribute hierarchy for the condition of eight attributes. Table 3-5 presents the total number of attribute profiles of each attribute hierarchy model for the condition of both six and eight attributes. 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. Since the convergent and the divergent hierarchies could have varying structures, the numbers of possible attribute profiles were different for various structures.

Real Data Analysis

The data used in this study included both real and simulated data. The real data 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.

Description of Data

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 3-6 summarizes the difference between Booklet 1 and Booklet 2.

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 3-7 provides the attributes modified form the CCSS and their corresponding TIMSS items. To illustrate, Table 3-8 shows one item from each booklet with the attributes being measured. The Q-matrices of B1 and B2 are shown in Table 3-9 and 3-10, 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 3-2 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 3.3 and 3.4, 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.

Analysis

This section describes the steps conducted in the model estimation. The first part lists the conditions, the second part briefly reviews the models used in this study, the third part describes the model estimation process using the EM algorithm, and the final part describes the evaluation indices.

Conditions

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 for the real data analysis.

Models

The real data were analyzed by fitting the DINA, DINA-H, DINO, and DINO-H models, using the expectation-maximization (EM) algorithm based on the marginal maximum likelihood estimation. The EM approach was chosen because its results were reproducible and could be replicated. The DINA and DINO models were described in detail in Chapter II. Both models require a specified binary Q-matrix, which identifies the attributes measured by the items. The DINA-H and DINO-H models were described earlier in the chapter.

Parameter Estimation Process

To estimate examinee attribute profiles and item parameters, the procedure outlined by de la Torre (2009) was followed. The EM algorithm was implemented by first setting initial parameter values, then estimating the expected value of the unknown variables, and giving the current parameter estimates. The final step was to re-estimate the posterior distribution, by maximizing the likelihood of the data by computing the marginal maximum likelihood estimation given the expected estimates of the unknown variables. The estimation steps were repeated until convergence was achieved. Because the estimation process was initialized with a flat prior distribution, the prior was updated in each iteration using the expected proportions of examinees in each latent class (Huebner & Wang, 2011). This method is referred to as the empirical Bayesian method in Carlin and Louis (1996). The estimations began with all the slip and guessing parameters set to 0.2 (Huebner, 2009; Huebner & Wang, 2011). It was shown in a simulation study that the results from the EM estimation were not sensitive to the initial values of the parameters as long as the true guess and slip parameters reasonably fall between 0 and 0.5, and it was demonstrated that the differences in results at varying levels of initial values was negligible (Huebner, 2009). In the DINA and DINA-H models, the η-matrix is used to indicate whether examinee *i* has mastered all of the required skills for item *j*, whereas in the DINO and DINO-H models the ω-matrix ($\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}}$) is used. 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 outlined as follows, using the notation for the DINA model as an example. The first step in computing the posterior matrix was to calculate the matrix of marginalized maximum likelihood estimation likelihood. For the observed data *X* and the attribute profiles α :

$$\text{Likelihood}(X) = \prod_{i=1}^{I} \text{Likelihood}(X_i) = \prod_{i=1}^{I} \sum_{l=1}^{L} \text{Likelihood}(X_i \mid \alpha_l) p(\alpha_l), \tag{16}$$

where $\text{Likelihood}(X_i)$ is the marginalized likelihood of the response vector of examinee *i*, and $p(\alpha_l)$ is the prior probability of the attribute profile vector α_l .

The next step toward computing the posterior matrix was to multiply the columns of the likelihood matrix by the prior probability for the corresponding skill pattern with a flat (non-informative) prior distribution, meaning that each skill pattern had a probability of 1/*L*, where *L* equals 2^K in the conventional DINA and DINO models and *L* equals the number of all possible attribute profiles specified for each DINA-H or DINO-H model.

Parameter estimation based on the marginalized likelihood (i.e., the marginal maximum likelihood estimation) was implemented using the EM algorithm. To obtain the maximum likelihood estimate, the following is maximized:

$$\ln(X) = \log \prod_{i=1}^{I} \text{Likelihood}(X_i) = \sum_{i=1}^{I} \log \text{Likelihood}(X_i). \tag{17}$$

The expected number of examinees with attribute profile α_l is computed from

$$I_l = \sum_{i=1}^{I} p(\alpha_l \mid X_i), \tag{18}$$

where $p(\alpha_l \mid X_i)$ is the posterior probability that examinee *i* has the attribute profile α_l . Moreover, the expected number of examinees with attribute profile α_l answering item *j* correctly is defined as:

$$R_{jl} == \sum_{i=1}^{I} P(\alpha_l \mid X_i) X_{ij}. \tag{19}$$

Finally, item parameters were estimated when $\eta = 0$, using

$$\hat{g}_j = \frac{R_{jl}^{(0)}}{I_{jl}^{(0)}}, \tag{20}$$

$$\text{and } \hat{s}_j = \frac{[I_{jl}^{(1)} - R_{jl}^{(1)}]}{I_{jl}^{(1)}} \;, \tag{21}$$

where $I_{jl}^{(0)}$ is the expected number of examinees lacking at least one of the required attributes for item *j*, and $R_{jl}^{(0)}$ is the expected number of examinees among $I_{jl}^{(0)}$ correctly answering item *j*. $I_{jl}^{(1)}$ and $R_{jl}^{(1)}$ have the same interpretation except that they pertain to the examinees with all the required attributes for item *j*. $I_{jl}^{(0)} + I_{jl}^{(1)}$ is equal to I_l for all *j*.

As mentioned earlier in this section, the algorithm started with initial values for *g* and *s* both equal to 0.2. Next, $I_{jl}^{(0)}$, $R_{jl}^{(0)}$, $I_{jl}^{(1)}$ and $R_{jl}^{(1)}$ were computed based on the current values of *g* and *s*. Then, the values of *g* and *s* were found and updated. The steps were repeated until convergence was achieved. The criterion for convergence was set to be smaller than 0.001 for both real data and simulation studies. The number of iteration cycles was smaller than 100 across all the conditions of the simulation.

The study modified scripts written for use in the R software environment (R Development Core Team, 2011) to incorporate the hierarchy in the DINA-H and DINO-H models for statistically conducting the real data and the simulation analyses.

Evaluation Indices

The study evaluated and compared model fit and item fit for each condition in the real data analysis for the DINA, DINA-H, DINO, and DINO-H models. The details for assessing each evaluation index are described in the next section.

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. These values were computed for each condition in the real data analysis.

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:

$$\mathrm{AIC} = -2\ln(\mathrm{Likelihood}) + 2p \ , \tag{22}$$

where ln(Likelihood) is the log-likelihood of the data under the model (see Equation 16 and 17) 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 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:

$$\mathrm{BIC} = -2\ln(\mathrm{Likelihood}) + p\ln(N)\,, \tag{23}$$

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 was provided in equation 5 in Chapter II.

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:

$$IDI_j = 1 - \frac{g_j}{1 - s_j} \quad . \tag{24}$$

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

Simulation Study

The simulation studies attempted to address the accuracy of the item parameter recovery when the cognitive skills were in a specified hierarchy. Varying estimation models were applied. The simulation analysis examined and compared the impacts of the misspecification of a skill hierarchy on various estimation models under their varying assumptions of independent attributes or hierarchically related attributes. The simulation design described in the next session was used to investigate the relationship between the numbers of attributes, test lengths, sample sizes, and various CDMs. Attention was also paid to the conditions while analyzing the hierarchically structured CDMs with smaller sample sizes and contrasting them with the conventional CDMs.

Simulation Design

This section includes two parts. The first part of the simulation procedure describes each factor manipulated in the simulation study, and the second part of simulation procedure describes the steps carried out in the simulation process.

Simulation Factors

The simulation design manipulated five factors: the number of attributes, the test lengths, the data generating models, the sample sizes, and the estimation models. Table 3-11 displays the values of these simulation factors.

The first simulation factor was the number of attributes. There were six or eight attributes assessed in each test, which were within the usual range of those found with current applications of CDMs (Rupp & Templin, 2008). A review of the literature on multiple classification models showed that most application examples used about four to eight attributes (Hartz, 2002; Maris, 1999; Rupp & Templin, 2008b). Six and eight attributes were chosen in the current study because the intention of the study was to maintain consistency with the previous studies using six attributes (Gierl, et. al., 2007; Leighton et. al., 2004; Rupp, et. al., 2010), and also to understand the effect of reducing the sample size requirements for tests measuring more attributes.

The second simulation factor was test length (i.e., the number of items). There were 12 or 30 items in each test. These numbers were chosen because the usual range of items measured in the most current applications of CDMs is two to four items for every single attribute (i.e., all columns of the Q-matrix sum up to 2 or 4) (Rupp & Templin, 2008a).

The third simulation factor was the identity of the data-generating model. To demonstrate the proposed model and to effectively answer the research questions, the study chose two of the skill hierarchies to apply in the simulation analysis, which were the linear hierarchy and the unstructured hierarchy. The linear hierarchy was the most constrained hierarchy with the least number of possible attribute profiles. The unstructured hierarchy was the least restricted hierarchy with about half of the number of possible attribute profiles of the conventional DINA/DINO model. Six different models were considered: the DINA model, the DINA model with linear hierarchy (DINA-H_L), the DINA model with unstructured hierarchy (DINA-H_U), the DINO model, the DINO model with linear hierarchy (DINO-H_L), and the DINO model with unstructured hierarchy (DINO-H_U). The baseline DINA and DINO models represented the conditions in which no constraint was imposed on the relationships among attributes. The skill hierarchies and the attribute profiles for the condition of six attributes were as described in Chapter II, and for the conditions of eight attributes were described in the model section in Chapter III.

The fourth simulation factor was the sample size. Three values were used: 300, 1,000, and 3,000 examinees. Most simulation studies typically use larger sample sizes to estimate models, with the samples ranging from 500 to 10,000 respondents (e.g., Rupp & Templin, 2008a). The smallest sample size in the current study was chosen to examine the performance of the proposed DINA-H and DINO-H models under small sample-size conditions.

The fifth factor was the identity of the estimation model. Six different models were considered: the DINA, DINA-H_L, DINA-H_U, DINO, DINO-H_L, and DINO-H_U models. While estimating with each skill hierarchical model, the specified possible attribute profiles of the model was applied to be $\boldsymbol{\alpha_i} = \{ \alpha_{ik} \}$ (the examinee's binary skills vector) in the EM algorism, and to replace the amount of 2^K attribute profiles in the conventional DINA and DINO models. In addition, the study only considered the cross comparisons of estimating data generated from the DINA, DINA-H_L, and DINA-H_U models by using DINA-based models, and estimating data generated from the DINO, DINO-H_L, and DINO-H_U models by using the DINO-based models. Cross estimations between the DINA and DINO models were not considered.

Simulation Procedure

The simulation steps were listed below.

1. The study first simulated four item-by-attribute Q-matrices (i.e., 12 items measuring 6 attributes, 12 items measuring 8 attributes, 30 items measuring 6 attributes, and 30 items measuring 8 attributes).

To be closer to the practical testing, the distribution of the percentage of items measuring varying numbers of attributes from TIMSS data was used as the guideline in simulating the Q-matrices. The current study adapted the distribution based on the Q-matrix in Park, Lee, and Choi (2010) and Choi (2011). Their Q-matrix was developed by using the attributes form the National Council of Teachers of Mathematics (NCTM) *Principles and Standards for School Mathematics* (NCTM, 2000) for coding, and from the consensus of three mathematics educators. In Park et al. (2010)'s and Choi (2011)'s studies, 39.4% of the coded attributes measured number and operations, 28.2% of the coded attributes measured algebra, 16.9% of the coded attributes measured geometry, 5.6% of the coded attributes measured measurement; and 9.9% of the coded attributes measured data analysis and probability. Table 3-12 shows the Q-matrix from Park et al. (2010) and Choi (2011) for the eighth grade TIMSS 2007 mathematics test. In their Q-matrix, three (10%) of the items measured one attribute, 13 (45%) of the items measured two attributes, ten (35%) of the items measured three attributes, and three (10%) of the items measured four attributes. This Q-matrix consists of 29 items measuring 12 attributes which were coded 71 times in total. Hence, in the current study the distribution of item percentages was set to 30% of items measuring one attribute, 60% of items measuring two attributes, and 10% of items measuring three attributes for the condition of six attributes in the Q-matrix. The distribution of item percentages was set to 10% of items measuring one attribute, 45% of items measuring two attributes, 35% of items measuring three attributes, and 10% of items measuring four attributes for the condition of eight attributes in the Q-matrix. Each row of the Q-matrix sums up to at least 1, which means that each item should have assessed at least one attribute.

2. The true item parameters of guessing and slip for each corresponding Q-matrix were simulated from a random uniform distribution with a lower bound of 0.05 and an upper bound of 0.4. The true item parameters were simulated differently for each unique Q-matrix, but were set the same for various models, different sample sizes, and across replications.

3. Based on the generated Q-matrix, the simulated guessing and slip parameters, and the α-matrix, the η-matrix was computed using Equation 4 in Chapter 2 for the N examinees (i.e., 300, 1000, or 3000) for the DINA(-H) model, and the ω-matrix was

computed using Equation 10 for the DINO(-H) model. Once the η-matrix and ω-matrix were obtained, the probability to answer each item correctly (P, from Equation 3 for the DINO(-H) and Equation 9 for the DINO(-H) model) was computed for each examinee. Based on these probabilities, the 0/1 data were simulated randomly from the binomial distribution for each examinee under different simulation conditions. Each data were simulated differently for each replication under each condition; however, their corresponding true item parameters were set the same for the 50 replications under each condition to make comparisons easier.

4. While simulating the data for the DINA and DINO models, the number of possible attribute profiles (L) was set to 2^K, and α was a matrix consisting of all the possible combinations of 0s and 1s for all attributes. While simulating the data for the DINA-H and DINO-H models, the number of possible attribute profiles (L) was set to the pre-specified reasonable examinee attribute profiles from various cognitive skill hierarchies, and α was a matrix consisting of all the pre-specified reasonable examinee attribute profiles from various cognitive skill hierarchies.

In sum, the 2 (the numbers of attributes: 6 and 8) x 2 (the numbers of items: 12 and 30) x 3 (data generating models: conventional, H-linear, H-unstructured) x 3 (sample sizes: 300, 1000, 3000) x 3 (estimation models: conventional, H-linear, H-unstructured) x 2 (CDMs: DINA and DINO) design results in a total of 216 conditions. Each condition was replicated 50 times. Tables 3-13 and 3-14 list all the conditions for the DINA/DINA-H and DINO/DINO-H models, respectively. The DINA/DINO-H_L models referred to the DINA/DINO models with the linear hierarchy, and the DINA/DINO-H_U models referred to the DINA/DINO models with the unstructured hierarchy.

These simulation studies examined the question of which simulated data best fit the evaluating models, based on different skill hierarchies under different conditions. Using simulated data, model parameter recovery was evaluated. Special attention was paid to the robustness of the models, and whether they were able to fit data that were consistent or inconsistent with the assumptions of the relationship among attributes held by the evaluating models.

Analysis

Model and Parameter Estimation

Six models, specifically the DINA, DINA-$H_{L,}$ DINA-H_U, DINO, DINO-H_L, and DINO-$H_{U,}$ were used to analyze the simulated data. Analysis was performed under the expectation-maximization (EM) algorithm based on the marginal maximum likelihood estimation. All models and their estimation procedures were as described earlier in the chapter under the section of real data analysis.

Evaluation Indices

Fit Indices

Model fit evaluations were to address the question of whether the data simulated based on different hierarchical models under varying conditions fit the evaluating models. Again, special attention was paid to the robustness of the model to fit the simulated data that were not consistent with the evaluating model. The mean of AIC and BIC for each condition are reported in the result section in which the terms MAIC and MBIC are used.

Summary Statistics

Three bias indices were used to evaluate model parameter recovery: the average squared bias (ASB), the average variance (AVAR), and the average mean-square error (AMSE) of item parameter estimates (*s* and *g*) for each condition. Under each simulation condition, the bias for each item was defined as the difference between the average item parameter estimates over replications and their corresponding true generating values. The following formulas use the guessing parameter as an example.

$$Bias_j = \bar{g}_j - g_j^* = (\frac{1}{R}\sum_{r=1}^{R} g_{jr}) - g_j^* \tag{25}$$

where $\bar{g}_j$ is the average *g* parameter estimate for item *j* over replications, g_j^* is the true generating value for item *j*, and *R* refers to the number of replications under each condition. The ASB for each condition is defined as the average squared bias:

$$ASB(g) = \frac{1}{J}\sum_{j=1}^{J} Bias_j^2 = \frac{1}{J}\sum_{j=1}^{J} (\bar{g}_j - g_j^*)^2 \,. \tag{26}$$

Furthermore, AVAR is defined as the average variance of an item parameter across replications for each condition as:

$$AVAR(g) = \frac{1}{J}\sum_{j=1}^{J} \frac{(g_{jr} - \bar{g}_j)^2}{R} \,, \tag{27}$$

where g_{jr} is the guessing parameter estimate for item *j* for replication *r*.

Finally, since the mean squared error is equal to the squared bias plus variance, the AMSE is regarded as a combination of information from variance and bias, and is defined as:

$$AMSE(g) = \frac{1}{J}\sum_{j=1}^{J}\sum_{r=1}^{R}\frac{(g_{jr} - g_j^*)^2}{R} \ . \qquad (28)$$

Table 3-1. Linear Attribute Hierarchy Profiles of Eight Attributes

	Attribute							
	1	2	3	4	5	6	7	8
Profile 1	0	0	0	0	0	0	0	0
Profile 2	1	0	0	0	0	0	0	0
Profile 3	1	1	0	0	0	0	0	0
Profile 4	1	1	1	0	0	0	0	0
Profile 5	1	1	1	1	0	0	0	0
Profile 6	1	1	1	1	1	0	0	0
Profile 7	1	1	1	1	1	1	0	0
Profile 8	1	1	1	1	1	1	1	0
Profile 9	1	1	1	1	1	1	1	1

Table 3-2. Convergent Attribute Hierarchy Profiles of Eight Attributes

	Attribute							
	1	2	3	4	5	6	7	8
Profile 1	0	0	0	0	0	0	0	0
Profile 2	1	0	0	0	0	0	0	0
Profile 3	1	1	0	0	0	0	0	0
Profile 4	1	1	1	0	0	0	0	0
Profile 5	1	1	0	1	0	0	0	0
Profile 6	1	1	0	1	1	0	0	0
Profile 7	1	1	1	0	1	0	0	0
Profile 8	1	1	1	1	0	0	0	0
Profile 9	1	1	1	1	1	0	0	0
Profile 10	1	1	0	1	1	1	0	0
Profile 11	1	1	1	0	1	1	0	0
Profile 12	1	1	1	1	1	1	0	0
Profile 13	1	1	0	1	1	0	1	0
Profile 14	1	1	1	0	1	0	1	0
Profile 15	1	1	1	1	1	0	1	0
Profile 16	1	1	0	1	1	1	1	0
Profile 17	1	1	1	0	1	1	1	0
Profile 18	1	1	1	1	1	1	1	0
Profile 19	1	1	0	1	1	1	0	1
Profile 20	1	1	1	0	1	1	0	1
Profile 21	1	1	1	1	1	1	0	1
Profile 22	1	1	0	1	1	0	1	1
Profile 23	1	1	1	0	1	0	1	1
Profile 24	1	1	1	1	1	0	1	1
Profile 25	1	1	0	1	1	1	1	1
Profile 26	1	1	1	0	1	1	1	1
Profile 27	1	1	1	1	1	1	1	1

Table 3-3. Divergent Attribute Hierarchy Profiles of Eight Attributes

	Attribute							
	1	2	3	4	5	6	7	8
Profile 1	0	0	0	0	0	0	0	0
Profile 2	1	0	0	0	0	0	0	0
Profile 3	1	1	0	0	0	0	0	0
Profile 4	1	1	1	0	0	0	0	0
Profile 5	1	0	0	1	0	0	0	0
Profile 6	1	0	0	1	1	0	0	0
Profile 7	1	0	0	1	0	1	0	0
Profile 8	1	0	0	1	1	1	0	0
Profile 9	1	1	0	1	0	0	0	0
Profile 10	1	1	0	1	1	0	0	0
Profile 11	1	1	0	1	0	1	0	0
Profile 12	1	1	0	1	1	1	0	0
Profile 13	1	1	1	1	0	0	0	0
Profile 14	1	1	1	1	1	0	0	0
Profile 15	1	1	1	1	0	1	0	0
Profile 16	1	1	1	1	1	1	0	0
Profile 17	1	1	1	0	0	0	1	0
Profile 18	1	1	1	1	0	0	1	0
Profile 19	1	1	1	1	1	0	1	0
Profile 20	1	1	1	1	0	1	1	0
Profile 21	1	1	1	1	1	1	1	0
Profile 22	1	1	1	0	0	0	0	1
Profile 23	1	1	1	1	0	0	0	1
Profile 24	1	1	1	1	1	0	0	1
Profile 25	1	1	1	1	0	1	0	1
Profile 26	1	1	1	1	1	1	0	1
Profile 27	1	1	1	0	0	0	1	1
Profile 28	1	1	1	1	0	0	1	1
Profile 29	1	1	1	1	1	0	1	1
Profile 30	1	1	1	1	0	1	1	1
Profile 31	1	1	1	1	1	1	1	1

Table 3-4. Unstructured Attribute Hierarchy Profiles of Eight Attributes

	Attribute							
	1	2	3	4	5	6	7	8
Profile 1	0	0	0	0	0	0	0	0
Profile 2	1	0	0	0	0	0	0	0
Profile 3	1	1	0	0	0	0	0	0
Profile 4	1	0	1	0	0	0	0	0
Profile 5	1	0	0	1	0	0	0	0
Profile 6	1	0	0	0	1	0	0	0
Profile 7	1	0	0	0	0	1	0	0
Profile 8	1	0	0	0	0	0	1	0
Profile 9	1	0	0	0	0	0	0	1
Profile 10	1	1	1	0	0	0	0	0
Profile 11	1	1	0	1	0	0	0	0
Profile 12	1	1	0	0	1	0	0	0
Profile 13	1	1	0	0	0	1	0	0
Profile 14	1	1	0	0	0	0	1	0
Profile 15	1	1	0	0	0	0	0	1
Profile 16	1	0	1	1	0	0	0	0
Profile 17	1	0	1	0	1	0	0	0
Profile 18	1	0	1	0	0	1	0	0
Profile 19	1	0	1	0	0	0	1	0
Profile 20	1	0	1	0	0	0	0	1
Profile 21	1	0	0	1	1	0	0	0
Profile 22	1	0	0	1	0	1	0	0
Profile 23	1	0	0	1	0	0	1	0
Profile 24	1	0	0	1	0	0	0	1
Profile 25	1	0	0	0	1	1	0	0
Profile 26	1	0	0	0	1	0	1	0
Profile 27	1	0	0	0	1	0	0	1
Profile 28	1	0	0	0	0	1	1	0
Profile 29	1	0	0	0	0	1	0	1
Profile 30	1	0	0	0	0	0	1	1
Profile 31	1	1	1	1	0	0	0	0
Profile 32	1	1	1	0	1	0	0	0
Profile 33	1	1	1	0	0	1	0	0
Profile 34	1	1	1	0	0	0	1	0
Profile 35	1	1	1	0	0	0	0	1
Profile 36	1	1	0	1	1	0	0	0
Profile 37	1	1	0	1	0	1	0	0

Table 3-4.––continued

	Attribute							
	1	2	3	4	5	6	7	8
Profile 38	1	1	0	1	0	0	1	0
Profile 39	1	1	0	1	0	0	0	1
Profile 40	1	1	0	0	1	1	0	0
Profile 41	1	1	0	0	1	0	1	0
Profile 42	1	1	0	0	1	0	0	1
Profile 43	1	1	0	0	0	1	1	0
Profile 44	1	1	0	0	0	1	0	1
Profile 45	1	1	0	0	0	0	1	1
Profile 46	1	0	1	1	1	0	0	0
Profile 47	1	0	1	1	0	1	0	0
Profile 48	1	0	1	1	0	0	1	0
Profile 49	1	0	1	1	0	0	0	1
Profile 50	1	0	1	0	1	1	0	0
Profile 51	1	0	1	0	1	0	1	0
Profile 52	1	0	1	0	1	0	0	1
Profile 53	1	0	1	0	0	1	1	0
Profile 54	1	0	1	0	0	1	0	1
Profile 55	1	0	1	0	0	0	1	1
Profile 56	1	0	0	1	1	1	0	0
Profile 57	1	0	0	1	1	0	1	0
Profile 58	1	0	0	1	1	0	0	1
Profile 59	1	0	0	1	0	1	1	0
Profile 60	1	0	0	1	0	1	0	1
Profile 61	1	0	0	1	0	0	1	1
Profile 62	1	0	0	0	1	1	1	0
Profile 63	1	0	0	0	1	1	0	1
Profile 64	1	0	0	0	1	0	1	1
Profile 65	1	0	0	0	0	1	1	1
Profile 66	1	1	1	1	1	0	0	0
Profile 67	1	1	1	1	0	1	0	0
Profile 68	1	1	1	1	0	0	1	0
Profile 69	1	1	1	1	0	0	0	1
Profile 70	1	1	1	0	1	1	0	0
Profile 71	1	1	1	0	1	0	1	0
Profile 72	1	1	1	0	1	0	0	1
Profile 73	1	1	1	0	0	1	1	0
Profile 74	1	1	1	0	0	1	0	1

Table 3-4.––continued

	Attribute							
	1	2	3	4	5	6	7	8
Profile 75	1	1	1	0	0	0	1	1
Profile 76	1	1	0	1	1	1	0	0
Profile 77	1	1	0	1	1	0	1	0
Profile 78	1	1	0	1	1	0	0	1
Profile 79	1	1	0	1	0	1	1	0
Profile 80	1	1	0	1	0	1	0	1
Profile 81	1	1	0	1	0	0	1	1
Profile 82	1	1	0	0	1	1	1	0
Profile 83	1	1	0	0	1	1	0	1
Profile 84	1	1	0	0	1	0	1	1
Profile 85	1	1	0	0	0	1	1	1
Profile 86	1	0	1	1	1	1	0	0
Profile 87	1	0	1	1	1	0	1	0
Profile 88	1	0	1	1	1	0	0	1
Profile 89	1	0	1	1	0	1	1	0
Profile 90	1	0	1	1	0	1	0	1
Profile 91	1	0	1	1	0	0	1	1
Profile 92	1	0	1	0	1	1	1	0
Profile 93	1	0	1	0	1	1	0	1
Profile 94	1	0	1	0	1	0	1	1
Profile 95	1	0	1	0	0	1	1	1
Profile 96	1	0	0	1	1	1	1	0
Profile 97	1	0	0	1	1	1	0	1
Profile 98	1	0	0	1	1	0	1	1
Profile 99	1	0	0	1	0	1	1	1
Profile 100	1	0	0	0	1	1	1	1
Profile 101	1	1	1	1	1	1	0	0
Profile 102	1	1	1	1	1	0	1	0
Profile 103	1	1	1	1	1	0	0	1
Profile 104	1	1	1	1	0	1	1	0
Profile 105	1	1	1	1	0	1	0	1
Profile 106	1	1	1	1	0	0	1	1
Profile 107	1	1	1	0	1	1	1	0
Profile 108	1	1	1	0	1	1	0	1
Profile 109	1	1	1	0	1	0	1	1
Profile 110	1	1	1	0	0	1	1	1
Profile 111	1	1	0	1	1	1	1	0

Table 3-4.––continued

	Attribute							
	1	2	3	4	5	6	7	8
Profile 112	1	1	0	1	1	1	0	1
Profile 113	1	1	0	1	1	0	1	1
Profile 114	1	1	0	1	0	1	1	1
Profile 115	1	1	0	0	1	1	1	1
Profile 116	1	0	1	1	1	1	1	0
Profile 117	1	0	1	1	1	1	0	1
Profile 118	1	0	1	1	1	0	1	1
Profile 119	1	0	1	1	0	1	1	1
Profile 120	1	0	1	0	1	1	1	1
Profile 121	1	0	0	1	1	1	1	1
Profile 122	1	1	1	1	1	1	1	0
Profile 123	1	1	1	1	1	1	0	1
Profile 124	1	1	1	1	1	0	1	1
Profile 125	1	1	1	1	0	1	1	1
Profile 126	1	1	1	0	1	1	1	1
Profile 127	1	1	0	1	1	1	1	1
Profile 128	1	0	1	1	1	1	1	1
Profile 129	1	1	1	1	1	1	1	1

Table 3-5. The Possible Number of Attribute Profiles for Each Hierarchy Model

	Possible number of attribute profile	
Attribute hierarchy	6 attributes	8 attributes
Linear	7	9
Convergent	12	27
Divergent	16	31
Unstructured	33	129
Baseline	$2^6 = 64$	$2^8 = 256$

Table 3-6. 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	10th	12th
Number of possible attribute profiles	726	690

Table 3-7. 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 multistep 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 3-8. 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?	2. 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 3-9. 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 3-10. 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 3-11. The List of Simulation Factors

	Simulation factor		
Number of attributes	6	8	
Test length	12	30	
Data generating model	DINA	DINA-H_L	DINA-H_U
	DINO	DINO-H_L	DINO-H_U
Sample size	300	1000	3000
Estimation model	DINA	DINA-H_L	DINA-H_U
	DINO	DINO-H_L	DINO-H_U

Table 3-12. Q-Matrix for the Eighth Grade TIMSS 2007 Mathematics Test

Attribute Item	1	2	3	4	5	6	7	8	9	10	11	12	Total Number
1	1	0	0	0	0	0	0	0	0	0	0	0	1
2	1	0	1	0	0	0	0	0	0	0	0	0	2
3	0	0	1	0	1	0	0	0	0	0	0	0	2
4	0	1	0	0	1	1	0	0	0	0	0	0	3
5	1	1	1	0	0	0	0	0	0	0	0	0	3
6	1	1	1	0	0	0	0	0	0	0	0	0	3
7	1	0	1	1	0	1	0	0	0	0	0	0	4
8	1	0	1	0	0	0	0	0	0	0	0	0	2
9	0	0	1	1	1	0	0	0	0	0	0	0	3
10	0	0	0	1	1	0	0	0	0	0	0	0	2
11	0	0	0	1	1	0	0	0	0	0	0	0	2
12	0	0	0	0	0	0	1	0	1	0	0	0	2
13	0	0	1	0	0	0	1	0	0	0	1	0	3
14	0	0	0	0	0	0	1	0	1	1	0	0	3
15	0	0	1	0	0	0	0	0	0	0	0	1	2
16	0	0	0	0	0	0	0	0	0	0	0	1	1
17	0	0	0	0	0	0	0	0	0	0	0	1	1
18	1	1	1	0	0	0	0	0	0	0	0	0	3
19	0	0	0	1	1	0	0	0	0	0	0	0	2
20	0	0	1	1	1	1	0	0	0	0	0	0	4
21	0	0	0	0	0	0	1	1	1	0	0	0	3
22	0	0	0	0	0	0	0	1	1	0	1	0	3
23	0	0	1	1	1	0	0	0	0	0	0	0	3
24	0	0	1	0	0	0	0	0	0	0	0	1	2
25	0	0	1	0	0	0	0	0	0	0	0	1	2
26	1	0	0	0	0	0	0	0	0	0	0	1	2
27	0	0	0	1	0	0	0	0	0	0	0	1	2
28	0	0	1	0	0	0	1	0	1	0	1	0	4
29	1	0	0	1	0	0	0	0	0	0	0	0	2
Sum	9	4	15	9	8	3	5	2	5	1	3	7	

Note: The table was from Park et al. (2010) and Choi (2011).

Table 3-13. The List of Conditions for the DINA and DINA–H models

Condition	Number of attribute	Test length	Data generating model	Sample size	Estimation model
K6_J12_NA_300_NA	6	12	DINA	300	DINA
K6_J12_NA_300_HL					DINA-H_L
K6_J12_NA_300_HU					DINA-H_U
K6_J12_NA_1000_NA				1000	DINA
K6_J12_NA_1000_HL					DINA-H_L
K6_J12_NA_1000_HU					DINA-H_U
K6_J12_NA_3000_NA				3000	DINA
K6_J12_NA_3000_HL					DINA-H_L
K6_J12_NA_3000_HU					DINA-H_U
K6_J12_HL_300_NA			DINA-H_L	300	DINA
K6_J12_HL_300_HL					DINA-H_L
K6_J12_HL_300_HU					DINA-H_U
K6_J12_HL_1000_NA				1000	DINA
K6_J12_HL_1000_HL					DINA-H_L
K6_J12_HL_1000_HU					DINA-H_U
K6_J12_HL_3000_NA				3000	DINA
K6_J12_HL_3000_HL					DINA-H_L
K6_J12_HL_3000_HU					DINA-H_U
K6_J12_HU_300_NA			DINA-H_U	300	DINA
K6_J12_HU_300_HL					DINA-H_L
K6_J12_HU_300_HU					DINA-H_U
K6_J12_HU_1000_NA				1000	DINA
K6_J12_HU_1000_HL					DINA-H_L
K6_J12_HU_1000_HU					DINA-H_U
K6_J12_HU_3000_NA				3000	DINA
K6_J12_HU_3000_HL					DINA-H_L
K6_J12_HU_3000_HU					DINA-H_U
K6_J30_NA_300_NA	6	30	DINA	300	DINA
K6_J30_NA_300_HL					DINA-H_L
K6_J30_NA_300_HU					DINA-H_U
K6_J30_NA_1000_NA				1000	DINA
K6_J30_NA_1000_HL					DINA-H_L
K6_J30_NA_1000_HU					DINA-H_U
K6_J30_NA_3000_NA				3000	DINA
K6_J30_NA_3000_HL					DINA-H_L
K6_J30_NA_3000_HU					DINA-H_U
K6_J30_HL_300_NA			DINA-H_L	300	DINA
K6_J30_HL_300_HL					DINA-H_L
K6_J30_HL_300_HU					DINA-H_U

Table 3-13.––continued

Condition	Number of attribute	Test length	Data generating model	Sample size	Estimation model
K6_J30_HL_1000_NA					DINA
K6_J30_HL_1000_HL				1000	DINA-H_L
K6_J30_HL_1000_HU			DINA-H_L		DINA-H_U
K6_J30_HL_3000_NA					DINA
K6_J30_HL_3000_HL				3000	DINA-H_L
K6_J30_HL_3000_HU	6	30			DINA-H_U
K6_J30_HU_300_NA					DINA
K6_J30_HU_300_HL				300	DINA-H_L
K6_J30_HU_300_HU					DINA-H_U
K6_J30_HU_1000_NA					DINA
K6_J30_HU_1000_HL			DINA-H_U	1000	DINA-H_L
K6_J30_HU_1000_HU					DINA-H_U
K6_J30_HU_3000_NA					DINA
K6_J30_HU_3000_HL				3000	DINA-H_L
K6_J30_HU_3000_HU					DINA-H_U
K8_J12_NA_300_NA					DINA
K8_J12_NA_300_HL				300	DINA-H_L
K8_J12_NA_300_HU					DINA-H_U
K8_J12_NA_1000_NA					DINA
K8_J12_NA_1000_HL			DINA	1000	DINA-H_L
K8_J12_NA_1000_HU					DINA-H_U
K8_J12_NA_3000_NA					DINA
K8_J12_NA_3000_HL				3000	DINA-H_L
K8_J12_NA_3000_HU					DINA-H_U
K8_J12_HL_300_NA					DINA
K8_J12_HL_300_HL				300	DINA-H_L
K8_J12_HL_300_HU					DINA-H_U
K8_J12_HL_1000_NA					DINA
K8_J12_HL_1000_HL	8	12	DINA-H_L	1000	DINA-H_L
K8_J12_HL_1000_HU					DINA-H_U
K8_J12_HL_3000_NA					DINA
K8_J12_HL_3000_HL				3000	DINA-H_L
K8_J12_HL_3000_HU					DINA-H_U
K8_J12_HU_300_NA					DINA
K8_J12_HU_300_HL				300	DINA-H_L
K8_J12_HU_300_HU					DINA-H_U
K8_J12_HU_1000_NA					DINA
K8_J12_HU_1000_HL			DINA-H_U	1000	DINA-H_L
K8_J12_HU_1000_HU					DINA-H_U

Table 3-13.––continued

Condition	Number of attribute	Test length	Data generating model	Sample size	Estimation model
K8_J12_HU_3000_NA	8	12	DINA-H$_U$	3000	DINA
K8_J12_HU_3000_HL					DINA-H$_L$
K8_J12_HU_3000_HU					DINA-H$_U$
K8_J30_NA_300_NA	8	30	DINA	300	DINA
K8_J30_NA_300_HL					DINA-H$_L$
K8_J30_NA_300_HU					DINA-H$_U$
K8_J30_NA_1000_NA				1000	DINA
K8_J30_NA_1000_HL					DINA-H$_L$
K8_J30_NA_1000_HU					DINA-H$_U$
K8_J30_NA_3000_NA				3000	DINA
K8_J30_NA_3000_HL					DINA-H$_L$
K8_J30_NA_3000_HU					DINA-H$_U$
K8_J30_HL_300_NA			DINA-H$_L$	300	DINA
K8_J30_HL_300_HL					DINA-H$_L$
K8_J30_HL_300_HU					DINA-H$_U$
K8_J30_HL_1000_NA				1000	DINA
K8_J30_HL_1000_HL					DINA-H$_L$
K8_J30_HL_1000_HU					DINA-H$_U$
K8_J30_HL_3000_NA				3000	DINA
K8_J30_HL_3000_HL					DINA-H$_L$
K8_J30_HL_3000_HU					DINA-H$_U$
K8_J30_HU_300_NA			DINA-H$_U$	300	DINA
K8_J30_HU_300_HL					DINA-H$_L$
K8_J30_HU_300_HU					DINA-H$_U$
K8_J30_HU_1000_NA				1000	DINA
K8_J30_HU_1000_HL					DINA-H$_L$
K8_J30_HU_1000_HU					DINA-H$_U$
K8_J30_HU_3000_NA				3000	DINA
K8_J30_HU_3000_HL					DINA-H$_L$
K8_J30_HU_3000_HU					DINA-H$_U$

Table 3-14. The List of Conditions for the DINO and DINO–H models

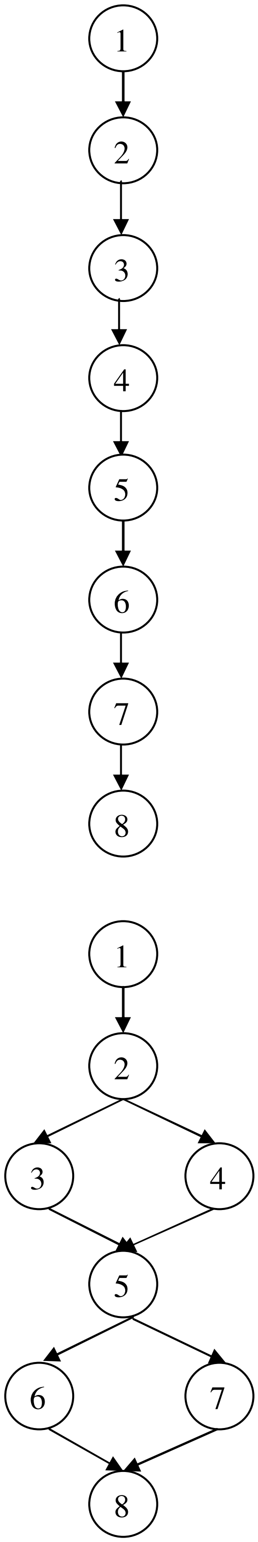
Condition	Number of attribute	Test length	Data generating model	Sample size	Estimation model
K6_J12_NO_300_NO	6	12	DINO	300	DINO
K6_J12_NO_300_HL					DINO-H$_{L}$
K6_J12_NO_300_HU					DINO-H$_{U}$
K6_J12_NO_1000_NO				1000	DINO
K6_J12_NO_1000_HL					DINO-H$_{L}$
K6_J12_NO_1000_HU					DINO-H$_{U}$
K6_J12_NO_3000_NO				3000	DINO
K6_J12_NO_3000_HL					DINO-H$_{L}$
K6_J12_NO_3000_HU					DINO-H$_{U}$
K6_J12_HL_300_NO			DINO-H$_{L}$	300	DINO
K6_J12_HL_300_HL					DINO-H$_{L}$
K6_J12_HL_300_HU					DINO-H$_{U}$
K6_J12_HL_1000_NO				1000	DINO
K6_J12_HL_1000_HL					DINO-H$_{L}$
K6_J12_HL_1000_HU					DINO-H$_{U}$
K6_J12_HL_3000_NO				3000	DINO
K6_J12_HL_3000_HL					DINO-H$_{L}$
K6_J12_HL_3000_HU					DINO-H$_{U}$
K6_J12_HU_300_NO			DINO-H$_{U}$	300	DINO
K6_J12_HU_300_HL					DINO-H$_{L}$
K6_J12_HU_300_HU					DINO-H$_{U}$
K6_J12_HU_1000_NO				1000	DINO
K6_J12_HU_1000_HL					DINO-H$_{L}$
K6_J12_HU_1000_HU					DINO-H$_{U}$
K6_J12_HU_3000_NO				3000	DINO
K6_J12_HU_3000_HL					DINO-H$_{L}$
K6_J12_HU_3000_HU					DINO-H$_{U}$
K6_J30_NO_300_NO	6	30	DINO	300	DINO
K6_J30_NO_300_HL					DINO-H$_{L}$
K6_J30_NO_300_HU					DINO-H$_{U}$
K6_J30_NO_1000_NO				1000	DINO
K6_J30_NO_1000_HL					DINO-H$_{L}$
K6_J30_NO_1000_HU					DINO-H$_{U}$
K6_J30_NO_3000_NO				3000	DINO
K6_J30_NO_3000_HL					DINO-H$_{L}$
K6_J30_NO_3000_HU					DINO-H$_{U}$
K6_J30_HL_300_NO			DINO-H$_{L}$	300	DINO
K6_J30_HL_300_HL					DINO-H$_{L}$
K6_J30_HL_300_HU					DINO-H$_{U}$

Table 3-14.––continued

Condition	Number of attribute	Test length	Data generating model	Sample size	Estimation model
K6_J30_HL_1000_NO					DINO
K6_J30_HL_1000_HL				1000	DINO-H_L
K6_J30_HL_1000_HU			DINO-H_L		DINO-H_U
K6_J30_HL_3000_NO					DINO
K6_J30_HL_3000_HL				3000	DINO-H_L
K6_J30_HL_3000_HU	6	30			DINO-H_U
K6_J30_HU_300_NO					DINO
K6_J30_HU_300_HL				300	DINO-H_L
K6_J30_HU_300_HU					DINO-H_U
K6_J30_HU_1000_NO					DINO
K6_J30_HU_1000_HL			DINO-H_U	1000	DINO-H_L
K6_J30_HU_1000_HU					DINO-H_U
K6_J30_HU_3000_NO					DINO
K6_J30_HU_3000_HL				3000	DINO-H_L
K6_J30_HU_3000_HU					DINO-H_U
K8_J12_NO_300_NO					DINO
K8_J12_NO_300_HL				300	DINO-H_L
K8_J12_NO_300_HU					DINO-H_U
K8_J12_NO_1000_NO					DINO
K8_J12_NO_1000_HL			DINO	1000	DINO-H_L
K8_J12_NO_1000_HU					DINO-H_U
K8_J12_NO_3000_NO					DINO
K8_J12_NO_3000_HL				3000	DINO-H_L
K8_J12_NO_3000_HU					DINO-H_U
K8_J12_HL_300_NO					DINO
K8_J12_HL_300_HL				300	DINO-H_L
K8_J12_HL_300_HU					DINO-H_U
K8_J12_HL_1000_NO					DINO
K8_J12_HL_1000_HL	8	12	DINO-H_L	1000	DINO-H_L
K8_J12_HL_1000_HU					DINO-H_U
K8_J12_HL_3000_NO					DINO
K8_J12_HL_3000_HL				3000	DINO-H_L
K8_J12_HL_3000_HU					DINO-H_U
K8_J12_HU_300_NO					DINO
K8_J12_HU_300_HL				300	DINO-H_L
K8_J12_HU_300_HU					DINO-H_U
K8_J12_HU_1000_NO					DINO
K8_J12_HU_1000_HL			DINO-H_U	1000	DINO-H_L
K8_J12_HU_1000_HU					DINO-H_U

Table 3-14.––continued

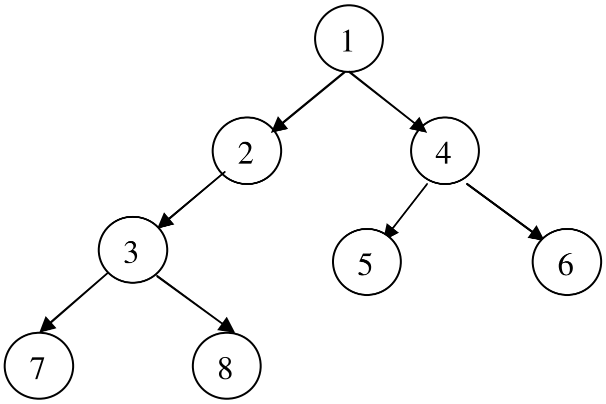
Condition	Number of attribute	Test length	Data generating model	Sample size	Estimation model
K8_J12_HU_3000_NO	8	12	DINO-H_U	3000	DINO
K8_J12_HU_3000_HL					DINO-H_L
K8_J12_HU_3000_HU					DINO-H_U
K8_J30_NO_300_NO	8	30	DINO	300	DINO
K8_J30_NO_300_HL					DINO-H_L
K8_J30_NO_300_HU					DINO-H_U
K8_J30_NO_1000_NO				1000	DINO
K8_J30_NO_1000_HL					DINO-H_L
K8_J30_NO_1000_HU					DINO-H_U
K8_J30_NO_3000_NO				3000	DINO
K8_J30_NO_3000_HL					DINO-H_L
K8_J30_NO_3000_HU					DINO-H_U
K8_J30_HL_300_NO			DINO-H_L	300	DINO
K8_J30_HL_300_HL					DINO-H_L
K8_J30_HL_300_HU					DINO-H_U
K8_J30_HL_1000_NO				1000	DINO
K8_J30_HL_1000_HL					DINO-H_L
K8_J30_HL_1000_HU					DINO-H_U
K8_J30_HL_3000_NO				3000	DINO
K8_J30_HL_3000_HL					DINO-H_L
K8_J30_HL_3000_HU					DINO-H_U
K8_J30_HU_300_NO			DINO-H_U	300	DINO
K8_J30_HU_300_HL					DINO-H_L
K8_J30_HU_300_HU					DINO-H_U
K8_J30_HU_1000_NO				1000	DINO
K8_J30_HU_1000_HL					DINO-H_L
K8_J30_HU_1000_HU					DINO-H_U
K8_J30_HU_3000_NO				3000	DINO
K8_J30_HU_3000_HL					DINO-H_L
K8_J30_HU_3000_HU					DINO-H_U



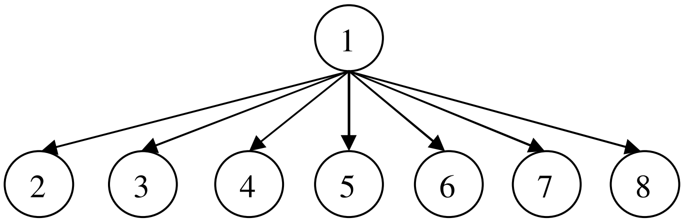
a. Linear

b. Convergent

Figure 3-1. Attribute hierarchies of eight attributes



c. Divergent



d. Unstructured

Figure 3-1.––continued

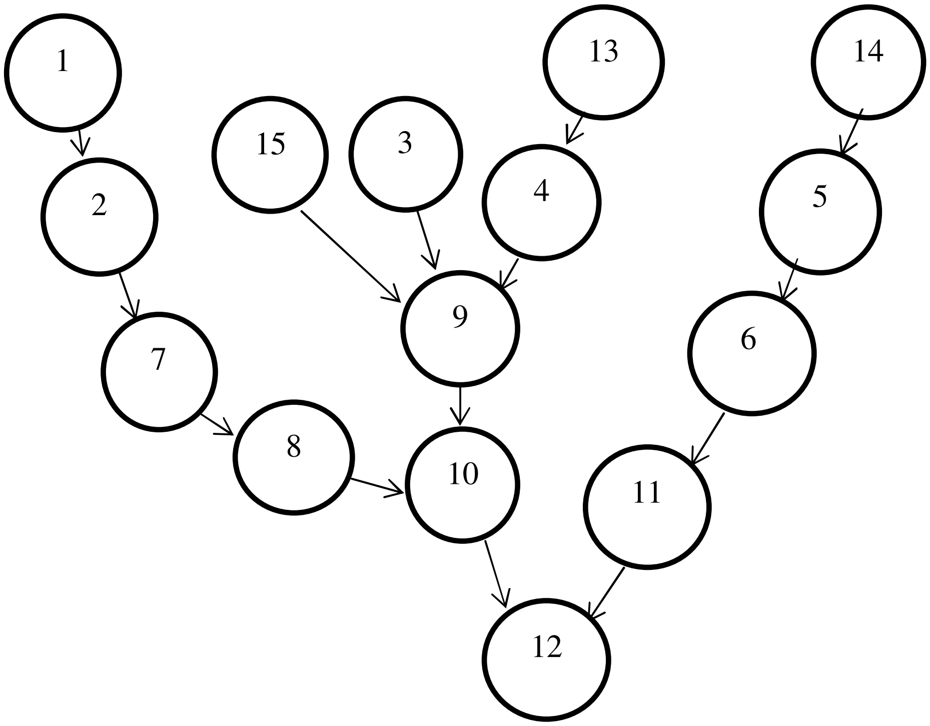


Figure 3-2. Hierarchical relationship among the attributes for the eighth grade TIMSS 2003 mathematics test

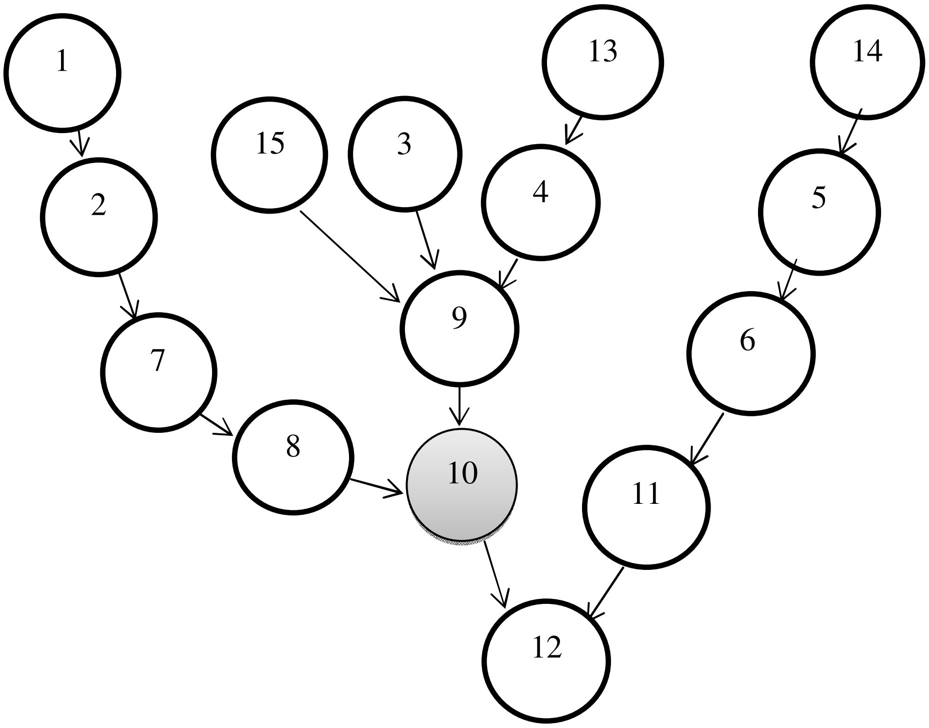


Figure 3-3. Hierarchical relationship among the attributes for booklet 1

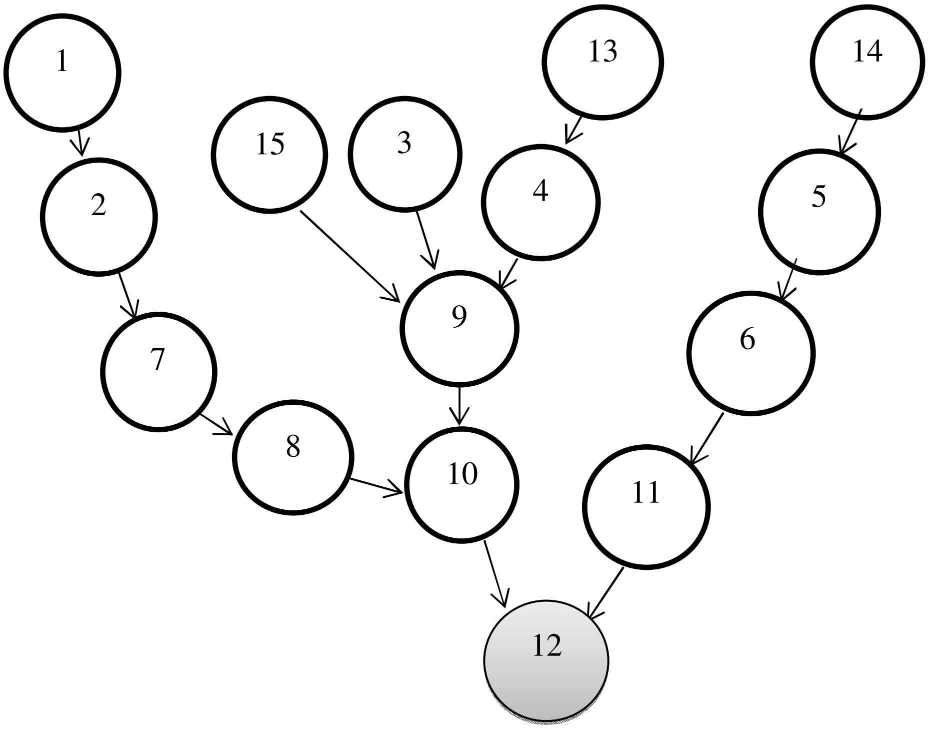


Figure 3-4. Hierarchical relationships among the attributes for booklet 2

CHAPTER IV

RESULTS

This chapter describes results from the study. It contains two main sections. The first section describes the results from the real data analysis. The second section presents the results from the simulation study.

Results of Real Data Analysis

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.

DINA and DINA-H

The following paragraphs provide the results of the model fit, item fit, and item parameter estimates from the real data analysis 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 4-1, 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 17), respectively. The AIC result for the DINA model is $-2\ln(Likelihood)+2p=(-2)\times(-11410)+2(2\times 29+2^{14}-1)=55702$. 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 4-2 to 4-5, 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 (see the highlighted cells in the tables), 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 4-2). 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 4-3). For TIMSS B2, about 20% of the items have higher δ values under the DINA-H model for the U.S. sample, and about 20% for the benchmark sample (see Table 4-4). 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 4-5). 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 4-2 to 4-5. 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 4-2). 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 4-3). 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 4-4). 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 4-5).

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 4-6. 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 4-7 to 4-10, 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.

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 4-1, the values of differences in Table 4-11 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 to73, using the same 0.001 criteria.

Item Fit

Tables 4-12 to 4-15 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 4-12). 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 4-13). 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 4-14). 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 4-15). 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 4-12 to 4-15, 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 4-12). 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 4-13). 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 4-14). 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 4-15). 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 4-16. 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 4-17 to 4-20 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.

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 4-21. 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 4-1 to Table 4-11, 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 4-22 to 4-25. 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 4-22). 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 4-23). 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 4-24). 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 4-25). 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.

Results of the Simulation Study

This section describes the detailed results of the fit indices and summary statistics for varying conditions under the DINA(O) and DINA(O)-H models from the simulation study. The conditions in the tables refer to the data generating models and the estimation models, respectively. For example, the condition of HL_NA refers to the condition of using the DINA-H-linear data generating model and the conventional DINA estimation model. The same logic of notation naming is used throughout the whole paper.

DINA and DINA-H

This part shows the results of the model fit and summary statistics from the DINA and DINA-H models in the simulation analysis. It includes: The main effects of model consistency, numbers of attributes, test lengths, and sample sizes; the interaction effects of test length by attribute, sample size by attribute, and sample size by test length; and the three-way interaction effect of sample size by test length by attribute.

Main Effect of Model Consistency

The following paragraphs present the results of the fit indices and summary statistics for the main effect of model consistency testing varying estimation models, in terms of their being consistent or inconsistent with the specifications on the skill hierarchies, under the DINA and DINA-H models from the simulation study. The values used for comparisons were computed by averaging over other conditions of sample sizes, test lengths, and numbers of attributes. The gray highlighted values mean that the conditions show the best results (i.e., the least MAIC and MBIC, and the least ASB and AMSE of guessing and slip parameter estimates).

Fit Indices

For the model fit indices MAIC and MBIC, all the conditions that used the estimation models consistent with their data generating models show better model fit results (i.e., smaller MAICs and smaller MBICs) than the conditions that used the estimation models inconsistent with their data generating models, as shown in Table 4-

26. The results confirm the assumption that a better model fit can be obtained from the calibration results when the relationships among attributes specified in the data generating model are consistent with those in the estimation model. The results also confirm that when skills are ordered hierarchically, the proposed DINA-H models perform better.

When the data generating model is DINA-H_L, using the DINA-H_U model to calibrate has better model fit results than using the conventional DINA model (see Table 4-26). When the true model is DINA-H_U, using the DINA estimation model produces smaller MAIC and MBIC values than using the DINA-H_L model. If data are generated via the conventional DINA model, the DINA-H_U model shows better calibration results than the DINA-H_L model. The results are due to different levels of dependency among the hierarchically ordered skills. The DINA-H_L model has the highest dependency of skills among the three hierarchical models, but the convention DINA model does not assume any dependency among skills.

Summary Statistics

The results of the summary statistics for the main effect of model consistency show that using the consistent estimation model with the data generating model obtained better item parameter recovery (i.e., the smaller ASB and AMSE of guessing and slip parameter estimates) when compared to the results of using the inconsistent model, as shown in Table 4-27. The only two exceptions are that the smaller AVAR of the guessing parameter estimates appears when using the DINA model to calibrate other DINA-H model data. Similar to the results from the evaluation of the fit indices, the results of the summary statistics generally confirm the assumption that better item parameter recovery can be obtained from the calibration results when the relationships among attributes specified in the data generating model are consistent with those in the estimation model, and also confirm that the proposed DINA-H models should be used to calibrate items when cognitive skills are ordered hierarchically.

Main Effect of Numbers of Attributes

Since the main effect of model consistency was confirmed and supported, the following main effects of other study variables and their interaction effects were examined by comparing only the results of using the consistent models in generating and estimating data. Note also that magnitudes of the model fit indices (i.e., MAIC and MBIC) depend heavily on the numbers of items, attributes, and examinees (see Equations 22 and 23), and thus the comparison of the fit indices across various conditions does not provide any meaningful information. The MAIC and MBIC values are not compared for the subsequent main and interaction effects. This section presents the results of the summary statistics for the main effect of numbers of attributes under the DINA and DINA-H models from the simulation study.

The results of the summary statistics for the main effect of numbers of attributes show that the conditions of eight attributes obtain better item parameter recovery with smaller ASB, AVAR, and AMSE when compared to the results of six attributes (see Table 4-28). The only exception is that AVAR of slip parameter estimates is smaller for the condition of six attributes. For both conditions of six and eight attributes, the guessing parameter estimates show better recover than the slip parameter estimates. Additionally, increasing the number of attribute from six to eight improves the recovery of guessing parameter estimates, with the amount of ASB, AVAR, and AMSE decreasing. Although slip parameter estimates show the same pattern, the amount of the improvement of AMSE from K=6 to K=8 is only about half of that for the guessing parameter estimates. For both the guessing and slip parameter estimates and for both conditions of numbers of attributes, the condition of using the DINA-H_L model shows the more accurate calibration results than the DINA and DINA-H_U models.

Main Effect of Test Lengths

The results of the summary statistics for the main effect of test lengths show that the condition of 12 items in a test has smaller ASB and AMSE for both the guessing and slip parameter estimates when compared to the results of 30 items, as shown in Table 4-29. The smaller AVAR values appear for the condition of 30 items for both guessing and slip parameter estimates. In general, the results of the summary statistics confirm that tests that consisted of 12 items obtained better item parameter recovery than those that used 30 items although there is larger variance when using 30 items in a test. When reducing numbers of items from 30 to 12 items, more reduction in estimation error (i.e., AMSE and ASB) is found for the guessing parameter estimates than for the slip parameter estimates. For both the conditions of 12 and 30 items, guessing parameter estimates are more accurate than slip parameter estimates. For both the conditions of 12 and 30 items, the results for the DINA-H_L model show better item parameter recovery than the DINA and DINA-H_U models.

Main Effect of Sample Sizes

The results of the summary statistics for the main effect of sample sizes consistently show that using N=3000 has the least ASB, AVAR, and AMSE for both the slip and guessing parameter estimates, compared to the results of using N=300 or 1000 (see Table 4-30). It means that using the larger sample size would result in better item parameter recovery. The results of the summary statistics support the assumption that better item parameter recovery can be obtained from the calibration results when large sample sizes are applied. The magnitude of AVAR consistently decreases when the number of sample sizes increase for both the guessing and slip parameter estimates. For all conditions of sample sizes, the DINA-H_L condition shows better item parameter recovery results of smaller AMSE than the DINA and DINA-H_U models. The only exception is that the smaller ASB appears when N=1000 and 3000 for the guessing parameter estimates under the DINA model. The guessing parameter estimates show more accurate item parameter recovery results than the slip parameter estimates for all conditions of sample sizes.

Interaction Effect of Test Lengths by Numbers of Attributes

The results of the summary statistics for the interaction effect of different test lengths by varying numbers of attributes show that the condition of 30 items with eight attributes performed the best (see Tables 4-31 and 4-32) although the results of the main effects show that the condition of 12 items and the condition of 8 attributes tend to perform better than other conditions (see Tables 4-28 and 4-29). Using 30 items measuring eight attributes show the smallest mean ASB, AVAR, and AMSE for both guessing and slip parameter estimates, except for the AVAR for slip. As shown in Figure 4-1, the interaction effect between test lengths and numbers of attributes can be observed under the DINA(-H) model.

Although the interaction effect is found for 30 items with eight attributes achieving better, the item parameter recovery results are slightly different between the guessing and slip parameters. For the guessing parameter estimates, twelve items with eight attributes performed better in item parameter recovery than 12 items with six attributes and 30 items with six attributes, except for the smallest ASB for the condition of 12 items with six attributes. For the slip parameter estimates, twelve items with six attributes performed better in item parameter recovery than 12 items with eight attributes and 30 items with six attributes.

Comparing to the NA_NA and HU_HU conditions, the condition of HL_HL obtains the best item parameter recovery results, except for the larger AMSE and ASB under the condition of 30 items with six attributes for both guessing and slip. Both AMSE and ASB increase when the number of items increases under the condition of six attributes, whereas the error decreases when the number of items increases under the condition of eight attributes for both the guessing and slip parameter estimates. Except for the AMSE for the condition of J=12 and K=8 and AVAR for the condition of J=30 and K=8, all other summary statistics results show that the guessing parameter recovery is better than the slip parameter recovery.

Interaction Effect of Sample Sizes by Numbers of Attributes

Tables 4-33 and 4-34 and Figure 4-2 present the ASB, AVAR, and AMSE of guessing and slip parameter estimates for the interaction effect of N by K for the DINA and DINA-H models. The results of summary statistics for the guessing and slip parameter estimates are different. The smallest mean AMSE was shown in the condition of N=3000 and K=8 for the guessing parameter estimates while the smallest mean ASB appears in the condition of N=300 and K=6. For the slip parameter estimates, the best item parameter recovery was observed in the condition of N=3000 and K=6. For the main effect, the condition of eight attributes and the condition of N=3000 perform better (see Tables 4-28 and 4-30). In terms of interaction effect, the condition of N=3000 and K=8 provides the best results for the guessing parameter estimates, and the condition of N=3000 and K=6 provides the best results for the slip parameter estimates (see Tables 4-33 and 4-34).

In terms of AMSE for the guessing parameter estimates, the condition of N=3000 and K=8 performs the best, next the condition of N=3000 and K=6 followed by the condition of N=1000 and K=8. For the slip parameter estimates, the top three conditions are the condition of N=3000 and K=6, the condition of N=3000 and K=8, and the condition of N=1000 and K=8. The item parameter recovery results are better for the guessing parameters than for the slip parameter estimates. The only exception is that the smaller AVAR appears under the condition of N=1000 and K=6 for the slip parameter estimates rather than guessing.

Comparing the conditions of different estimation models, for the guessing parameter estimates the DINA-H_L model shows better item parameter recovery of the smaller AMSE than the DINA and DINA-H_U models, although slightly larger ASB is shown for the conditions of N=1000 and K=6, N=1000 and K=8, N=3000 and K=6, and N=3000 and K=8. The NA_NA condition is better than the HU_HU condition for varying sample sizes when K=6, but this is not so when K=8. For the slip parameter estimates, the DINA-H_L model performs the best, and the conventional DINA model obtains the poorest item parameter recovery results. Consistently across all conditions, the guessing parameter results are better than the slip parameter results.

Interaction Effect of Sample Sizes by Test Lengths

Tables 4-35 and 4-36 list the ASB, AVAR, and AMSE of guessing and slip parameter estimates for the interaction effect of sample sizes by test lengths for the DINA and DINA-H models. For both parameter estimates, the condition of N=3000 and J=30 obtains the best item parameter recovery results and the next better results are shown under the condition of N=3000 and J=12, while the condition of N=1000 and J=30 obtains the worst item parameter recovery results. But, for the main effects, the condition of N=3000 and the condition of J=12 produce the best item parameter recovery results (see Tables 4-29 and 4-30).

The AMSE, ASB, and AVAR decrease when sample size increases under the condition of J=12 for both guessing and slip parameter estimates (see Tables 4-35 and 4-36). When the sample size is 300 or 3000, the conditions with more items produce more accurate item parameter recovery results for both guessing and slip parameter estimates (see Figure 4-3). The guessing parameters are recovered better than the slip parameters with smaller error across all conditions.

The DINA-H_L model shows the best item parameter recovery results under all conditions for both the guessing and slip parameter estimates. In general, for the guessing parameter estimates, the conventional DINA model recovers item parameter better (i.e., smaller AMSE) than the DINA-H_U model with some exceptions for ASB and AVAR (see Table 4-35). For the slip parameter estimates, the DINA-H_U model shows better item parameter recovery than the conventional DINA model under all conditions (see Table 4-36).

Interaction Effect of N by J by K

Tables 4-37 and 4-38 list the ASB, AVAR, and AMSE of guessing and slip parameter estimates for the N by J by K three-way interaction effect under the DINA and DINA-H models. These three-way interaction tables also show a comprehensive picture of which combinations of various conditions provide the best results. The values shown in the tables are averaging over the values with consistent data-generating and estimation models for each N by J by K condition.

For the main effects of both the guessing and slip parameter estimates, best results were obtained from the conditions of N=3000, J=12, and K=8 (see Tables 4-28 to 4-30). For the three-way interaction effect for the guessing parameter estimates, across all samples size conditions, the condition of J=30 and K=8 obtained the best item parameter recovery with the smallest ASB, AVAR, and AMSE appearing under the condition of N=3000 by J=30 by K=8, followed by the condition of N=3000 by J=30 by K=6, and then the condition of N=1000 by J=30 by K=8 (see Table 4-37 and Figures 4-4 and 4-5). For the slip parameter estimates, the condition of N=3000 by J=30 by K=6 obtained the best item parameter recovery with the smallest ASB, AVAR, and AMSE. The next better results are shown under the condition of N=3000 by J=30 by K=8 and then the condition of N=3000 by J=12 by K=6 (see Table 4-38). Generally speaking, the guessing parameter estimates show better item parameter recovery results than the slip parameter estimates, with a few exceptions of smaller ASB and AVAR appearing for the slip parameter estimates, but not the AMSE values.

In Appendix A, Tables A1 to A9 are listed to show the comprehensive results of model fit and item parameter recoveries across all conditions for the DINA(-H) model. In general, the results confirmed what is found under the main effect of model consistency. In these tables, the rows of means show the values of averaging over nine different data generating and estimation conditions.

DINO and DINO-H

This section describes the detailed results of the fit indices and summary statistics for varying conditions under the DINO and DINO-H models from the simulation study. Just as with the DINA and DINA-H models, this section includes the main effects of model consistency, numbers of attributes, test lengths, and sample sizes; the interaction effects of test length by attribute, sample size by attribute, and sample size by test length; and the three-way interaction effect of sample size by test length by attribute.

Main Effect of Model Consistency

This part discusses the results of the fit indices and summary statistics for the main effect of model consistency when using varying estimation models consistent or inconsistent with the specifications of the skill hierarchies under the DINO and DINO-H models from the simulation study.

Fit Indices

For the model fit indices MAIC and MBIC, all conditions, except one, that used the estimation models consistent with their data generating models show better model fit results than the conditions that used the estimation models inconsistent with their data generating models (see Table 4-39). The only exception is the smaller MBIC that occurs when using the DINO-H_L model to estimate data generated via the DINO-H_U model. Generally speaking, the results confirm the main effect of model consistency that better model fit can be obtained from the calibration results when the relationships among attributes specified in the data generating model are consistent with those in the estimation model. Similar to what was found under the DINA-H model, the results also confirm that when skills are ordered hierarchically, the proposed DINO-H model performs better and is preferred to be applied, in terms of the model fit.

Due to the different levels of dependency among the hierarchically ordered skills, using the conventional DINO model to estimate data generated via the DINO-H_L model produces poorer model fit results than those estimated by the DINO-H_U model. However, the results do not show any consistent pattern for the data generated via the conventional DINO or the DINO-H_U models.

Summary Statistics

The results of the summary statistics for the main effect of model consistency show that using the consistent estimation model with the data generating model produces more accurate results of guessing and slip parameter estimates with the least AMSE and ASB, compared to the results of using the inconsistent model (see Table 4-40). The two exceptions are that the smaller AVAR values appear when using the DINO model to estimate the DINO-H_L model data for the guessing parameter estimates and when using the DINO-H_L model to estimate the DINO-H_U model data for the slip parameter estimates. Similar to the results from the evaluation of the fit indices, when the relationships among attributes specified in the data generating model are consistent with those in the estimation model, better item parameter recovery results can be obtained. The results of the summary statistics under the DINO-H models are consistent to what were found under the DINA-H models. The proposed DINO-H models should be applied to calibrate items when cognitive skills are in a certain hierarchical structure.

Main Effect of Numbers of Attributes

As with the DINA(-H) model, the results of model fit are not comparable here because the values of MAICs and MBICs are more sensitive to the number of attributes, test lengths, and sample sizes. Hence, only the summary statistics are evaluated for the following main and interaction effects.

The results of the summary statistics for the main effect of varying numbers of attributes show that measuring six attributes in a Q-matrix generates better item parameter recovery when compared to the results of measuring eight attributes for the guessing parameter estimates, as shown in Table 4-41. For the slip parameter estimates under the DINO(-H) model, the condition of eight attributes performs better in terms of item parameter recovery than the condition of six attributes. The results of the summary statistics show that the main effect of numbers of attributes is different for the guessing and slip parameter estimates. The reduction of the estimation error in the guessing parameter estimates when decreasing the numbers of attributes from eight to six is much larger than the improvement in the slip parameter estimates when increasing the numbers of attributes from six to eight.

For both conditions of six and eight attributes, the condition of HL_HL shows the least amount of error for both guessing and slip parameter estimates. The condition of NO_NO shows more accurate estimation results with smaller AMSE and AVAR than the condition of HU_HU for the guessing parameter estimates, although the smaller ASB appears under the condition of HU_HU. In contrast, for the slip parameter estimates the condition of HU_HU produces better item parameter recovery results than the condition of NO_NO, with only one exception of smaller ASB appearing for the condition of NO_NO when K=6.

Main Effect of Test Lengths

The results of the summary statistics for the main effect of varying test lengths show that the condition of 30 items consistently shows better item parameter recovery than the condition of 12 items (see Table 4-42). The condition of 30 items produces smaller AMSE, ASB, and AVAR than the condition of 12 items for both the guessing and slip parameter estimates. The improvement of estimation accuracy is better for the guessing parameter estimates than for the slip parameter estimates when increasing number of items from 12 to 30.

The slip parameter estimates show better item parameter recovery results than the guessing parameter estimates. For both parameter estimates under both the conditions of 12 and 30 items, the DINO_H_L model consistently performs better than the DINO or DINO-H_U models. The only exception is that the smaller ASB appears under the conventional DINO model rather than under the DINO_H_L model for the slip parameter estimates when J=30. For the guessing parameter estimates, the NO_NO condition shows better item parameter recovery than the condition of HU_HU for both conditions of 12 and 30 items although the larger ASB is found under the NO_NO condition when J=12. For the slip parameter estimates, the HU_HU condition is better when J=12 and the NO_NO condition is better when J=30.

Main Effect of Sample Sizes

All the results of the summary statistics for the main effect of varying numbers of sample sizes show that the condition of the largest sample sizes (i.e., N=3000) produced the smallest ASBs, AVARs, and AMSEs for both the guessing and slip parameter estimates, as shown in Table 4-43. The values of ASBs, AVARs, and AMSEs for both the guessing and slip parameter estimates decrease when the number of sample sizes increases. The slip parameter estimates consistently perform better than the guessing parameter estimates for all three conditions of sample sizes. In terms of the mean values of AMSE, ASB, and AVAR, the reduction of the estimation errors when increasing sample sizes from 300 to 1000 and from 1000 to 3000 for the guessing parameter estimates is much larger than the improvement for the slip parameter estimates. Across various sample sizes, the condition of DINO-H_L model shows the less estimation error than the DINO and DINO-H_U models.

Interaction Effect of Test Lengths by Numbers of Attributes

The results of the summary statistics for the interaction effect of test lengths with varying numbers of attributes show that using more items measuring fewer attributes obtained better item parameter recovery, compared to the results of using fewer items measuring more attributes. As shown in Tables 4-44 and 4-45, the condition of 30 items with 6 attributes produces the smallest ASBs, AVAR, and AMSEs for both the guessing and slip parameter estimates. The next best condition is 12 items with six attributes for the guessing parameter estimates, whereas it is the condition of 30 items with eight attributes for the slip parameter estimates. The main effect results show that the conditions of 30 items and six attributes are preferred for estimating the guessing parameters while the conditions of 30 items and eight attributes are preferred for estimating the slip parameters (see Tables 4-41 and 4-42). The interaction effect can be observed in Figure 4-6 showing that the conditions of more items with fewer attributes produce better item parameter recovery than the conditions of fewer items with more attributes.

Generally speaking, the slip parameter estimates show better item parameter recovery results than the guessing parameter estimates, except for the smaller ASB for the guessing parameter estimates when J=12 and K=6. For the guessing parameter estimates, the condition of DINO-H_L model produced less estimation error than the conditions of DINO and DINO-H_U models across all combinations of conditions of different test lengths by varying numbers of attributes. For the slip parameter estimates, the condition of DINO-H_L model is better than the DINO and DINO-H_U models under the conditions of J=12 with K=6, J=12 with K=8, and J=30 with K=8. However, the DINO model shows better slip parameter recovery results under the condition of J=30 with K=6.

Interaction Effect of Sample Sizes by Numbers of Attributes

The ASBs, AVARs, and AMSEs of guessing and slip parameter estimates for the interaction effect of N by K for the DINO and DINO-H models are presented in Tables 4-46 and 4-47, respectively. The results of summary statistics for the guessing parameter estimates show that the condition of N=3000 with K=6 obtained the best item parameter recovery results with the least mean AMSE and AVAR, although the condition of N=1000 with K=6 produced the least mean ASB. The difference of mean ASB between the conditions of N=1000 with K=6 and N=3000 with K=6 is relatively small (i.e., 0.00002). The results of summary statistics for the slip parameter estimates show that the condition of N=3000 with K=8 obtained the least mean AVAR and AMSE, and the condition of N=1000 with K=8 produced the least mean ASB. The difference of mean ASB between the conditions of N=1000 with K=8 and N=3000 with K=6 is 0.000014.

For the main effect, the condition of six attributes performs better than eight attributes in estimating the guessing parameters, but the condition of eight attributes provides better results for the slip parameter estimates (see Table 4-41), and the condition of N=3000 provides the best results for both the guessing and slip parameter estimates (see Table 4-43). Generally speaking, the interaction effect shows that the condition of larger sample sizes with fewer attributes in the Q-matrix show better item parameter recovery for the guessing parameter estimates, but the condition of larger sample size with more attributes is better for the slip parameter estimates, as shown in Figure 4-7.

The slip parameters are, in general, recovered better than the guessing parameters, except for the ASB values under the conditions of N=300 with K=6 and N=1000 with K=6. In terms of AMSE for the guessing parameter estimates, the DINO-H_L model shows better estimation accuracy than the DINO and DINO-H_U models. The only exception is that the smaller AMSE occurs when N=1000 with K=6. For the slip parameter estimates, the DINO-H_L model consistently shows better estimation accuracy than the DINO and DINO-H_U models across all different conditions of sample sizes with varying numbers of attributes.

Interaction Effect of Sample Sizes by Test Lengths

Tables 4-48 and 4-49 list the ASBs, AVARs, and AMSEs of guessing and slip parameter estimates for the interaction effect of sample sizes by test lengths for the DINO and DINO-H models, respectively. For both guessing and slip parameter estimates, the conditions of N=3000 with J=30 produced the least mean ASBs, AVARs, and AMSEs across all conditions. In terms of AMSE, the next best conditions are N=3000 with J=12 and N=1000 with J=30 for both guessing and slip parameter estimates. It shows that having larger sample sizes with more items in a test can obtain more accurate item parameter recovery results. The role of sample size is more important than the role of number of items in reducing estimation error. The values of ASBs, AVARs, and AMSEs increase when sample sizes decrease. The values of ASBs, AVARs, and AMSEs increase when test lengths decrease, except for the condition of N=300 for the guessing parameter estimates. The trend of the main effects of sample sizes and test lengths is clearer than the interaction effect between the two variables (see Figure 4-8), as the main effects show that the condition of N=3000 and the condition of J=30 are preferred (see Tables 4-42 and 4-43).

The slip parameter estimates show better item parameter recovery results than the guessing parameter estimates across all conditions of different sample sizes by various test lengths. The DINO-H_L model produces more accurate estimation results than using the DINO and DINO-HU models.

Interaction Effect of N by J by K

Tables 4-50 and 4-51 present the results of summary statistics for the N by J by K three-way interaction effect for the guessing and slip parameter estimates under the DINO and DINO-H models, respectively. The values were averaging over the three values with consistent data-generating and estimation models for each N by J by K condition.

For the main effects under the DINO(-H) model, the conditions of N=3000, J=30, and K=6 show better item parameter recovery for the guessing parameter estimates, and the conditions of N=3000, J=30, and K=8 show better item parameter recovery for the slip parameter estimates (see Tables 4-41 to 4-43). For the three-way interaction effect for the guessing parameter estimates, the condition of N=3000 by J=30 by K=6 produced the best item parameter recovery results with the least mean AMSE, ASB, and AVAR (see Table 4-50). For the slip parameter estimates, the condition of N=3000 by J=30 by K=8 shows the best item parameter recovery results with the least mean AMSE and AVAR, while the condition of N=3000 by J=30 by K=6 has the least mean ASB (see Table 4-51). With smaller sample sizes, the values of summary statistics are more sensitive to the numbers of attributes and test lengths. When sample size is smaller, test length is shorter and the numbers of attributes are more, item parameter recovery is poorer with larger ASB and AMSE for the slip parameter estimates (see Figures 4-9 and 4-10). In general, the slip parameter estimates show more accurate estimation results than the guessing parameter estimates, except for the conditions of N=1000 by J=12 by K=6 and N=300 by J=12 by K=6 in which the guessing parameter estimates performed better.

Tables A10 to A18 in Appendix A list the more comprehensive results of item parameter recoveries across all conditions for additional information. As with the DINA(-H) model, the results generally confirmed what was found under the main effect of model consistency.

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

This section compares the results of the fit indices and summary statistics between the DINA(-H) and DINO(-H) models from the simulation study. The comparisons include the main effects of model consistency, numbers of attributes, test lengths, and sample sizes.

Main Effect of Model Consistency

This part contrasts the results of the fit indices and summary statistics between the DINA(-H) and DINO(-H) models from the simulation study for the main effect of using varying estimation models consistent or inconsistent with the specifications on the skill hierarchies.

Fit Indices

Generally speaking, for the main effect of model consistency, both the DINA(-H) and DINO(-H) models confirm that using the estimation models consistent with the data generating models show better model fit results than the conditions using the estimation models inconsistent with their data generating models, as shown in Tables 4-26 and 4-39. The only exception for both the DINA(-H) and DINO(-H) models is that the HL_HL condition did not obtain better item fit results when compared to the HL_NA(O) and HL_HU conditions. The DINA, DINA-H_U, DINO, and DINO-H_U models all confirm that using the estimation models consistent with their data generating models show better item fit results.

Table 4-52 shows the differences of model fit between the DINA(-H) and DINO(-H) models. The differences found in MAIC and MBIC between the DINA(-H) and DINO(-H) models were computed by subtracting the DINA(-H) condition values from those of the DINO(-H). The positive values in the differences of MAIC and MBIC, thus, indicate that the DINA(-H) model performs better than the DINO(-H) model (see the highlighted cells). All the differences of the model fit indices show that the DINO(-H) model has better model fit with smaller MAICs and MBICs than those in the DINA(-H) model. In terms of model fit, the DINO(-H) model performs better.

Summary Statistics

All the results of summary statistics for the DINA(-H) and DINO(-H) models' main effect of model consistency show that using the estimation model consistent with the data generating model obtains better item parameter recovery (see Tables 4-27 and 4-40).

The differences of ASB, AVAR, and AMSE for the guessing and slip parameters between the DINA(-H) and DINO(-H) models, as shown in Table 4-53, were also computed from using the values in the DINO(-H) conditions minus those in the DINA(-H) conditions. The positive values mean that the values in the DINA(-H) conditions are smaller than the values in the DINO(-H) conditions, indicating that the DINA(-H) model performs better than the DINO(-H) model. Comparing the summary statistics between the DINA(-H) model and the DINO(-H) model, the DINA(-H) model outperforms the DINO(-H) model with smaller AMSEs under two-thirds of the conditions for the guessing parameter estimates. The DINO(-H) model performs better than the DINA(-H) model with smaller AMSE values under all conditions for the slip parameter estimates. This finding generally reconfirms that the DINO(-H) model obtains better item parameter recovery than the DINA(-H) model, especially when the data generating model is consistent with the estimation model.

Main Effect of Numbers of Attributes

The condition of eight attributes produces more accurate item parameter recovery results for both the guessing and slip parameter estimates under the DINA(-H) model and for the slip parameter estimates under the DINO(-H) model, as shown in Tables 4-28 and 4-41, while the condition of six attribute is deemed to be a better condition in terms of the guessing parameter estimation. Comparing the differences of AMSEs between the DINA(-H) and DINO(-H) models, the DINO(-H) model displays better item parameter recovery than the DINA(-H) model in estimating the slip parameters for both the condition of six and eight attributes (See Table 4-54). The DINA(-H) model outperforms the DINO(-H) model for the guessing parameter estimates when K=8, in terms of the magnitude of AMSE.

Main Effect of Test Lengths

The results of the summary statistics for the main effect of test lengths show that using 12 items in a test obtains better item parameter recovery under the DINA(-H) model and using 30 items in a test obtains better item parameter recovery under the DINO(-H) model (see Tables 4-29 and 4-42).

Generally speaking, the DINO model outperforms the DINA model in estimating slip parameters for both conditions of J=12 and J=30, whereas the DINA model outperforms the DINO model in obtaining better guessing parameter recovery with smaller AMSE when J=12 (see Table 4-55).

Main Effect of Sample Sizes

All the results of the summary statistics for the main effect of sample sizes under both the DINA(-H) and DINO(-H) models show that using the larger sample size (i.e., N=3000) would result in better item parameter recovery for both the guessing and slip parameter estimates (see Tables 4-30 and 4-43).

Some mixed findings appear in the differences of item parameter recovery between the DINA(-H) and DINO(-H) models. For the guessing parameter recovery, the DINA model performs better in the conditions of sample sizes of 300 and 3000, whereas the DINO model performs better in the condition of sample size of 1000 (see Table 4-56). The DINO model outperforms the DINA model in obtaining the better slip parameter recovery, regardless of the sample sizes.

The comparisons of the summary statistics for all interaction effects between the DINA(-H) and DINO(-H) models are shown in Tables A19 to A26 in Appendix.

Table 4-1. 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 4-2. Results of Item Fit Index- δ for TIMSS B1 Data under the DINA and DINA-H Models

	U.S. Sample			Benchmark Sample			Difference (Benchmark -U.S.)	
Item	DINA	DINA-H	DINA-DINA-H	DINA	DINA-H	DINA-DINA-H	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 4-3. Results of Item Fit Index-IDI for TIMSS B1 Data under the DINA and DINA-H Models

	U.S. Sample			Benchmark Sample			Difference (Benchmark -U.S.)	
Item	DINA	DINA-H	DINA-DINA-H	DINA	DINA-H	DINA-DINA-H	DINA	DINA-H
1	0.705	0.657	-0.047	0.566	0.58735	0.021	-0.139	-0.070
2	-4.730	-5.054	-0.324	-14.689	-6.54564	8.143	-9.959	-1.491
3	0.767	0.714	-0.053	0.799	0.63321	-0.166	0.033	-0.080
4	0.323	0.273	-0.051	0.251	0.24241	-0.008	-0.073	-0.031
5	0.459	0.353	-0.106	0.483	0.34513	-0.138	0.024	-0.008
6	-21.590	-14.246	7.344	-8.912	-4.26875	4.643	12.678	9.977
7	0.485	0.437	-0.048	0.805	0.43607	-0.369	0.320	-0.001
8	-1.590	-1.551	0.039	-0.458	-0.99634	-0.538	1.132	0.555
9	0.480	0.482	0.001	0.428	0.38097	-0.047	-0.052	-0.101
10	0.528	0.507	-0.020	0.167	0.39345	0.226	-0.360	-0.114
11	0.587	0.546	-0.041	0.354	0.39315	0.040	-0.234	-0.153
12	0.599	0.583	-0.016	0.375	0.51674	0.142	-0.224	-0.067
13	0.723	0.559	-0.164	0.667	0.48893	-0.178	-0.056	-0.070
14	0.647	0.549	-0.098	0.744	0.53181	-0.212	0.097	-0.017
15	0.710	0.702	-0.007	0.909	0.63998	-0.269	0.199	-0.062
16	0.998	0.694	-0.304	1.000	0.65927	-0.341	0.002	-0.035
17	0.725	0.711	-0.014	0.719	0.67247	-0.047	-0.006	-0.038
18	0.580	0.346	-0.234	1.000	0.71877	-0.281	0.420	0.372
19	0.627	0.503	-0.123	1.000	0.56142	-0.439	0.373	0.058
20	0.724	0.699	-0.024	0.592	0.60638	0.014	-0.131	-0.093
21	0.481	0.395	-0.085	0.548	0.40041	-0.148	0.067	0.005
22	0.832	0.803	-0.029	0.721	0.71554	-0.006	-0.110	-0.087
23	0.947	0.908	-0.039	0.966	0.88485	-0.081	0.019	-0.023
24	0.983	0.968	-0.015	0.953	0.93139	-0.021	-0.030	-0.036
25	0.429	0.432	0.003	0.184	0.22197	0.038	-0.245	-0.210
26	1.000	0.996	-0.004	1.000	0.99989	0.000	0.000	0.003
27	0.982	0.983	0.001	0.931	0.46868	-0.463	-0.051	-0.515
28	0.995	0.985	-0.011	0.914	0.73072	-0.183	-0.082	-0.254
29	0.516	0.519	0.003	0.139	0.26469	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 4-4. Results of Item Fit Index- δ for TIMSS B2 Data under the DINA and DINA-H Models

	U.S. Sample			Benchmark Sample			Difference (Benchmark -U.S.)	
Item	DINA	DINA-H	DINA-DINA-H	DINA	DINA-H	DINA-DINA-H	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 4-5. Results of Item Fit Index-IDI for TIMSS B2 Data under the DINA and DINA-H Models

	U.S. Sample			Benchmark Sample			Difference (Benchmark -U.S.)	
Item	DINA	DINA-H	DINA-DINA-H	DINA	DINA-H	DINA-DINA-H	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 4-6. 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 4-7. 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 4-8. 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 4-9. 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 4-10. 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 4-11. 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 4-12. Results of Item Fit Index- δ for TIMSS B1 Data under the DINO and DINO-H Models

	U.S. Sample			Benchmark Sample			Difference (Benchmark -U.S.)	
Item	DINO	DINO-H	DINO-DINO-H	DINO	DINO-H	DINO-DINO-H	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 4-13. Results of Item Fit Index-IDI for TIMSS B1 Data under the DINO and DINO-H Models

	U.S. Sample			Benchmark Sample			Difference (Benchmark -U.S.)	
Item	DINO	DINO-H	DINO-DINO-H	DINO	DINO-H	DINO-DINO-H	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 4-14. Results of Item Fit Index- δ for TIMSS B2 Data under the DINO and DINO-H Models

	U.S. Sample			Benchmark Sample			Difference (Benchmark -U.S.)	
Item	DINO	DINO-H	DINO-DINO-H	DINO	DINO-H	DINO-DINO-H	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 4-15. Results of Item Fit Index-IDI for TIMSS B2 Data under the DINO and DINO-H Models

	U.S. Sample			Benchmark Sample			Difference (Benchmark -U.S.)	
Item	DINO	DINO-H	DINO-DINO-H	DINO	DINO-H	DINO-DINO-H	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 4-16. 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 4-17. 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 4-18. 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 4-19. 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 4-20. 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 4-21. 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 4-22. 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 4-23. Differences of Item Fit Index-IDI between the DINA(-H) and DINO(-H) Models for TIMSS B1 Data

IDI	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 4-24. Differences of Item Fit Index- δ 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	-0.034	-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 4-25. 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

Table 4-26. Results of Fit Indices of the Main Effect of Model Consistency for DINA(-H)

Overall	MAIC	MBIC
HL_NA	33567	34546
HL_HL	**33271**	**33510**
HL_HU	33412	34007
HU_NA	35683	36662
HU_HL	36740	36978
HU_HU	**35529**	**36123**
NA_NA	**36129**	**37108**
NA_HL	37150	37389
NA_HU	36553	37147

Note. The bold characters indicate the conditions whose data-generating model is consistent with its estimation model. The gray highlighted values indicate the best results (i.e., the smallest MAIC and MBIC) among each comparison. These rules apply to all the following tables.

Table 4-27. Summary Statistics for the Main Effect of Model Consistency for DINA(-H)

Overall	ASB(g)	AVAR(g)	AMSE(g)	ASB(s)	AVAR(s)	AMSE(s)
HL_NA	0.00281	0.00061	0.00342	0.00270	0.00155	0.00425
HL_HL	**0.00162**	**0.00062**	**0.00225**	**0.00166**	**0.00149**	**0.00315**
HL_HU	0.00243	0.00079	0.00321	0.00274	0.00156	0.00430
HU_NA	0.00294	0.00178	0.00473	0.00215	0.00298	0.00513
HU_HL	0.01255	0.00295	0.01549	0.05011	0.00433	0.05444
HU_HU	**0.00164**	**0.00203**	**0.00367**	**0.00193**	**0.00291**	**0.00484**
NA_NA	**0.00170**	**0.00105**	**0.00275**	**0.00220**	**0.00379**	**0.00598**
NA_HL	0.01179	0.00162	0.01341	0.05307	0.00549	0.05856
NA_HU	0.00660	0.00119	0.00779	0.00945	0.00449	0.01394

Table 4-28. Summary Statistics for the Main Effect of Numbers of Attributes for DINA(-H)

		ASB(g)	AVAR(g)	AMSE(g)	ASB(s)	AVAR(s)	AMSE(s)
K=6	HL_HL	0.003129	0.000806	0.003935	0.003279	0.001404	0.004683
	HU_HU	0.003164	0.003034	0.006198	0.003376	0.002143	0.005519
	NA_NA	0.003143	0.000994	0.004137	0.003649	0.002835	0.006484
	Mean	**0.003145**	**0.001611**	**0.004756**	**0.003435**	**0.002127**	**0.005562**
K=8	HL_HL	0.000118	0.000438	0.000556	0.000035	0.001582	0.001617
	HU_HU	0.000126	0.001020	0.001146	0.000493	0.003671	0.004164
	NA_NA	0.000252	0.001104	0.001356	0.000745	0.004737	0.005483
	Mean	**0.000165**	**0.000854**	**0.001019**	**0.000424**	**0.003330**	**0.003754**

Table 4-29. Summary Statistics for the Main Effect of Test Lengths for DINA(-H)

		ASB(g)	AVAR(g)	AMSE(g)	ASB(s)	AVAR(s)	AMSE(s)
J=12	HL_HL	0.000167	0.000697	0.000864	0.000041	0.001836	0.001877
	HU_HU	0.000180	0.003069	0.003249	0.000619	0.003740	0.004359
	NA_NA	0.000301	0.001548	0.001850	0.000970	0.004888	0.005858
	Mean	**0.000216**	**0.001772**	**0.001988**	**0.000543**	**0.003488**	**0.004031**
J=30	HL_HL	0.003080	0.000546	0.003626	0.003273	0.001149	0.004423
	HU_HU	0.003110	0.000984	0.004094	0.003250	0.002074	0.005324
	NA_NA	0.003093	0.000549	0.003643	0.003425	0.002684	0.006109
	Mean	**0.003094**	**0.000693**	**0.003788**	**0.003316**	**0.001969**	**0.005285**

Table 4-30. Summary Statistics for the Main Effect of Sample Sizes for DINA(-H)

		ASB(g)	AVAR(g)	AMSE(g)	ASB(s)	AVAR(s)	AMSE(s)
N=300	HL_HL	0.00005	0.00134	0.00139	0.00007	0.00328	0.00335
	HU_HU	0.00017	0.00376	0.00393	0.00083	0.00640	0.00723
	NA_NA	0.00033	0.00235	0.00268	0.00137	0.00793	0.00930
	Mean	**0.00018**	**0.00249**	**0.00267**	**0.00076**	**0.00587**	**0.00663**
N=1000	HL_HL	0.00469	0.00040	0.00509	0.00489	0.00091	0.00581
	HU_HU	0.00467	0.00183	0.00650	0.00483	0.00174	0.00657
	NA_NA	0.00467	0.00062	0.00530	0.00502	0.00256	0.00758
	Mean	**0.00468**	**0.00095**	**0.00563**	**0.00492**	**0.00174**	**0.00665**
N=3000	HL_HL	0.00012	0.00013	0.00025	0.00001	0.00029	0.00030
	HU_HU	0.00010	0.00049	0.00058	0.00014	0.00058	0.00072
	NA_NA	0.00009	0.00017	0.00026	0.00021	0.00086	0.00107
	Mean	**0.00010**	**0.00026**	**0.00037**	**0.00012**	**0.00058**	**0.00069**

Table 4-31. Summary Statistics for the Guessing Parameter for the Interaction Effect of J by K for DINA(-H)

		K=6			K=8		
		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
J=12	HL_HL	0.00011	0.00101	0.00113	0.00022	0.00038	0.00060
	HU_HU	0.00015	0.00468	0.00483	0.00021	0.00146	0.00167
	NA_NA	0.00014	0.00146	0.00160	0.00047	0.00163	0.00210
	Mean	**0.00013**	**0.00239**	**0.00252**	**0.00030**	**0.00116**	**0.00146**
J=30	HL_HL	0.00614	0.00060	0.00674	0.00002	0.00049	0.00051
	HU_HU	0.00618	0.00139	0.00756	0.00004	0.00058	0.00062
	NA_NA	0.00615	0.00052	0.00667	0.00004	0.00058	0.00061
	Mean	**0.00616**	**0.00084**	**0.00699**	**0.00003**	**0.00055**	**0.00058**

Table 4-32. Summary Statistics for the Slip Parameter for the Interaction Effect of J by K for DINA(-H)

		K=6			K=8		
		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
J=12	HL_HL	0.00004	0.00178	0.00181	0.00004	0.00190	0.00194
	HU_HU	0.00041	0.00293	0.00334	0.00082	0.00455	0.00538
	NA_NA	0.00090	0.00403	0.00494	0.00103	0.00574	0.00678
	Mean	**0.00045**	**0.00291**	**0.00336**	**0.00063**	**0.00406**	**0.00470**
J=30	HL_HL	0.00652	0.00103	0.00755	0.00003	0.00127	0.00129
	HU_HU	0.00634	0.00136	0.00770	0.00016	0.00279	0.00295
	NA_NA	0.00639	0.00164	0.00803	0.00046	0.00373	0.00419
	Mean	**0.00642**	**0.00134**	**0.00776**	**0.00021**	**0.00260**	**0.00281**

Table 4-33. Summary Statistics for the Guessing Parameter for the Interaction Effect of N by K for DINA(-H)

		K=6			K=8		
		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
N=300	HL_HL	0.00002	0.00173	0.00175	0.00008	0.00095	0.00104
	HU_HU	0.00010	0.00526	0.00536	0.00025	0.00226	0.00251
	NA_NA	0.00008	0.00222	0.00230	0.00057	0.00249	0.00306
	Mean	**0.00007**	**0.00307**	**0.00313**	**0.00030**	**0.00190**	**0.00220**
N=1000	HL_HL	0.00928	0.00053	0.00981	0.00011	0.00027	0.00038
	HU_HU	0.00929	0.00306	0.01235	0.00004	0.00060	0.00064
	NA_NA	0.00927	0.00058	0.00985	0.00008	0.00066	0.00075
	Mean	**0.00928**	**0.00139**	**0.01067**	**0.00008**	**0.00051**	**0.00059**
N=3000	HL_HL	0.00009	0.00016	0.00025	0.00016	0.00009	0.00025
	HU_HU	0.00010	0.00078	0.00089	0.00009	0.00019	0.00028
	NA_NA	0.00008	0.00018	0.00026	0.00010	0.00016	0.00027
	Mean	**0.00009**	**0.00037**	**0.00047**	**0.00012**	**0.00015**	**0.00027**

Table 4-34. Summary Statistics for the Slip Parameter for the Interaction Effect of N by K for DINA(-H)

		K=6			K=8		
		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
N=300	HL_HL	0.000056	0.003026	0.003082	0.000080	0.003530	0.003610
	HU_HU	0.000365	0.004889	0.005254	0.001302	0.007912	0.009214
	NA_NA	0.001001	0.006007	0.007008	0.001732	0.009861	0.011593
	Mean	**0.000474**	**0.004641**	**0.005115**	**0.001038**	**0.007101**	**0.008139**
N=1000	HL_HL	0.009772	0.000901	0.010674	0.000016	0.000925	0.000941
	HU_HU	0.009601	0.001154	0.010755	0.000067	0.002325	0.002392
	NA_NA	0.009770	0.001947	0.011718	0.000267	0.003181	0.003448
	Mean	**0.009715**	**0.001334**	**0.011049**	**0.000117**	**0.002143**	**0.002260**
N=3000	HL_HL	0.000010	0.000283	0.000293	0.000009	0.000291	0.000300
	HU_HU	0.000162	0.000387	0.000549	0.000109	0.000777	0.000886
	NA_NA	0.000176	0.000552	0.000727	0.000237	0.001171	0.001407
	Mean	**0.000116**	**0.000407**	**0.000523**	**0.000118**	**0.000746**	**0.000864**

Table 4-35. Summary Statistics for the Guessing Parameter for the Interaction Effect of N by J for DINA(-H)

		J=12			J=30		
		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
N=300	HL_HL	0.00008	0.00155	0.00163	0.00002	0.00113	0.00115
	HU_HU	0.00024	0.00590	0.00614	0.00010	0.00163	0.00173
	NA_NA	0.00060	0.00351	0.00411	0.00006	0.00119	0.00125
	Mean	**0.00031**	**0.00365**	**0.00396**	**0.00006**	**0.00132**	**0.00138**
N=1000	HL_HL	0.00018	0.00041	0.00059	0.00921	0.00039	0.00960
	HU_HU	0.00013	0.00256	0.00269	0.00920	0.00110	0.01030
	NA_NA	0.00013	0.00090	0.00103	0.00922	0.00034	0.00956
	Mean	**0.00015**	**0.00129**	**0.00144**	**0.00921**	**0.00061**	**0.00982**
N=3000	HL_HL	0.00024	0.00013	0.00038	0.00001	0.00012	0.00013
	HU_HU	0.00016	0.00075	0.00091	0.00003	0.00022	0.00025
	NA_NA	0.00018	0.00023	0.00041	0.00001	0.00011	0.00012
	Mean	**0.00020**	**0.00037**	**0.00057**	**0.00001**	**0.00015**	**0.00017**

Table 4-36. Summary Statistics for the Slip Parameter for the Interaction Effect of N by J for DINA(-H)

		J=12			J=30		
		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
N=300	HL_HL	0.00010	0.00401	0.00411	0.00004	0.00254	0.00259
	HU_HU	0.00145	0.00839	0.00985	0.00022	0.00441	0.00462
	NA_NA	0.00212	0.01028	0.01240	0.00061	0.00559	0.00620
	Mean	**0.00122**	**0.00756**	**0.00878**	**0.00029**	**0.00418**	**0.00447**
N=1000	HL_HL	0.00002	0.00115	0.00117	0.00977	0.00067	0.01044
	HU_HU	0.00016	0.00210	0.00226	0.00951	0.00138	0.01089
	NA_NA	0.00042	0.00331	0.00372	0.00962	0.00182	0.01144
	Mean	**0.00020**	**0.00219**	**0.00238**	**0.00963**	**0.00129**	**0.01092**
N=3000	HL_HL	0.00001	0.00034	0.00035	0.00001	0.00023	0.00024
	HU_HU	0.00025	0.00072	0.00097	0.00002	0.00044	0.00046
	NA_NA	0.00037	0.00108	0.00145	0.00004	0.00064	0.00068
	Mean	**0.00021**	**0.00072**	**0.00093**	**0.00002**	**0.00044**	**0.00046**

Table 4-37. Summary Statistics for the Guessing Parameter for the Three-Way Interaction Effect of N by J by K for DINA(-H)

		K=6			K=8		
		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
N=300	J=12	0.000079	0.004693	0.004772	0.000533	0.002612	0.003145
	J=30	0.000051	0.001445	0.001496	0.000067	0.001191	0.001258
N=1000	J=12	0.000158	0.001904	0.002062	0.000137	0.000678	0.000815
	J=30	0.018400	0.000876	0.019276	0.000020	0.000344	0.000363
N=3000	J=12	0.000164	0.000560	0.000724	0.000226	0.000183	0.000409
	J=30	0.000020	0.000189	0.000208	0.000009	0.000116	0.000125

Table 4-38. Summary Statistics for the Slip Parameter for the Three-Way Interaction Effect of N by J by K for DINA(-H)

		K=6			K=8		
		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
N=300	J=12	0.000874	0.006396	0.007270	0.001570	0.008727	0.010297
	J=30	0.000074	0.002885	0.002959	0.000506	0.005475	0.005981
N=1000	J=12	0.000264	0.001804	0.002069	0.000129	0.002572	0.002701
	J=30	0.019165	0.000864	0.020029	0.000104	0.001715	0.001819
N=3000	J=12	0.000216	0.000536	0.000752	0.000205	0.000894	0.001099
	J=30	0.000016	0.000278	0.000294	0.000031	0.000598	0.000629

Table 4-39. Results of Fit Indices of the Main Effect of Model Consistency for DINO(-H)

Overall	MAIC	MBIC
HL_NO	32872	33851
HL_HL	**32576**	**32815**
HL_HU	32721	33315
HU_NO	33278	34257
HU_HL	33445	33684
HU_HU	**33128**	**33723**
NO_NO	**35127**	**36106**
NO_HL	36218	36457
NO_HU	36065	36659

Table 4-40. Summary Statistics for the Main Effect of Model Consistency for DINO(-H)

Overall	ASB(g)	AVAR(g)	AMSE(g)	ASB(s)	AVAR(s)	AMSE(s)
HL_NO	0.00054	0.00172	0.00198	0.00134	0.00074	0.00207
HL_HL	**0.00003**	**0.00143**	**0.00146**	**0.00004**	**0.00061**	**0.00065**
HL_HU	0.00011	0.00147	0.00158	0.00144	0.00064	0.00208
HU_NO	0.00525	0.01150	0.01676	0.00034	0.00111	0.00145
HU_HL	0.02297	0.01368	0.03665	0.00732	0.00079	0.00811
HU_HU	**0.00055**	**0.01448**	**0.01503**	**0.00022**	**0.00092**	**0.00113**
NO_NO	**0.00063**	**0.00403**	**0.00466**	**0.00022**	**0.00103**	**0.00124**
NO_HL	0.04939	0.00672	0.05611	0.01022	0.00162	0.01184
NO_HU	0.01658	0.00756	0.02413	0.00759	0.00110	0.00869

Table 4-41. Summary Statistics for the Main Effect of Numbers of Attributes for DINO(-H)

		ASB(g)	AVAR(g)	AMSE(g)	ASB(s)	AVAR(s)	AMSE(s)
K=6	HL_HL	0.000021	0.001280	0.001302	0.000051	0.000800	0.000851
	HU_HU	0.000209	0.006030	0.006239	0.000289	0.001171	0.001460
	NO_NO	0.000330	0.002982	0.003311	0.000281	0.001273	0.001554
	Mean	**0.000187**	**0.003431**	**0.003617**	**0.000207**	**0.001081**	**0.001288**
K=8	HL_HL	0.000035	0.001579	0.001614	0.000030	0.000427	0.000457
	HU_HU	0.000894	0.022922	0.023816	0.000148	0.000661	0.000810
	NO_NO	0.000938	0.005070	0.006008	0.000153	0.000779	0.000931
	Mean	**0.000622**	**0.009857**	**0.010479**	**0.000110**	**0.000622**	**0.000733**

Table 4-42. Summary Statistics for the Main Effect of Test Lengths for DINO(-H)

		ASB(g)	AVAR(g)	AMSE(g)	ASB(s)	AVAR(s)	AMSE(s)
J=12	HL_HL	0.000034	0.001682	0.001716	0.000030	0.000738	0.000768
	HU_HU	0.000504	0.013632	0.014136	0.000391	0.001242	0.001633
	NO_NO	0.000856	0.005301	0.006157	0.000406	0.001513	0.001919
	Mean	**0.000465**	**0.006871**	**0.007336**	**0.000275**	**0.001164**	**0.001440**
J=30	HL_HL	0.000023	0.001177	0.001200	0.000051	0.000489	0.000540
	HU_HU	0.000599	0.015320	0.015919	0.000046	0.000590	0.000637
	NO_NO	0.000411	0.002751	0.003162	0.000028	0.000539	0.000567
	Mean	**0.000344**	**0.006416**	**0.006760**	**0.000042**	**0.000539**	**0.000581**

Table 4-43. Summary Statistics for the Main Effect of Sample Sizes for DINO(-H)

		ASB(g)	AVAR(g)	AMSE(g)	ASB(s)	AVAR(s)	AMSE(s)
N=300	HL_HL	0.00006	0.00304	0.00310	0.00008	0.00134	0.00142
	HU_HU	0.00114	0.03251	0.03364	0.00039	0.00206	0.00244
	NO_NO	0.00138	0.00864	0.01002	0.00046	0.00233	0.00279
	Mean	**0.00086**	**0.01473**	**0.01559**	**0.00031**	**0.00191**	**0.00222**
N=1000	HL_HL	0.00002	0.00093	0.00095	0.00002	0.00037	0.00040
	HU_HU	0.00028	0.00781	0.00809	0.00013	0.00054	0.00067
	NO_NO	0.00026	0.00262	0.00288	0.00010	0.00058	0.00068
	Mean	**0.00019**	**0.00378**	**0.00397**	**0.00008**	**0.00050**	**0.00058**
N=3000	HL_HL	0.00001	0.00032	0.00033	0.00002	0.00013	0.00014
	HU_HU	0.00024	0.00311	0.00335	0.00014	0.00015	0.00029
	NO_NO	0.00026	0.00082	0.00108	0.00009	0.00016	0.00025
	Mean	**0.00017**	**0.00142**	**0.00159**	**0.00008**	**0.00015**	**0.00023**

Table 4-44. Summary Statistics for the Guessing Parameter for the Interaction Effect of J by K for DINO(-H)

		K=6			K=8		
		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
J=12	HL_HL	0.00002	0.00166	0.00168	0.00004	0.001707	0.001752
	HU_HU	0.00033	0.00864	0.00896	0.00068	0.018628	0.019308
	NO_NO	0.00057	0.00419	0.00476	0.00114	0.006409	0.007552
	Mean	**0.00031**	**0.00483**	**0.00514**	**0.00062**	**0.008914**	**0.009537**
J=30	HL_HL	0.00002	0.00090	0.00092	0.00003	0.001450	0.001475
	HU_HU	0.00009	0.00342	0.00351	0.00111	0.027216	0.028324
	NO_NO	0.00009	0.00177	0.00186	0.00073	0.003731	0.004463
	Mean	**0.00007**	**0.00203**	**0.00210**	**0.00062**	**0.010799**	**0.011421**

Table 4-45. Summary Statistics for the Slip Parameter for the Interaction Effect of J by K for DINO(-H)

		K=6			K=8		
		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
J=12	HL_HL	0.00005	0.00106	0.00111	0.00001	0.00041	0.00042
	HU_HU	0.00055	0.00179	0.00234	0.00023	0.00069	0.00093
	NO_NO	0.00055	0.00204	0.00259	0.00026	0.00099	0.00125
	Mean	**0.00038**	**0.00163**	**0.00201**	**0.00017**	**0.00070**	**0.00087**
J=30	HL_HL	0.00006	0.00053	0.00059	0.00005	0.00044	0.00049
	HU_HU	0.00003	0.00055	0.00058	0.00006	0.00063	0.00069
	NO_NO	0.00001	0.00051	0.00052	0.00004	0.00057	0.00061
	Mean	**0.00003**	**0.00053**	**0.00056**	**0.00005**	**0.00055**	**0.00060**

Table 4-46. Summary Statistics for the Guessing Parameter for the Interaction Effect of N by K for DINO(-H)

		K=6			K=8		
		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
N=300	HL_HL	0.00004	0.00273	0.00277	0.00008	0.00335	0.00343
	HU_HU	0.00035	0.01225	0.01260	0.00192	0.05276	0.05469
	NO_NO	0.00073	0.00664	0.00737	0.00202	0.01064	0.01267
	Mean	**0.00037**	**0.00721**	**0.00758**	**0.00134**	**0.02225**	**0.02360**
N=1000	HL_HL	0.00002	0.00083	0.00085	0.00002	0.00102	0.00104
	HU_HU	0.00013	0.00458	0.00471	0.00043	0.01103	0.01146
	NO_NO	0.00010	0.00181	0.00191	0.00043	0.00342	0.00385
	Mean	**0.00008**	**0.00241**	**0.00249**	**0.00029**	**0.00516**	**0.00545**
N=3000	HL_HL	0.00001	0.00028	0.00028	0.00001	0.00036	0.00037
	HU_HU	0.00015	0.00126	0.00140	0.00032	0.00497	0.00530
	NO_NO	0.00016	0.00049	0.00065	0.00036	0.00114	0.00151
	Mean	**0.00010**	**0.00068**	**0.00078**	**0.00023**	**0.00216**	**0.00239**

Table 4-47. Summary Statistics for the Slip Parameter for the Interaction Effect of N by K for DINO(-H)

		K=6			K=8		
		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
N=300	HL_HL	0.000108	0.001740	0.001848	0.000058	0.000940	0.000998
	HU_HU	0.000540	0.002614	0.003153	0.000236	0.001498	0.001734
	NO_NO	0.000671	0.002911	0.003582	0.000253	0.001751	0.002005
	Mean	**0.000439**	**0.002422**	**0.002861**	**0.000183**	**0.001396**	**0.001579**
N=1000	HL_HL	0.000025	0.000494	0.000519	0.000019	0.000256	0.000275
	HU_HU	0.000201	0.000697	0.000898	0.000061	0.000380	0.000441
	NO_NO	0.000101	0.000719	0.000820	0.000098	0.000450	0.000547
	Mean	**0.000109**	**0.000637**	**0.000746**	**0.000059**	**0.000362**	**0.000421**
N=3000	HL_HL	0.000020	0.000166	0.000186	0.000012	0.000085	0.000098
	HU_HU	0.000125	0.000202	0.000328	0.000148	0.000106	0.000254
	NO_NO	0.000072	0.000189	0.000261	0.000107	0.000135	0.000241
	Mean	**0.000073**	**0.000186**	**0.000258**	**0.000089**	**0.000109**	**0.000198**

Table 4-48. Summary Statistics for the Guessing Parameter for the Interaction Effect of N by J for DINO(-H)

		J=12			J=30		
		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
N=300	HL_HL	0.00007	0.00362	0.00368	0.00005	0.00247	0.00252
	HU_HU	0.00070	0.02576	0.02645	0.00158	0.03925	0.04083
	NO_NO	0.00177	0.01147	0.01323	0.00099	0.00582	0.00681
	Mean	**0.00084**	**0.01361**	**0.01446**	**0.00087**	**0.01585**	**0.01672**
N=1000	HL_HL	0.00002	0.00106	0.00108	0.00001	0.00079	0.00081
	HU_HU	0.00041	0.01101	0.01142	0.00015	0.00461	0.00476
	NO_NO	0.00040	0.00350	0.00390	0.00012	0.00174	0.00186
	Mean	**0.00028**	**0.00519**	**0.00547**	**0.00009**	**0.00238**	**0.00247**
N=3000	HL_HL	0.00001	0.00037	0.00038	0.00001	0.00027	0.00027
	HU_HU	0.00040	0.00413	0.00453	0.00007	0.00210	0.00217
	NO_NO	0.00040	0.00094	0.00134	0.00012	0.00070	0.00082
	Mean	**0.00027**	**0.00181**	**0.00208**	**0.00007**	**0.00102**	**0.00109**

Table 4-49. Summary Statistics for the Slip Parameter for the Interaction Effect of N by J for DINO(-H)

		J=12			J=30		
		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
N=300	HL_HL	0.00005	0.00161	0.00167	0.00011	0.00107	0.00118
	HU_HU	0.00066	0.00282	0.00348	0.00011	0.00130	0.00141
	NO_NO	0.00085	0.00348	0.00434	0.00007	0.00118	0.00125
	Mean	**0.00052**	**0.00264**	**0.00316**	**0.00010**	**0.00118**	**0.00128**
N=1000	HL_HL	0.00002	0.00045	0.00047	0.00003	0.00030	0.00033
	HU_HU	0.00024	0.00072	0.00096	0.00002	0.00036	0.00038
	NO_NO	0.00019	0.00084	0.00103	0.00001	0.00033	0.00034
	Mean	**0.00015**	**0.00067**	**0.00082**	**0.00002**	**0.00033**	**0.00035**
N=3000	HL_HL	0.00002	0.00015	0.00017	0.00001	0.00010	0.00011
	HU_HU	0.00027	0.00019	0.00046	0.00001	0.00011	0.00012
	NO_NO	0.00017	0.00022	0.00039	0.00000	0.00011	0.00011
	Mean	**0.00015**	**0.00019**	**0.00034**	**0.00001**	**0.00011**	**0.00012**

Table 4-50. Summary Statistics for the Guessing Parameter for the Three-Way Interaction Effect of N by J by K for DINO(-H)

		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
		K=6			K=8		
N=300	J=12	0.000625	0.010269	0.010894	0.001061	0.016958	0.018020
	J=30	0.000123	0.004143	0.004266	0.001623	0.027549	0.029171
N=1000	J=12	0.000106	0.003351	0.003457	0.000453	0.007026	0.007479
	J=30	0.000058	0.001468	0.001525	0.000132	0.003291	0.003423
N=3000	J=12	0.000190	0.000865	0.001055	0.000354	0.002759	0.003113
	J=30	0.000019	0.000488	0.000507	0.000111	0.001557	0.001668

Table 4-51. Summary Statistics for the Slip Parameter for the Three-Way Interaction Effect of N by J by K for DINO(-H)

		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
		K=6			K=8		
N=300	J=12	0.000805	0.003692	0.004497	0.000242	0.001582	0.001825
	J=30	0.000074	0.001152	0.001226	0.000123	0.001210	0.001333
N=1000	J=12	0.000200	0.000944	0.001144	0.000099	0.000393	0.000492
	J=30	0.000018	0.000329	0.000347	0.000020	0.000331	0.000351
N=3000	J=12	0.000140	0.000258	0.000398	0.000167	0.000116	0.000283
	J=30	0.000005	0.000113	0.000119	0.000011	0.000101	0.000112

Table 4-52. Differences of Fit Indices between the DINA(-H) and DINO(-H) Models for the Main Effect of Model Consistency

DINO-DINA	MAIC	MBIC
HL_NO - HL_NA	-695	-695
HL_HL - HL_HL	**-695**	**-695**
HL_HU - HL_HU	-691	-692
HU_NO - HU_NA	-2405	-2405
HU_HL - HU_HL	-3295	-3294
HU_HU - HU_HU	**-2401**	**-2400**
NO_NO - NA_NA	**-1002**	**-1002**
NO_HL - NA_HL	-932	-932
NO_HU - NA_HU	-488	-488

Table 4-53. Differences of Summary Statistics between the DINA(-H) and DINO(-H) Models for the Main Effect of Model Consistency

DINO-DINA	ASB(g)	AVAR(g)	AMSE(g)	ASB(s)	AVAR(s)	AMSE(s)
HL_NO - HL_NA	-0.00227	0.00111	-0.00144	-0.00136	-0.00081	-0.00218
HL_HL - HL_HL	**-0.00159**	**0.00081**	**-0.00079**	**-0.00162**	**-0.00088**	**-0.00250**
HL_HU - HL_HU	-0.00232	0.00068	-0.00163	-0.00130	-0.00092	-0.00222
HU_NO - HU_NA	0.00231	0.00972	0.01203	-0.00181	-0.00187	-0.00368
HU_HL - HU_HL	0.01042	0.01073	0.02116	-0.04279	-0.00354	-0.04633
HU_HU - HU_HU	**-0.00109**	**0.01245**	**0.01136**	**-0.00171**	**-0.00199**	**-0.00371**
NO_NO - NA_NA	**-0.00107**	**0.00298**	**0.00191**	**-0.00198**	**-0.00276**	**-0.00474**
NO_HL - NA_HL	0.03760	0.00510	0.04270	-0.04285	-0.00387	-0.04672
NO_HU - NA_HU	0.00998	0.00637	0.01634	-0.00186	-0.00339	-0.00525

Note. The highlighted values indicated that the DINA(-H) model performs better than its DINO(-H) counterpart (i.e., the smaller ASB, AVAR, and AMSE in the DINA(-H) model). The bold characters indicate the conditions whose data-generating model is consistent with its estimation model. These rules are the same for the following tables.

Table 4-54. Differences of Summary Statistics between the DINA(-H) and DINO(-H) Models for the Main Effect of Numbers of Attributes

	DINO-DINA	ASB(g)	AVAR(g)	AMSE(g)	ASB(s)	AVAR(s)	AMSE(s)
K=6	HL_HL - HL_HL	-0.0031	0.0005	-0.0026	-0.0032	-0.0006	-0.0038
	HU_HU -HU_HU	-0.0030	0.0030	0.0000	-0.0031	-0.0010	-0.0041
	NO_NO - NA_NA	-0.0028	0.0020	-0.0008	-0.0034	-0.0016	-0.0049
	Mean	**-0.0030**	**0.0018**	**-0.0011**	**-0.0032**	**-0.0010**	**-0.0043**
K=8	HL_HL - HL_HL	-0.0001	0.0011	0.0011	0.0000	-0.0012	-0.0012
	HU_HU -HU_HU	0.0008	0.0219	0.0227	-0.0003	-0.0030	-0.0034
	NO_NO - NA_NA	0.0007	0.0040	0.0047	-0.0006	-0.0040	-0.0046
	Mean	**0.0005**	**0.0090**	**0.0095**	**-0.0003**	**-0.0027**	**-0.0030**

Table 4-55. Differences of Summary Statistics between the DINA(-H) and DINO(-H) Models for the Main Effect of Test Lengths

	DINO-DINA	ASB(g)	AVAR(g)	AMSE(g)	ASB(s)	AVAR(s)	AMSE(s)
J=12	HL_HL - HL_HL	-0.0001	0.0010	0.0009	0.0000	-0.0011	-0.0011
	HU_HU -HU_HU	0.0003	0.0106	0.0109	-0.0002	-0.0025	-0.0027
	NO_NO - NA_NA	0.0006	0.0038	0.0043	-0.0006	-0.0034	-0.0039
	Mean	**0.0002**	**0.0051**	**0.0053**	**-0.0003**	**-0.0023**	**-0.0026**
J=30	HL_HL - HL_HL	-0.0031	0.0006	-0.0024	-0.0032	-0.0007	-0.0039
	HU_HU -HU_HU	-0.0025	0.0143	0.0118	-0.0032	-0.0015	-0.0047
	NO_NO - NA_NA	-0.0027	0.0022	-0.0005	-0.0034	-0.0021	-0.0055
	Mean	**-0.0028**	**0.0057**	**0.0030**	**-0.0033**	**-0.0014**	**-0.0047**

Table 4-56. Differences of Summary Statistics between the DINA(-H) and DINO(-H) Models for the Main Effect of Sample Sizes

	DINO-DINA	ASB(g)	AVAR(g)	AMSE(g)	ASB(s)	AVAR(s)	AMSE(s)
N=300	HL_HL - HL_HL	0.00001	0.00170	0.00171	0.00001	-0.00194	-0.00193
	HU_HU -HU_HU	0.00097	0.02875	0.02971	-0.00044	-0.00434	-0.00479
	NO_NO - NA_NA	0.00105	0.00629	0.00734	-0.00091	-0.00560	-0.00651
	Mean	**0.00068**	**0.01225**	**0.01292**	**-0.00045**	**-0.00396**	**-0.00441**
N=1000	HL_HL - HL_HL	-0.00467	0.00053	-0.00414	-0.00487	-0.00054	-0.00541
	HU_HU -HU_HU	-0.00439	0.00598	0.00159	-0.00470	-0.00120	-0.00590
	NO_NO - NA_NA	-0.00441	0.00200	-0.00242	-0.00492	-0.00198	-0.00690
	Mean	**-0.00449**	**0.00284**	**-0.00166**	**-0.00483**	**-0.00124**	**-0.00607**
N=3000	HL_HL - HL_HL	-0.00011	0.00019	0.00008	0.00001	-0.00016	-0.00016
	HU_HU -HU_HU	0.00014	0.00262	0.00277	0.00000	-0.00043	-0.00043
	NO_NO - NA_NA	0.00017	0.00065	0.00082	-0.00012	-0.00070	-0.00082
	Mean	**0.00007**	**0.00115**	**0.00122**	**-0.00004**	**-0.00043**	**-0.00047**

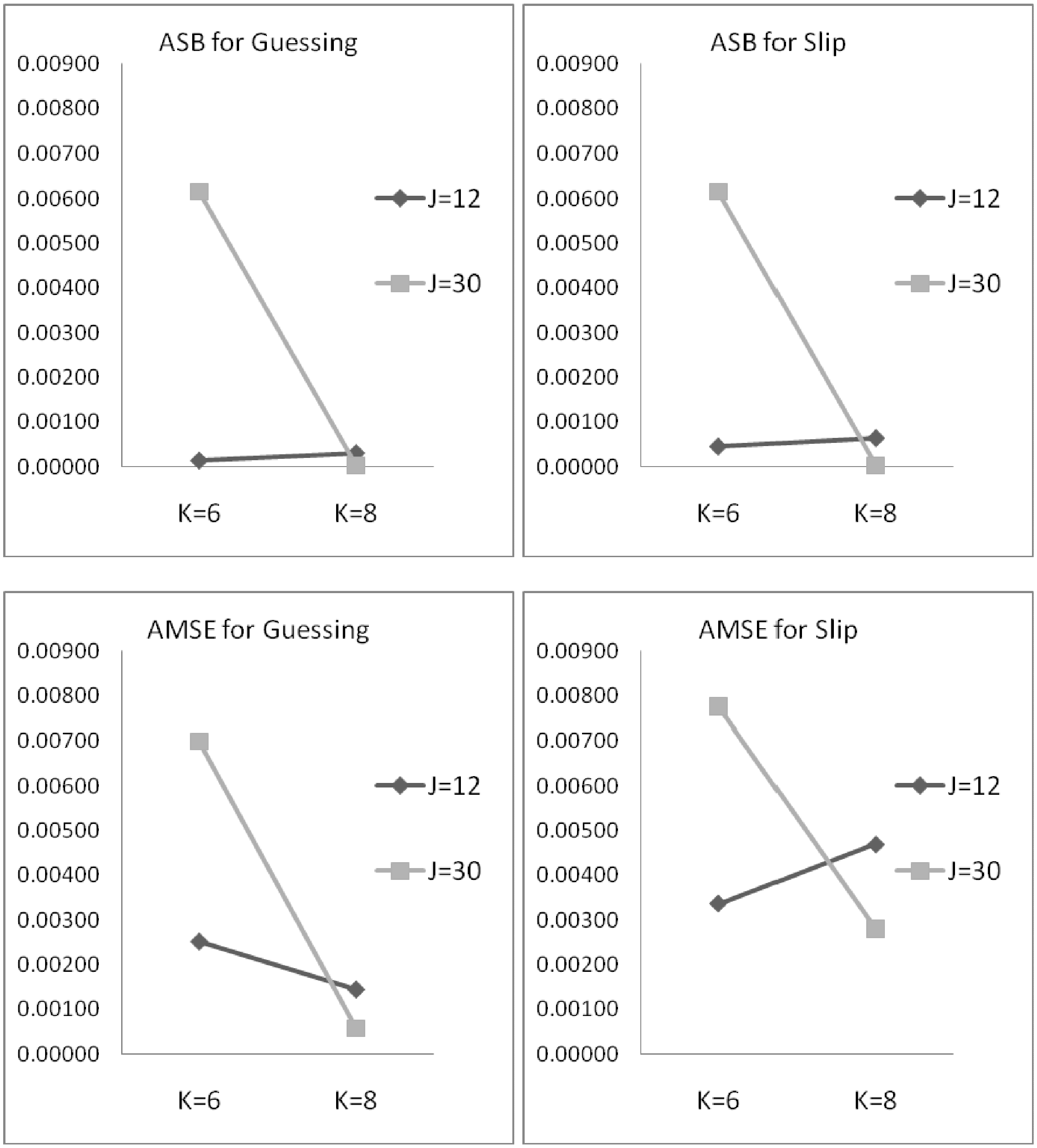


Figure 4-1. Summary statistics for the interaction effect of J by K for DINA(-H)

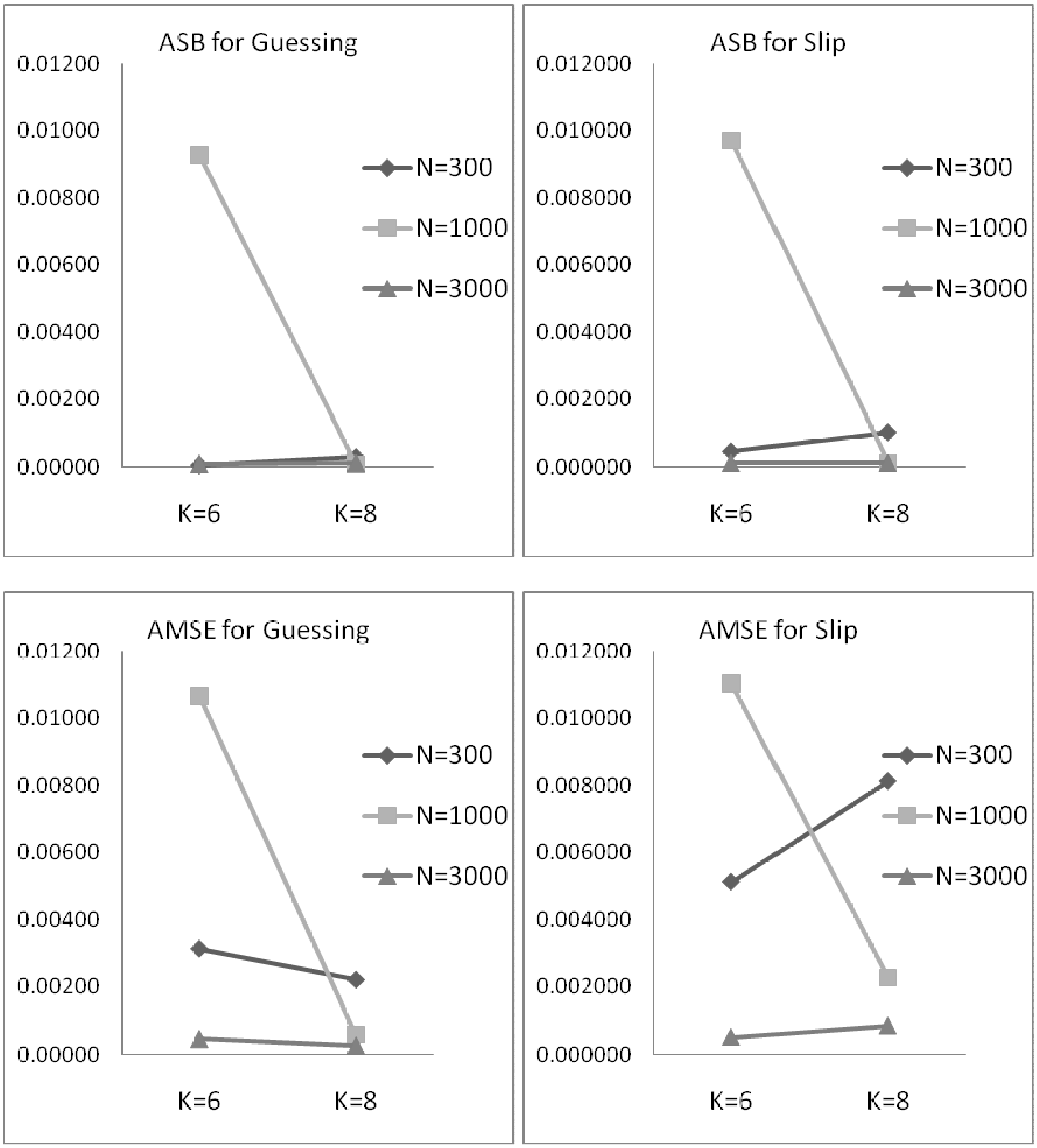


Figure 4-2. Summary statistics for the interaction effect of N by K for DINA(-H)

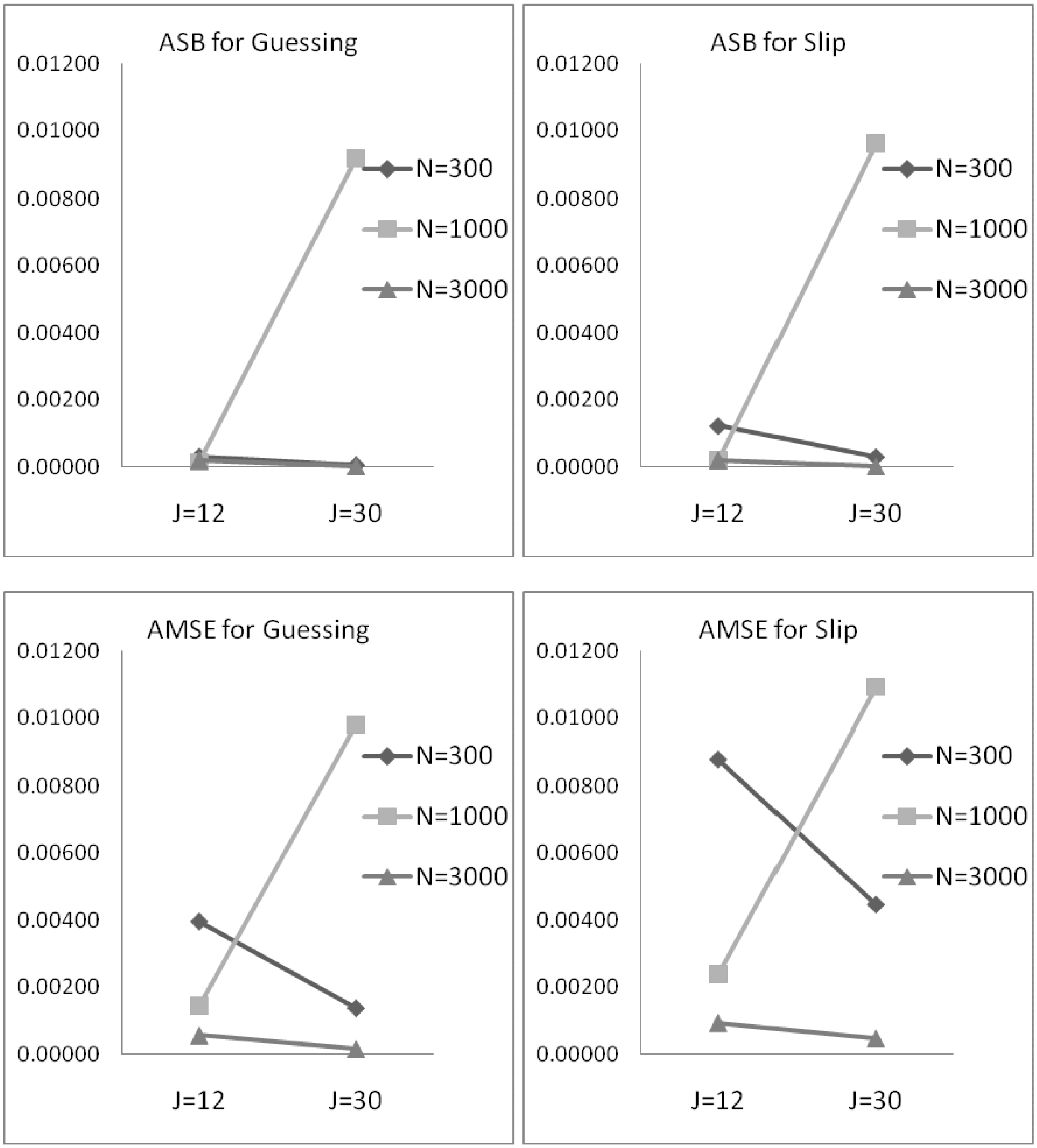


Figure 4-3. Summary statistics for the interaction effect of N by J for DINA(-H)

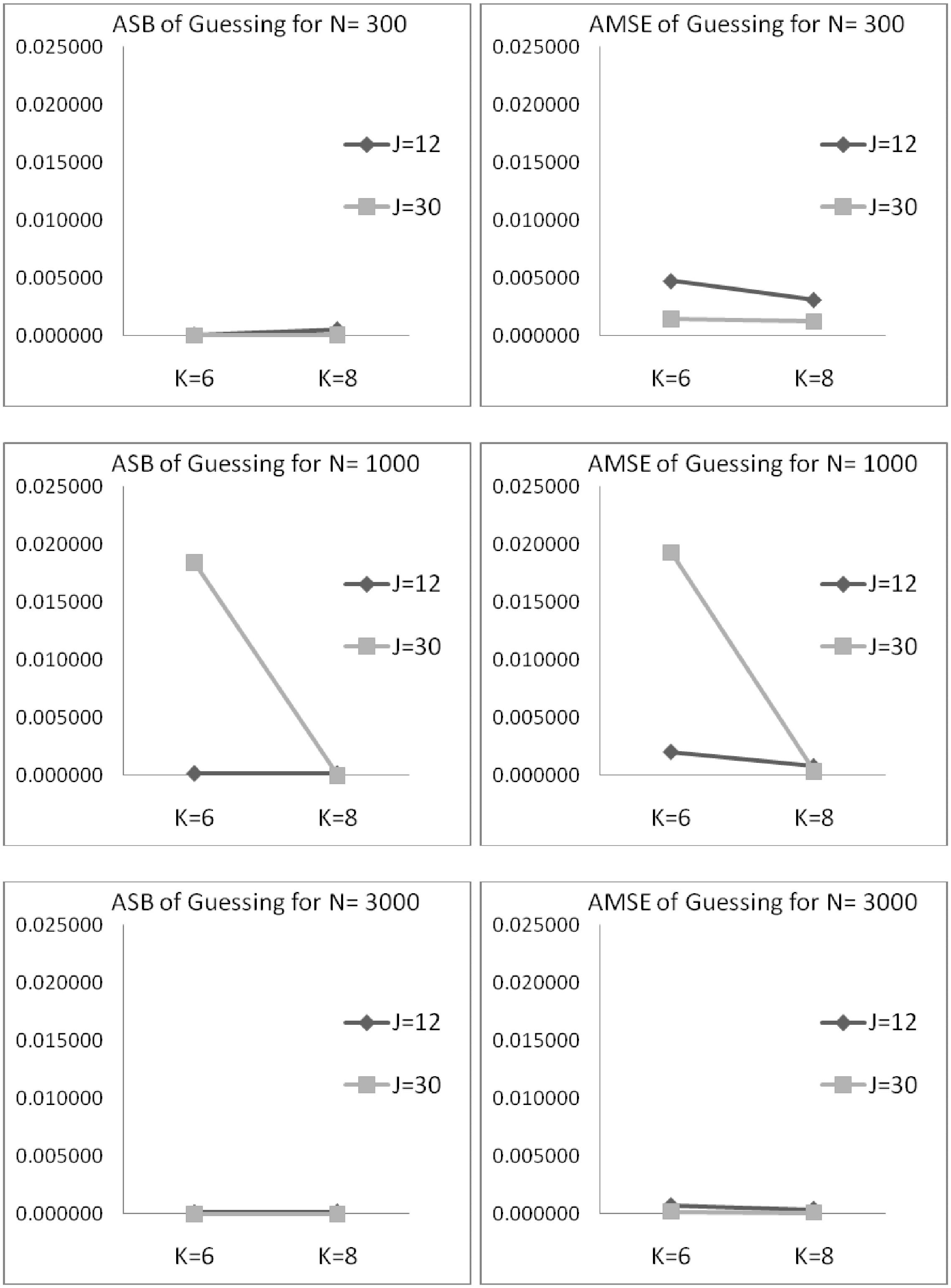


Figure 4-4. Summary statistics for the guessing parameter estimates for the three-way interaction effect of N by J by K for DINA(-H)

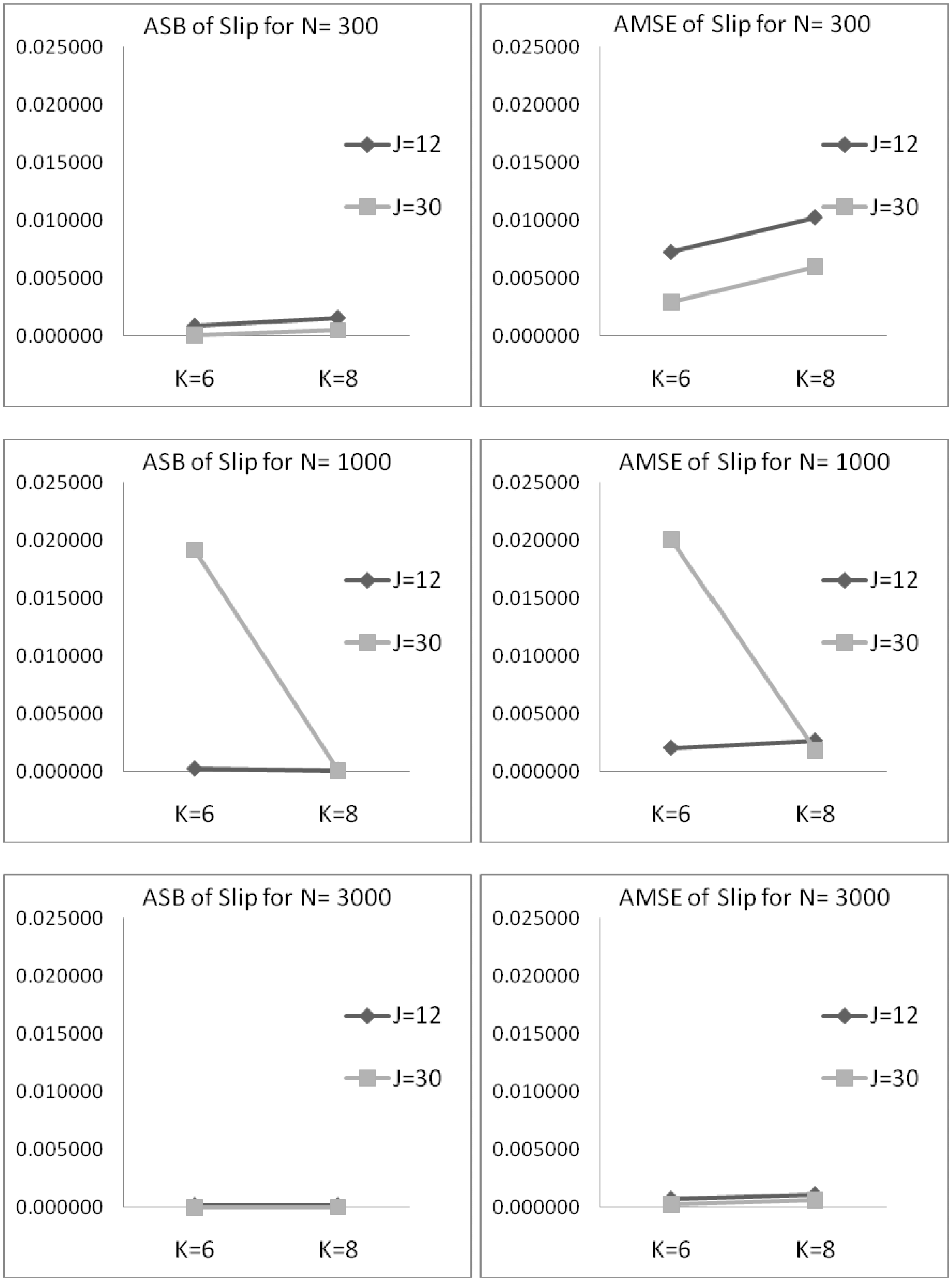


Figure 4-5. Summary statistics for the slip parameter estimates for the three-way interaction effect of N by J by K for DINA(-H)

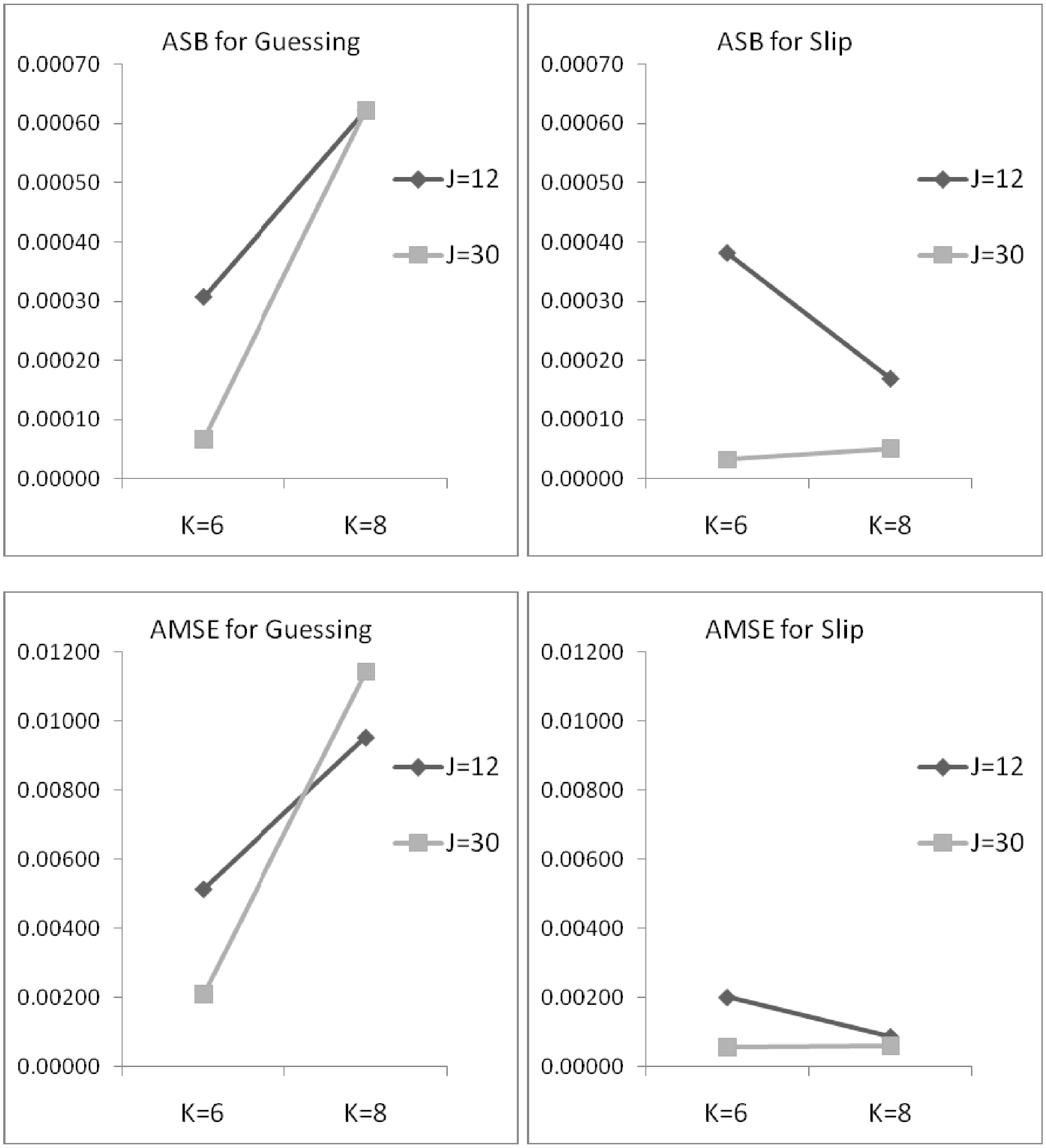


Figure 4-6. Summary statistics for the interaction effect of J by K for DINO(-H)

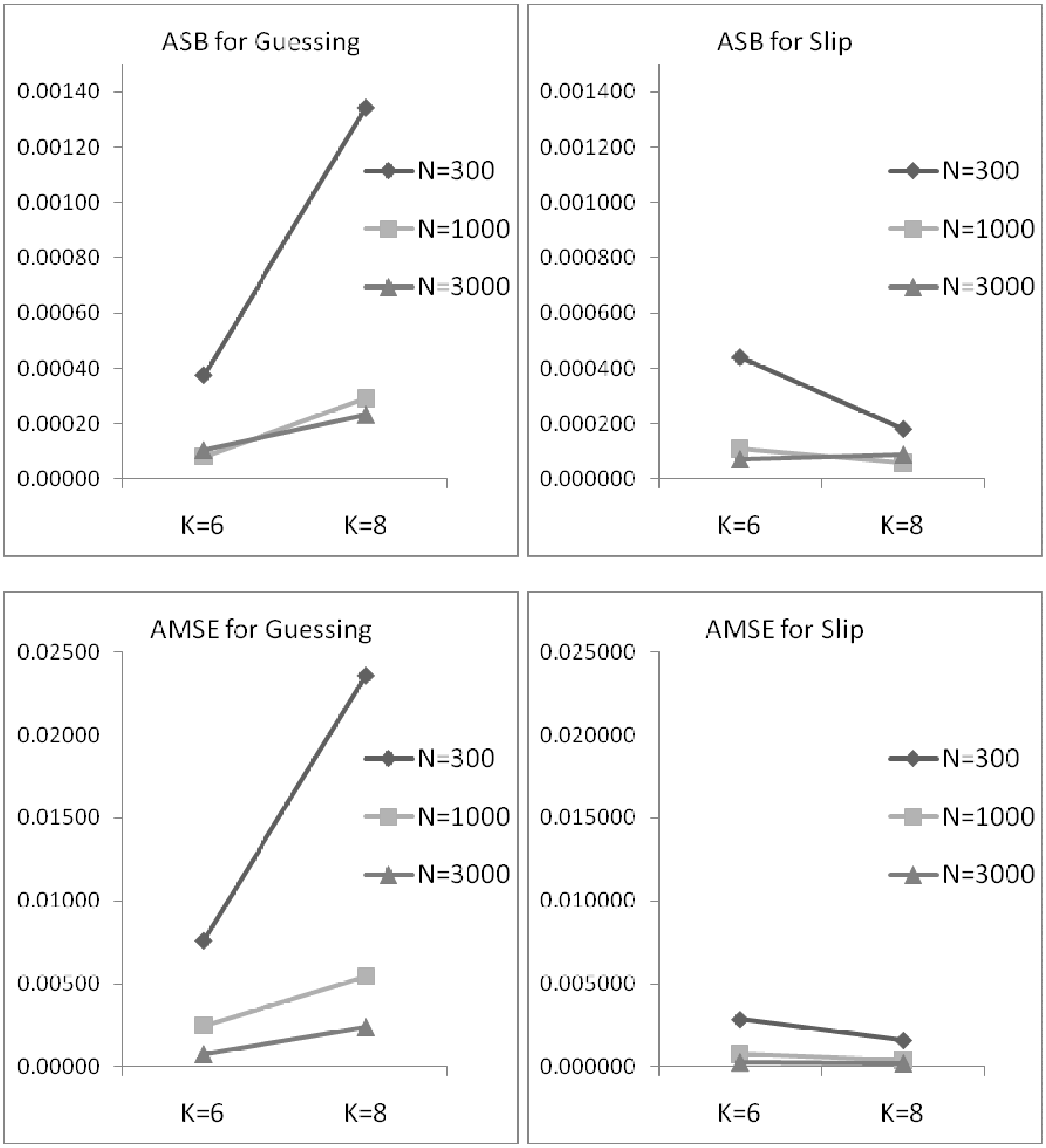


Figure 4-7. Summary statistics for the interaction effect of N by K for DINO(-H)

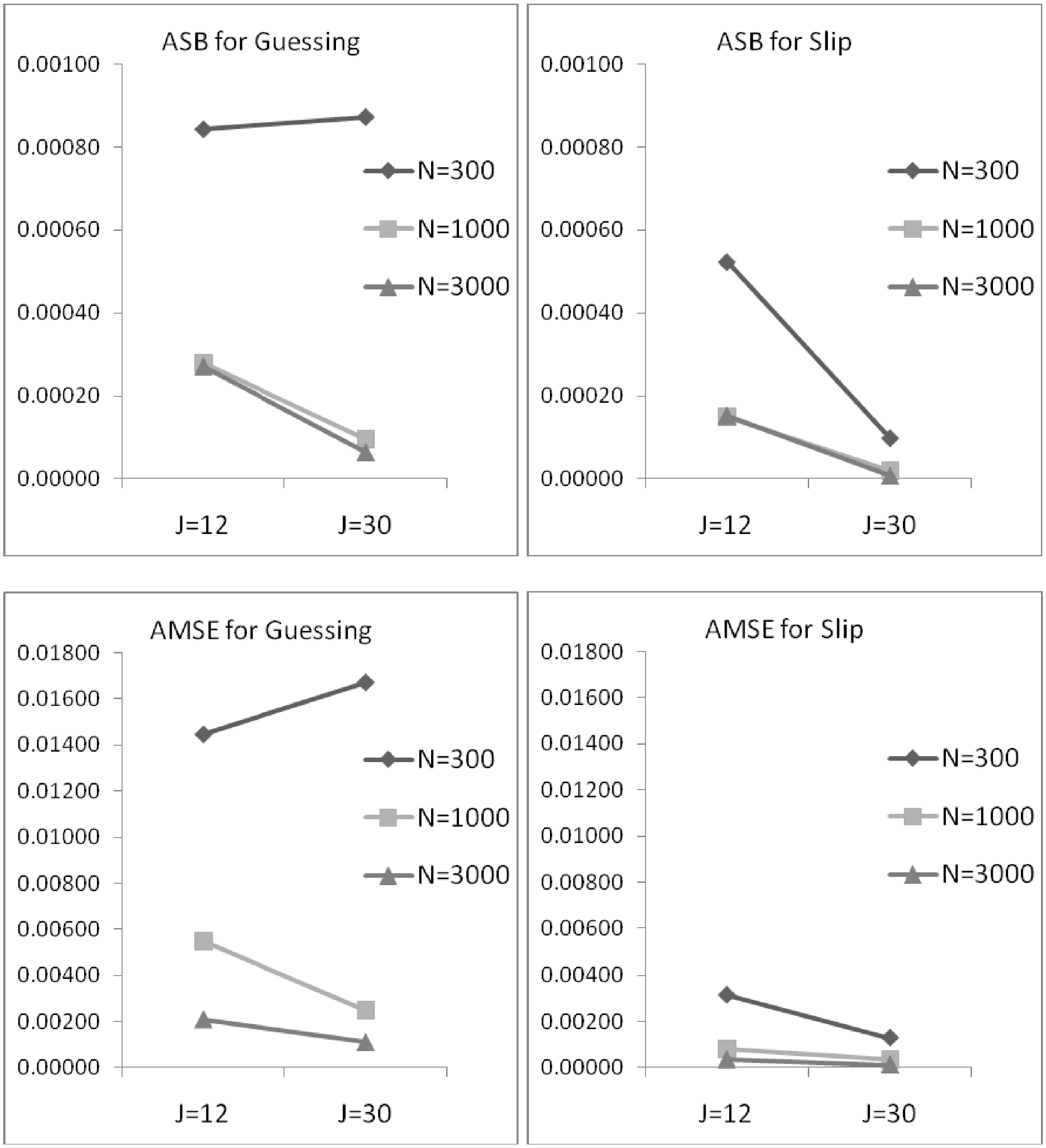


Figure 4-8. Summary statistics for the interaction effect of N by J for DINO(-H)

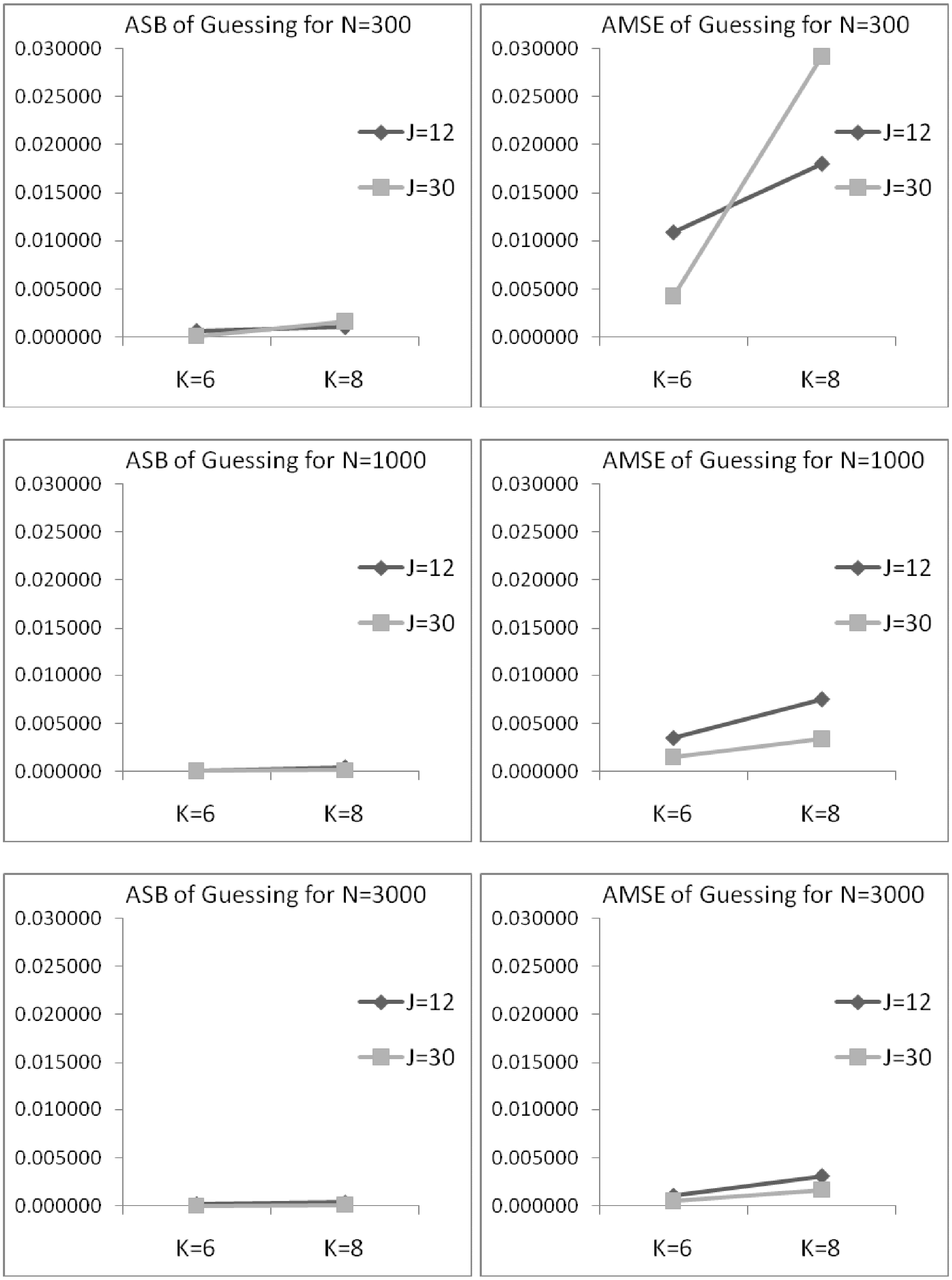
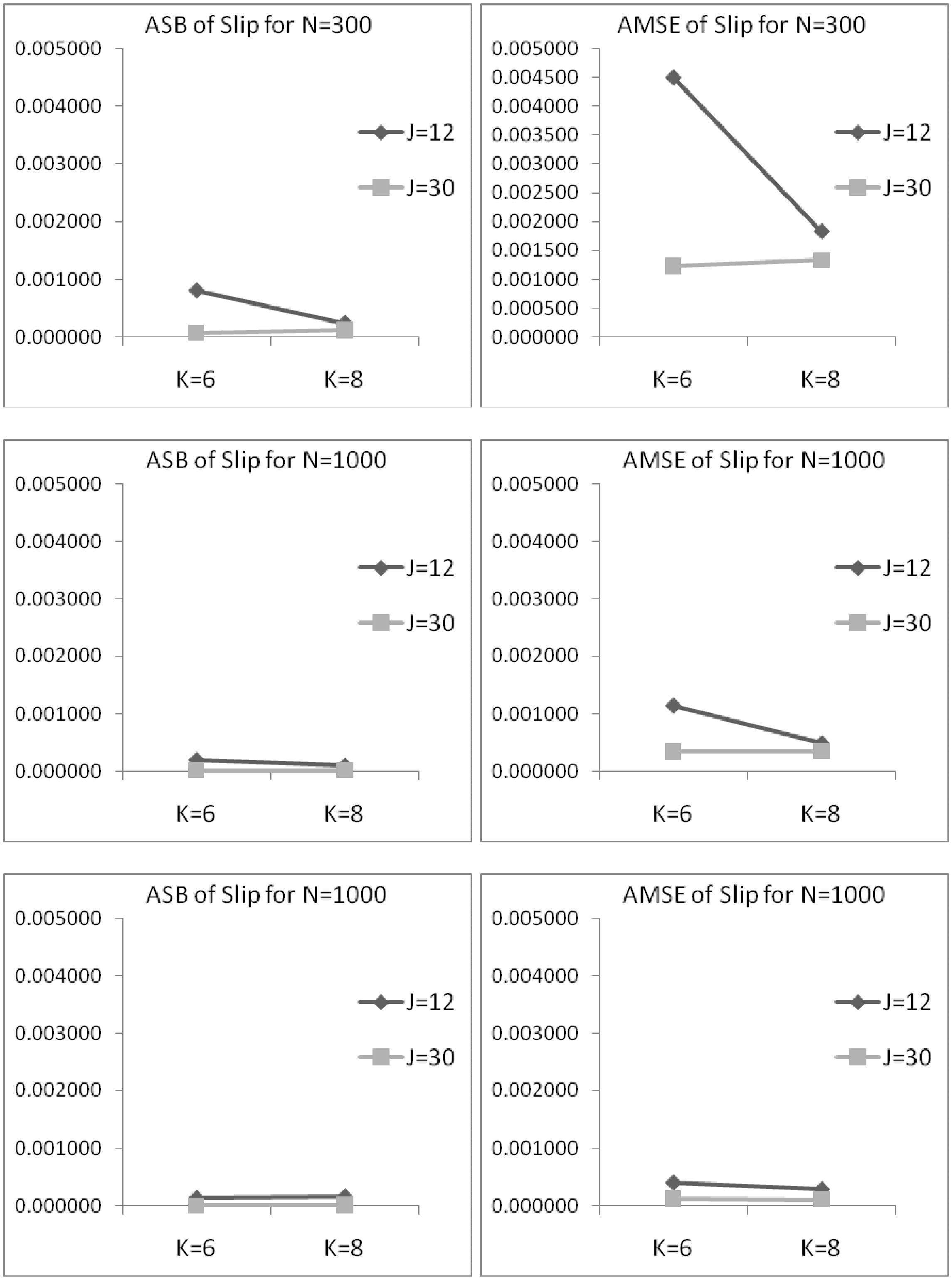


Figure 4-9. Summary statistics for the guessing parameter estimates for the three-way interaction effect of N by J by K for DINO(-H)



ASB of Slip for N=300

AMSE of Slip for N=300

ASB of Slip for N=1000

AMSE of Slip for N=1000

ASB of Slip for N=1000

AMSE of Slip for N=1000

Figure 4-10. Summary statistics for the slip parameter estimates for the three-way interaction effect of N by J by K for DINO(-H)

CHAPTER V

DISCUSSION

This chapter discusses the results of the study and its implications. It contains five main sections. The first section summarizes the results. The second section describes the practical implications. The third section discusses the limitations of the study. The fourth section suggests some future research questions. The fifth section concludes the study.

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 both simulated data and the retrofitted TIMSS 2003 eighth grade mathematics data. The study evaluated the model fit (MAIC and MBIC), item fit (Delta ($\overline{\delta}_j$) and IDI), and item parameter recovery (ASB, AVAR, and AMSE) of slip and guessing for each condition.

In general, the DINA-H and DINO-H models show better model fit, better item fit, and better item parameter recovery than the conventional DINA and DINO models when skills are hierarchically ordered. The misspecification of a skill hierarchy has a negative impact on all models. The item parameters are poorly recovered when a specified skill hierarchy is inconsistent with an estimation model. 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 major findings of the study are summarized in the order of the research questions proposed in Chapter I.

Research Question 1

1. How do the proposed DINA-H and DINO-H perform?

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, whereas the simulation results support that the DINO-H outperforms the DINA-H model. This is related to the different evaluation indices used in the studies. The model and item fits were examined in the real data analysis, while the item parameter recovery was evaluated in the simulation study. The findings of the simulation study that the DINO-H model is better than the DINA-H model when incooperating a skill hierarchy may be due to the assumption of the DINO model. One of the DINO model's main assumptions is to allow the compensation among the mastery of each attribute. This assumption may be more consistent with the hierarchically related cognitive skills because low levels on certain skills could be compensated for by high levels on other skills. 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.

Research Question 2

2. When skills are ordered hierarchically, how do the conventional DINA and DINO models and the proposed DINA-H and DINO-H models compare?

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. The main effect of model consistency from the simulation analysis shows that better results can be obtained when the relationships among attributes specified in the data generating model are consistent with those in the estimation model. Hence, when skills are ordered hierarchically, the proposed DINA-H and DINO-H models should be considered, rather than the conventional DINA and DINO models.

For the main effect of numbers of attributes, the condition of eight attributes recovers item parameter estimates more accurately under both the DINA(-H) and DINO(-H) models, except for the guessing parameter estimates under the DINO(-H) model. For the main effect of test lengths, using 12 items in a test obtains better results for the DINA model, and using 30 items is better for the DINO model. For the main effect of sample sizes, using the larger sample size would result in better item parameter recovery for all models. Generally speaking, having larger sample sizes with more items in a test can obtain more accurate item parameter recovery results. The role of sample size is more important than the role of number of items in reducing estimation error. With smaller sample sizes, the values of summary statistics are more sensitive to the numbers of attributes and test lengths. Using more items measuring fewer attributes displayed more accurate item parameter recovery under the DINO(-H) models. The conditions of larger sample sizes with fewer attributes in the Q-matrix show more accurate item parameter recovery results for the slip parameter estimates under the DINA(-H) model and for the guessing parameter estimates under the DINO(-H) model.

Research Question 3

3. What is the impact of misspecification of a skill hierarchy on each of the four models?

The calibration results of simulation analysis using varying estimation models consistent or inconsistent with the specifications of the skill hierarchies were compared to each other. The item parameters are poorly recovered when a specified skill hierarchy is inconsistent with an estimation model. This situation is especially worse with smaller sample sizes, and fewer items with more attributes in the Q-matrices. The misspecification of a skill hierarchy has a negative impact on all models across most of the conditions of varying numbers of attributes, test lengths, and sample sizes. The DINA-H and the DINO-H models provide stable item parameter estimates even with smaller sample sizes. The results support the assumption that the DINA-H and the DINO-H models, instead of the conventional DINA and DINO models, should be considered when skills are in a certain hierarchical structure, no matter how long the test is, how many attributes are measured, or how small the sample size is.

Implications of the Study

This study is unique in its incorporation of hierarchically structured skills into the estimation process of the conventional DINA/DINO models. In some school subjects, skills are ordered hierarchically. For example, number and operation, algebra, and geometry are not completely independent knowledge segments. The current DINA and DINO models assume independent skills and up to 2^K of attribute profiles, without considering situations where skills are hierarchically related in a certain structure. If the skills are hierarchically related and the conventional models are applied, the parameter estimates are biased and less accurate. There is a need for a model whose specifications, relationships among attributes, possible attribute profiles, and Q-matrix are consistent with the theoretical background and test blueprint. Additionally, the proposed approaches also reduce the number of possible reasonable latent classes in order to promote computing efficiency.

The simulation analysis provides valuable information about the inaccuracy of parameter estimates due to misspecification of the relationships among attributes and possible attribute profiles. The simulation study examined model fit and item parameter recovery when the data simulation models are different from the estimation models. Specifically, both the conventional DINA and DINO models and new hierarchical models are applied when skills are independent or dependent in a specified hierarchical structure. The simulation study confirms that model specification needs to be consistent with the assumptions and characteristics of skills in order to obtain better model fit, item fit, and item parameter recoveries.

The current study contributes to the examination of the performance of the proposed DINA-H and DINO-H models, and provides information about model fit and item parameter recovery under varying conditions of different numbers of attributes, different test lengths, and sample sizes. The results of the study demonstrate the feasibility of the proposed DINA-H and DINO-H models, facilitate the reduction of possible attribute profiles in analyzing a CDM, allow analysis of tests that assess more attributes in the future, promote computational efficiency, and also promote teaching improvement by providing diagnostic information about different cognitive levels and relationships among the skills.

Limitations

This section provides a list of limitations of this study. 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.

Due to the scope of the study, only a few variables and a limited number of conditions for each simulation factor were considered in the simulation design. Future research could incorporate other factors that were not considered in the current study. The comparisons between the DINA/DINA-H and the DINO/DINO-H models are only based on numeric data. In practical situations, it is crucial to consider the rationality of different assumptions of these two models. Additionally, other confounding factors such as the dependency among skills or the response patterns, dimensionalities and item difficulties, are not completely excluded in the analysis.

Future Research Questions

Some suggestions of how to address the limitations in the study and suggestions for future research are discussed below. First, the study only implemented two types of cognitive attribute hierarchies, the linear and unstructured hierarchies. Other types (e.g., the convergent and divergent hierarchies) could also be investigated and compared to the results from the current study. In addition, since a certain type of a hierarchy model could have various types of structures, except for the liner hierarchy model, other possible structures under a hierarchy could also be tried. Especially 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.

CDMs and IRT models differ in their assumptions, the scales representing examinees' true abilities, and the reporting of scoring results. In CDMs, examinees' attribute profiles are categorical latent variables, while IRT-based testing estimates examinees' ability on a continuous scale, the latent trait θ, and orders examinees along a latent dimension (or dimensions). Future research can be conducted on a variety of comparisons between CDMs and IRT models, such as estimation process, retrofitting, or ability estimates. Additionally, unidimensional IRT is commonly used in standardized testing because it is useful for the purposes of scaling and equating. How to implement CDMs in conducting scaling and equating is another direction of future research.

The study implemented the EM algorithm, which is one of the popular estimating methods in CDMs. Other algorithms, for example MCMC, may be applied and compared to the results of using the EM algorithm. Since initializing the EM algorithm requires setting up the starting values of the item parameters that need to be estimated, the impact of varying starting values on the parameter estimates can also be compared and investigated. Additionally, the current study initialized a flat (non-informative) prior, assigning a probability of 1/L to each of the L skill patterns. Future studies can compare different prior distributions or adapt the prior distribution obtained from different models with the same dataset.

In addition to the model fit, item fit, biases, and AMSEs, the differences in scoring and classifying examinee respondents between the conventional models and the proposed hierarchical models can be evaluated. The three common approaches in assigning an examinee into a latent class, MLE, MAP, and EAP, can be compared. As mentioned earlier, the common limitation in current CDMs is that they use 0/1 discrete variables to represent true person profiles, rather than a continuous variable. A different way of score reporting could be to report the probabilities or percentages of mastering each attribute based on the probability from EAP results in future studies. EAP calculates the probabilities of mastery for each attribute for an examinee and sets up a cutoff probability value (usually at 0.5) to classify the attribute into mastery or non-mastery (Huebner, & Wang, 2011). This cutoff can be altered based on different research purposes, for example, setting a lower cutoff value for lower grade students and a higher value for higher grade students. It is also possible to report more than two mastery categories, such as low, medium, high, and advanced levels. Future research can compare various CDMs when using with multiple mastery levels.

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. Future research could also incorporate a number of other factors that were not considered in the current study, for example, different groups of examinee ability levels, different proportions of items measuring varying numbers of attributes, more varieties of Q-matrices and sample sizes as well as different content subjects.

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 profiles, 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 and better item parameter recovery when calibrating items with hierarchically structured attributes and with smaller sample sizes.

The simulation analysis provides valuable information that using larger sample sizes, using more items measuring fewer attributes, and using larger sample sizes with more items in a test produce better item parameter recovery under both the DINA(-H) and DINO(-H) models. The real 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.

Specifying attribute profiles incorrectly would affect the accuracy of estimates of item and attribute parameters. When attributes are hierarchically ordered, the conventional CDMs might produce results that are less accurate. 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.

APPENDIX

Table A1. Results of Fit Indices of J by K for DINA(-H) when N=300

		K=6		K=8	
		MAIC	MBIC	MAIC	MBIC
J=12	HL_300_NA	4274	4596	4485	5519
	HL_300_HL	**4177**	**4288**	**4004**	**4122**
	HL_300_HU	4219	4427	4235	4798
	HU_300_NA	4526	4848	4783	5816
	HU_300_HL	4504	4615	4399	4517
	HU_300_HU	**4472**	**4679**	**4535**	**5098**
	NA_300_NA	**4585**	**4908**	**4750**	**5783**
	NA_300_HL	4562	4673	4358	4476
	NA_300_HU	4561	4768	4516	5079
	Mean	***4431***	***4644***	***4452***	***5023***
J=30	HL_300_NA	9973	10428	10385	10428
	HL_300_HL	**9868**	**10112**	**9913**	**10112**
	HL_300_HU	9915	10256	10137	10256
	HU_300_NA	10573	11029	11066	11029
	HU_300_HL	10960	11204	11128	11204
	HU_300_HU	**10515**	**10856**	**10821**	**10856**
	NA_300_NA	**10755**	**11210**	**11161**	**11210**
	NA_300_HL	11251	11496	11107	11496
	NA_300_HU	11056	11396	11040	11396
	Mean	***10541***	***10888***	***10751***	***10888***

Note. The highlighted values indicated that the DINA(-H) model performs better than its DINO(-H) counterpart (i.e., the smaller MAIC, MBIS, ASB, AVAR, and AMSE in the DINA(-H) model). The bold characters indicate the conditions whose data-generating model is consistent with its estimation model. In these tables, the rows of means show the values of averaging over nine different data generating and estimation conditions. These rules apply to all the following tables.

Table A2. Results of Fit Indices of J by K for DINA(-H) when N=1000

		K=6		K=8	
		MAIC	MBIC	MAIC	MBIC
J=12	HL_1000_NA	13968	14395	13765	15134
	HL_1000_HL	**13864**	**14012**	**13280**	**13437**
	HL_1000_HU	13911	14186	13514	14260
	HU_1000_NA	14875	15302	14909	16278
	HU_1000_HL	14985	15132	14736	14893
	HU_1000_HU	**14820**	**15095**	**14661**	**15407**
	NA_1000_NA	**14927**	**15354**	**14869**	**16239**
	NA_1000_HL	15053	15200	14637	14794
	NA_1000_HU	14970	15245	14656	15402
	Mean	***14597***	***14880***	***14337***	***15094***
J=30	HL_1000_NA	33334	33938	33321	34867
	HL_1000_HL	**33226**	**33550**	**32844**	**33178**
	HL_1000_HU	33278	33730	33071	33994
	HU_1000_NA	33493	34096	35780	37326
	HU_1000_HL	35576	35900	36797	37131
	HU_1000_HU	**33436**	**33888**	**35531**	**36454**
	NA_1000_NA	**34329**	**34932**	**36135**	**37681**
	NA_1000_HL	36615	36939	36807	37140
	NA_1000_HU	35256	35707	36248	37171
	Mean	***34283***	***34742***	***35171***	***36105***

Table A3. Results of Fit Indices of J by K for DINA(-H) when N=3000

		K=6		K=8	
		MAIC	MBIC	MAIC	MBIC
J=12	HL_3000_NA	41732	42255	40304	41979
	HL_3000_HL	**41616**	**41796**	**39806**	**39998**
	HL_3000_HU	41670	42006	40052	40965
	HU_3000_NA	44144	44667	43617	45293
	HU_3000_HL	44629	44810	44004	44196
	HU_3000_HU	**44083**	**44419**	**43366**	**44279**
	NA_3000_NA	**44610**	**45132**	**43496**	**45171**
	NA_3000_HL	45115	45295	43732	43924
	NA_3000_HU	44816	45152	43337	44250
	Mean	***43602***	***43948***	***42413***	***43340***
J=30	HL_3000_NA	98314	99052	98947	100839
	HL_3000_HL	**98203**	**98600**	**98451**	**98859**
	HL_3000_HU	98257	98809	98684	99813
	HU_3000_NA	104076	104814	106353	108245
	HU_3000_HL	108706	109103	110450	110858
	HU_3000_HU	**104016**	**104569**	**106091**	**107220**
	NA_3000_NA	**106359**	**107097**	**107568**	**109460**
	NA_3000_HL	112188	112584	110380	110789
	NA_3000_HU	109814	110366	108365	109494
	Mean	***104437***	***104999***	***105032***	***106175***

Table A4. Summary Statistics for the Guessing Parameter of J by K for DINA(-H) when N= 300

		K=6			K=8		
		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
J=12	HL_300_NA	0.003397	0.001739	0.005136	0.001877	0.001016	0.002893
	HL_300_HL	**0.000012**	**0.002256**	**0.002268**	**0.000146**	**0.000842**	**0.000988**
	HL_300_HU	0.000728	0.003603	0.004331	0.002823	0.001005	0.003827
	HU_300_NA	0.003221	0.007370	0.010591	0.000445	0.003674	0.004119
	HU_300_HL	0.006349	0.014933	0.021282	0.007866	0.002812	0.010677
	HU_300_HU	**0.000091**	**0.008503**	**0.008595**	**0.000395**	**0.003290**	**0.003684**
	NA_300_NA	**0.000134**	**0.003319**	**0.003453**	**0.001058**	**0.003704**	**0.004763**
	NA_300_HL	0.012215	0.006017	0.018232	0.006039	0.001809	0.007847
	NA_300_HU	0.009027	0.003963	0.012990	0.002212	0.003830	0.006042
	Mean	***0.003908***	***0.005745***	***0.009653***	***0.002540***	***0.002442***	***0.004982***
J=30	HL_300_NA	0.000064	0.001263	0.001327	0.000272	0.001150	0.001422
	HL_300_HL	**0.000026**	**0.001197**	**0.001223**	**0.000021**	**0.001065**	**0.001086**
	HL_300_HU	0.000036	0.001228	0.001264	0.000148	0.001110	0.001258
	HU_300_NA	0.000732	0.002633	0.003365	0.000106	0.001261	0.001367
	HU_300_HL	0.018832	0.002913	0.021744	0.008670	0.002341	0.011010
	HU_300_HU	**0.000100**	**0.002023**	**0.002123**	**0.000097**	**0.001240**	**0.001338**
	NA_300_NA	**0.000027**	**0.001115**	**0.001143**	**0.000083**	**0.001267**	**0.001350**
	NA_300_HL	0.017199	0.000786	0.017985	0.007462	0.001562	0.009024
	NA_300_HU	0.010637	0.001015	0.011652	0.000905	0.001588	0.002493
	Mean	***0.005295***	***0.001575***	***0.006870***	***0.001974***	***0.001398***	***0.003372***

Table A5. Summary Statistics for the Guessing Parameter of J by K for DINA(-H) when N= 1000

		K=6			K=8		
		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
J=12	HL_1000_NA	0.003293	0.000539	0.003832	0.001636	0.000254	0.001890
	HL_1000_HL	**0.000156**	**0.000585**	**0.000742**	**0.000202**	**0.000232**	**0.000433**
	HL_1000_HU	0.000754	0.000820	0.001574	0.002784	0.000258	0.003042
	HU_1000_NA	0.005131	0.002071	0.007201	0.000236	0.000892	0.001128
	HU_1000_HL	0.010067	0.003966	0.014033	0.009917	0.000975	0.010893
	HU_1000_HU	**0.000203**	**0.004299**	**0.004502**	**0.000063**	**0.000822**	**0.000885**
	NA_1000_NA	**0.000113**	**0.000828**	**0.000941**	**0.000148**	**0.000979**	**0.001127**
	NA_1000_HL	0.011700	0.005196	0.016896	0.005484	0.000719	0.006203
	NA_1000_HU	0.009219	0.001267	0.010486	0.000548	0.001021	0.001569
	Mean	***0.004515***	***0.002175***	***0.006690***	***0.002335***	***0.000684***	***0.003019***
J=30	HL_1000_NA	0.017897	0.000530	0.018427	0.000163	0.000319	0.000482
	HL_1000_HL	**0.018399**	**0.000477**	**0.018876**	**0.000021**	**0.000302**	**0.000322**
	HL_1000_HU	0.018191	0.000512	0.018703	0.000080	0.000317	0.000396
	HU_1000_NA	0.018062	0.001924	0.019986	0.000022	0.000382	0.000403
	HU_1000_HL	0.028313	0.004503	0.032816	0.009940	0.000625	0.010565
	HU_1000_HU	**0.018382**	**0.001812**	**0.020194**	**0.000022**	**0.000383**	**0.000405**
	NA_1000_NA	**0.018418**	**0.000341**	**0.018759**	**0.000016**	**0.000347**	**0.000363**
	NA_1000_HL	0.033480	0.000854	0.034334	0.006401	0.000643	0.007044
	NA_1000_HU	0.024919	0.000377	0.025296	0.000490	0.000500	0.000990
	Mean	***0.021785***	***0.001259***	***0.023044***	***0.001906***	***0.000424***	***0.002330***

Table A6. Summary Statistics for the Guessing Parameter of J by K for DINA(-H) when N= 3000

		K=6			K=8		
		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
J=12	HL_3000_NA	0.003368	0.000163	0.003530	0.001594	0.000084	0.001678
	HL_3000_HL	**0.000177**	**0.000191**	**0.000368**	**0.000309**	**0.000078**	**0.000387**
	HL_3000_HU	0.000870	0.000238	0.001107	0.002666	0.000084	0.002750
	HU_3000_NA	0.006446	0.000438	0.006884	0.000308	0.000256	0.000564
	HU_3000_HL	0.011147	0.001347	0.012494	0.009083	0.000220	0.009303
	HU_3000_HU	**0.000158**	**0.001240**	**0.001398**	**0.000172**	**0.000259**	**0.000432**
	NA_3000_NA	**0.000158**	**0.000248**	**0.000406**	**0.000197**	**0.000212**	**0.000408**
	NA_3000_HL	0.011791	0.000824	0.012615	0.004665	0.000694	0.005359
	NA_3000_HU	0.009324	0.000261	0.009585	0.000693	0.000263	0.000956
	Mean	***0.004826***	***0.000550***	***0.005377***	***0.002187***	***0.000239***	***0.002427***
J=30	HL_3000_NA	0.000029	0.000134	0.000163	0.000119	0.000115	0.000234
	HL_3000_HL	**0.000004**	**0.000129**	**0.000133**	**0.000009**	**0.000109**	**0.000118**
	HL_3000_HU	0.000012	0.000134	0.000146	0.000057	0.000115	0.000172
	HU_3000_NA	0.000604	0.000372	0.000976	0.000011	0.000123	0.000134
	HU_3000_HL	0.020104	0.000456	0.020560	0.010262	0.000291	0.010553
	HU_3000_HU	**0.000050**	**0.000324**	**0.000375**	**0.000007**	**0.000124**	**0.000131**
	NA_3000_NA	**0.000006**	**0.000112**	**0.000118**	**0.000010**	**0.000115**	**0.000125**
	NA_3000_HL	0.017997	0.000071	0.018068	0.007031	0.000265	0.007296
	NA_3000_HU	0.010636	0.000088	0.010724	0.000537	0.000152	0.000690
	Mean	***0.005493***	***0.000202***	***0.005696***	***0.002005***	***0.000157***	***0.002161***

Table A7. Summary Statistics for the Slip Parameter of J by K for DINA(-H) when N= 300

		K=6			K=8		
		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
J=12	HL_300_NA	0.002540	0.003976	0.006516	0.002314	0.004356	0.006670
	HL_300_HL	**0.000078**	**0.003765**	**0.003843**	**0.000113**	**0.004257**	**0.004370**
	HL_300_HU	0.002468	0.004188	0.006657	0.002429	0.004351	0.006780
	HU_300_NA	**0.001599**	**0.006980**	**0.008579**	**0.003026**	**0.010030**	**0.013056**
	HU_300_HL	0.049749	0.003558	0.053306	0.051879	0.018286	0.070165
	HU_300_HU	**0.000670**	**0.006876**	**0.007547**	**0.002231**	**0.009913**	**0.012144**
	NA_300_NA	**0.001874**	**0.008548**	**0.010421**	**0.002366**	**0.012010**	**0.014376**
	NA_300_HL	0.031799	0.008200	0.039999	0.041014	0.018950	0.059964
	NA_300_HU	0.004519	0.009809	0.014328	0.008539	0.012769	0.021308
	Mean	***0.010588***	***0.006211***	***0.016800***	***0.012657***	***0.010547***	***0.023204***
J=30	HL_300_NA	0.000059	0.002379	0.002439	0.000066	0.002958	0.003024
	HL_300_HL	**0.000034**	**0.002287**	**0.002321**	**0.000047**	**0.002802**	**0.002850**
	HL_300_HU	0.000054	0.002346	0.002400	0.000069	0.002945	0.003015
	HU_300_NA	**0.000065**	**0.002989**	**0.003053**	**0.000402**	**0.005964**	**0.006366**
	HU_300_HL	0.026209	0.004703	0.030912	0.054352	0.010332	0.064684
	HU_300_HU	**0.000059**	**0.002901**	**0.002961**	**0.000372**	**0.005911**	**0.006284**
	NA_300_NA	**0.000129**	**0.003467**	**0.003595**	**0.001098**	**0.007711**	**0.008809**
	NA_300_HL	**0.017853**	**0.006939**	**0.024792**	**0.052134**	**0.010723**	**0.062857**
	NA_300_HU	0.001724	0.006308	0.008032	0.006207	0.009734	0.015941
	Mean	***0.005132***	***0.003813***	***0.008945***	***0.012750***	***0.006565***	***0.019314***

Table A8. Summary Statistics for the Slip Parameter of J by K for DINA(-H) when N= 1000

		K=6			K=8		
		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
J=12	HL_1000_NA	0.002027	0.001311	0.003338	0.001866	0.001106	0.002972
	HL_1000_HL	**0.000020**	**0.001201**	**0.001222**	**0.000014**	**0.001107**	**0.001121**
	HL_1000_HU	0.002086	0.001311	0.003397	0.002094	0.001085	0.003179
	HU_1000_NA	**0.000403**	**0.001521**	**0.001924**	**0.000392**	**0.003079**	**0.003472**
	HU_1000_HL	0.046347	0.000737	0.047084	0.040181	0.004728	0.044909
	HU_1000_HU	**0.000264**	**0.001414**	**0.001678**	**0.000046**	**0.002793**	**0.002839**
	NA_1000_NA	**0.000508**	**0.002798**	**0.003306**	**0.000327**	**0.003815**	**0.004143**
	NA_1000_HL	0.041344	0.003131	0.044474	0.071287	0.005354	0.076641
	NA_1000_HU	0.006881	0.003321	0.010201	0.007458	0.003753	0.011212
	Mean	***0.011098***	***0.001860***	***0.012958***	***0.013741***	***0.002980***	***0.016721***
J=30	HL_1000_NA	0.019765	0.000612	0.020376	0.000022	0.000764	0.000786
	HL_1000_HL	**0.019524**	**0.000601**	**0.020126**	**0.000018**	**0.000742**	**0.000760**
	HL_1000_HU	0.019560	0.000605	0.020164	0.000023	0.000758	0.000781
	HU_1000_NA	**0.019178**	**0.000966**	**0.020143**	**0.000073**	**0.001847**	**0.001921**
	HU_1000_HL	0.097185	0.003159	0.100344	0.054063	0.003146	0.057209
	HU_1000_HU	**0.018939**	**0.000894**	**0.019833**	**0.000088**	**0.001856**	**0.001944**
	NA_1000_NA	**0.019033**	**0.001096**	**0.020129**	**0.000207**	**0.002546**	**0.002754**
	NA_1000_HL	**0.070075**	**0.003641**	**0.073716**	**0.081104**	**0.001756**	**0.082860**
	NA_1000_HU	0.027474	0.001191	0.028665	0.015232	0.002698	0.017930
	Mean	***0.034526***	***0.001418***	***0.035944***	***0.016759***	***0.001791***	***0.018549***

Table A9. Summary Statistics for the Slip Parameter of J by K for DINA(-H) when N= 3000

		K=6			K=8		
		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
J=12	HL_3000_NA	0.001913	0.000367	0.002279	0.001779	0.000327	0.002106
	HL_3000_HL	**0.000012**	**0.000360**	**0.000372**	**0.000007**	**0.000327**	**0.000334**
	HL_3000_HU	0.002024	0.000368	0.002392	0.002030	0.000322	0.002352
	HU_3000_NA	**0.000363**	**0.000493**	**0.000856**	**0.000218**	**0.001027**	**0.001245**
	HU_3000_HL	0.044386	0.000250	0.044636	0.043884	0.001584	0.045468
	HU_3000_HU	**0.000304**	**0.000497**	**0.000800**	**0.000197**	**0.000950**	**0.001146**
	NA_3000_NA	**0.000332**	**0.000752**	**0.001084**	**0.000411**	**0.001406**	**0.001818**
	NA_3000_HL	0.037354	0.000710	0.038064	0.065160	0.004994	0.070154
	NA_3000_HU	0.006457	0.000883	0.007340	0.009519	0.001687	0.011205
	Mean	***0.010349***	***0.000520***	***0.010869***	***0.013689***	***0.001403***	***0.015092***
J=30	HL_3000_NA	0.000009	0.000208	0.000217	0.000016	0.000260	0.000276
	HL_3000_HL	**0.000007**	**0.000206**	**0.000213**	**0.000010**	**0.000256**	**0.000266**
	HL_3000_HU	0.000009	0.000209	0.000218	0.000015	0.000259	0.000274
	HU_3000_NA	**0.000016**	**0.000278**	**0.000294**	**0.000030**	**0.000581**	**0.000611**
	HU_3000_HL	0.024084	0.000438	0.024521	0.068990	0.001090	0.070079
	HU_3000_HU	**0.000021**	**0.000277**	**0.000299**	**0.000021**	**0.000604**	**0.000626**
	NA_3000_NA	**0.000020**	**0.000351**	**0.000371**	**0.000062**	**0.000935**	**0.000997**
	NA_3000_HL	**0.023242**	**0.000930**	**0.024172**	**0.104491**	**0.000572**	**0.105063**
	NA_3000_HU	0.002005	0.000741	0.002746	0.017347	0.000974	0.018321
	Mean	***0.005490***	***0.000404***	***0.005895***	***0.021220***	***0.000615***	***0.021835***

Table A10. Results of Fit Indices of J by K for DINO(-H) when N=300

		K=6		K=8	
		MAIC	MBIC	MAIC	MBIC
J=12	HL_300_NO	4287	4609	4513	5546
	HL_300_HL	**4184**	**4296**	**4030**	**4148**
	HL_300_HU	4231	4439	4265	4828
	HU_300_NO	4450	4772	4410	5443
	HU_300_HL	4386	4497	3971	4089
	HU_300_HU	**4398**	**4605**	**4170**	**4733**
	NO_300_NO	**4646**	**4968**	**4596**	**5630**
	NO_300_HL	4630	4742	4192	4311
	NO_300_HU	4650	4857	4397	4960
	Mean	***4429***	***4643***	***4283***	***4854***
J=30	HL_300_NO	9711	10167	10099	10167
	HL_300_HL	**9606**	**9850**	**9630**	**9850**
	HL_300_HU	9653	9994	9856	9994
	HU_300_NO	10043	10499	9990	10499
	HU_300_HL	10149	10394	9715	10394
	HU_300_HU	**9984**	**10325**	**9767**	**10325**
	NO_300_NO	**10529**	**10984**	**10601**	**10984**
	NO_300_HL	11092	11336	10597	11336
	NO_300_HU	11050	11391	10715	11391
	Mean	***10202***	***10549***	***10108***	***10549***

Table A11. Results of Fit Indices of J by K for DINO(-H) when N=1000

		K=6		K=8	
		MAIC	MBIC	MAIC	MBIC
J=12	HL_1000_NO	13975	14402	13896	15265
	HL_1000_HL	**13869**	**14016**	**13411**	**13568**
	HL_1000_HU	13919	14194	13648	14394
	HU_1000_NO	14482	14909	13464	14833
	HU_1000_HL	14471	14618	13090	13247
	HU_1000_HU	**14426**	**14701**	**13218**	**13964**
	NO_1000_NO	**15238**	**15665**	**14121**	**15490**
	NO_1000_HL	15411	15558	13825	13982
	NO_1000_HU	15335	15610	13975	14721
	Mean	***14570***	***14853***	***13628***	***14385***
J=30	HL_1000_NO	31960	32564	32307	33853
	HL_1000_HL	**31850**	**32174**	**31832**	**32166**
	HL_1000_HU	31903	32355	32062	32985
	HU_1000_NO	33016	33620	32173	33719
	HU_1000_HL	33506	33830	32184	32517
	HU_1000_HU	**32959**	**33410**	**31937**	**32859**
	NO_1000_NO	**34797**	**35401**	**34189**	**35735**
	NO_1000_HL	36843	37167	34990	35323
	NO_1000_HU	36611	37062	34856	35778
	Mean	***33716***	***34176***	***32948***	***33882***

Table A12. Results of Fit Indices of J by K for DINO(-H) when N=3000

		K=6		K=8	
		MAIC	MBIC	MAIC	MBIC
J=12	HL_3000_NO	41648	42170	40634	42310
	HL_3000_HL	**41532**	**41712**	**40141**	**40333**
	HL_3000_HU	41593	41929	40384	41297
	HU_3000_NO	43260	43782	39565	41240
	HU_3000_HL	43391	43571	39384	39576
	HU_3000_HU	**43190**	**43526**	**39306**	**40219**
	NO_3000_NO	**45552**	**46075**	**41744**	**43420**
	NO_3000_HL	46230	46410	41852	42044
	NO_3000_HU	45912	46249	41836	42749
	Mean	***43590***	***43936***	***40538***	***41465***
J=30	HL_3000_NO	95530	96269	95904	97796
	HL_3000_HL	**95408**	**95804**	**95417**	**95825**
	HL_3000_HU	95477	96030	95659	96788
	HU_3000_NO	98748	99486	95734	97626
	HU_3000_HL	100430	100827	96666	97074
	HU_3000_HU	**98691**	**99243**	**95492**	**96622**
	NO_3000_NO	**104008**	**104746**	**101498**	**103390**
	NO_3000_HL	110333	110729	104623	105032
	NO_3000_HU	109560	110112	103878	105008
	Mean	***100909***	***101472***	***98319***	***99462***

Table A13. Summary Statistics for the Guessing Parameter of J by K for DINO(-H) when N= 300

		K=6			K=8		
		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
J=12	HL_300_NO	0.000532	0.003479	0.004011	0.002091	0.007003	0.005736
	HL_300_HL	**0.000034**	**0.003617**	**0.003651**	**0.000101**	**0.003613**	**0.003715**
	HL_300_HU	0.000108	0.003659	0.003767	0.000430	0.003807	0.004237
	HU_300_NO	0.003192	0.018329	0.021521	0.004745	0.029652	0.034397
	HU_300_HL	0.012502	0.018461	0.030963	0.019727	0.030671	0.050398
	HU_300_HU	**0.000565**	**0.017810**	**0.018375**	**0.000826**	**0.033708**	**0.034534**
	NO_300_NO	**0.001275**	**0.009379**	**0.010654**	**0.002257**	**0.013554**	**0.015810**
	NO_300_HL	0.055309	0.022947	0.078256	0.044388	0.018828	0.063216
	NO_300_HU	0.028057	0.011066	0.039123	0.004287	0.025403	0.029691
	Mean	***0.011286***	***0.012083***	***0.023369***	***0.008761***	***0.018471***	***0.026859***
J=30	HL_300_NO	0.000051	0.001875	0.001925	0.000109	0.003322	0.003431
	HL_300_HL	**0.000040**	**0.001842**	**0.001882**	**0.000056**	**0.003093**	**0.003150**
	HL_300_HU	0.000042	0.001863	0.001904	0.000095	0.003175	0.003269
	HU_300_NO	0.000213	0.007059	0.007273	0.022984	0.041719	0.064702
	HU_300_HL	0.017324	0.006002	0.023326	0.041491	0.068823	0.110314
	HU_300_HU	**0.000139**	**0.006686**	**0.006826**	**0.003020**	**0.071821**	**0.074842**
	NO_300_NO	**0.000191**	**0.003900**	**0.004090**	**0.001791**	**0.007732**	**0.009523**
	NO_300_HL	0.048921	0.003659	0.052580	0.060275	0.010228	0.070503
	NO_300_HU	0.029523	0.003665	0.033188	0.008272	0.023970	0.032242
	Mean	***0.010716***	***0.004061***	***0.014777***	***0.015344***	***0.025987***	***0.041331***

Table A14. Summary Statistics for the Guessing Parameter of J by K for DINO(-H) when N= 1000

		K=6			K=8		
		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
J=12	HL_1000_NO	0.000330	0.000992	0.001322	0.001542	0.001095	0.002637
	HL_1000_HL	**0.000025**	**0.001014**	**0.001038**	**0.000025**	**0.001105**	**0.001130**
	HL_1000_HU	0.000096	0.001033	0.001128	0.000208	0.001170	0.001378
	HU_1000_NO	0.001321	0.007538	0.008859	0.010550	0.012147	0.022697
	HU_1000_HL	0.013076	0.006588	0.019664	0.025124	0.013494	0.038617
	HU_1000_HU	**0.000155**	**0.006463**	**0.006618**	**0.000670**	**0.015550**	**0.016220**
	NO_1000_NO	**0.000138**	**0.002577**	**0.002715**	**0.000665**	**0.004423**	**0.005088**
	NO_1000_HL	0.044039	0.009002	0.053041	0.037152	0.005231	0.042384
	NO_1000_HU	0.025310	0.002484	0.027794	0.007288	0.006670	0.013959
	Mean	***0.009388***	***0.004188***	***0.013575***	***0.009247***	***0.006765***	***0.016012***
J=30	HL_1000_NO	0.000023	0.000663	0.000686	0.000042	0.000962	0.001005
	HL_1000_HL	**0.000016**	**0.000652**	**0.000668**	**0.000012**	**0.000936**	**0.000948**
	HL_1000_HU	0.000018	0.000659	0.000677	0.000033	0.000945	0.000978
	HU_1000_NO	0.000091	0.002887	0.002979	0.000765	0.008175	0.008940
	HU_1000_HL	0.025018	0.002443	0.027461	0.022902	0.006440	0.029342
	HU_1000_HU	**0.000102**	**0.002706**	**0.002807**	**0.000199**	**0.006510**	**0.006709**
	NO_1000_NO	**0.000055**	**0.001046**	**0.001101**	**0.000186**	**0.002427**	**0.002612**
	NO_1000_HL	0.049214	0.001225	0.050438	0.052152	0.003728	0.055880
	NO_1000_HU	0.025472	0.001114	0.026587	0.002643	0.008217	0.010860
	Mean	***0.011112***	***0.001488***	***0.012601***	***0.008770***	***0.004260***	***0.013030***

Table A15. Summary Statistics for the Guessing Parameter of J by K for DINO(-H) when N= 3000

		K=6			K=8		
		ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
J=12	HL_3000_NO	0.000250	0.000335	0.000584	0.001483	0.000415	0.001898
	HL_3000_HL	**0.000011**	**0.000339**	**0.000349**	**0.000009**	**0.000403**	**0.000411**
	HL_3000_HU	0.000092	0.000345	0.000437	0.000190	0.000429	0.000619
	HU_3000_NO	0.002955	0.002201	0.005156	0.015813	0.003344	0.019157
	HU_3000_HL	0.013542	0.001919	0.015461	0.025535	0.005295	0.030830
	HU_3000_HU	**0.000266**	**0.001633**	**0.001899**	**0.000544**	**0.006625**	**0.007169**
	NO_3000_NO	**0.000293**	**0.000624**	**0.000917**	**0.000509**	**0.001249**	**0.001758**
	NO_3000_HL	0.035575	0.001412	0.036988	0.048641	0.001992	0.050634
	NO_3000_HU	0.022389	0.000928	0.023318	0.009869	0.002874	0.012744
	Mean	***0.008375***	***0.001082***	***0.009457***	***0.011399***	***0.002514***	***0.013913***
J=30	HL_3000_NO	0.000007	0.000219	0.000226	0.000026	0.000328	0.000353
	HL_3000_HL	**0.000003**	**0.000217**	**0.000221**	**0.000008**	**0.000320**	**0.000329**
	HL_3000_HU	0.000005	0.000219	0.000224	0.000024	0.000323	0.000347
	HU_3000_NO	0.000029	0.000886	0.000914	0.000398	0.004066	0.004465
	HU_3000_HL	0.022249	0.000797	0.023045	0.037135	0.003221	0.040356
	HU_3000_HU	**0.000029**	**0.000881**	**0.000910**	**0.000104**	**0.003317**	**0.003421**
	NO_3000_NO	**0.000026**	**0.000364**	**0.000390**	**0.000220**	**0.001035**	**0.001255**
	NO_3000_HL	0.048375	0.000744	0.049119	0.068665	0.001665	0.070330
	NO_3000_HU	0.025923	0.000481	0.026404	0.009871	0.003825	0.013696
	Mean	***0.010739***	***0.000534***	***0.011273***	***0.012939***	***0.002011***	***0.014950***

Table A16. Summary Statistics for the Slip Parameter of J by K for DINO(-H) when N= 300

		K=6			K=8		
		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
J=12	HL_300_NO	0.003644	0.003135	0.006779	0.000569	0.001014	0.001583
	HL_300_HL	**0.000074**	**0.002323**	**0.002396**	**0.000034**	**0.000903**	**0.000937**
	HL_300_HU	0.003938	0.002219	0.006157	0.000432	0.000952	0.001384
	HU_300_NO	**0.000932**	**0.005221**	**0.006153**	**0.001374**	**0.002122**	**0.003496**
	HU_300_HL	0.009112	0.003976	0.013088	0.006340	0.000702	0.007042
	HU_300_HU	**0.001024**	**0.004030**	**0.005054**	**0.000302**	**0.001600**	**0.001903**
	NO_300_NO	**0.001317**	**0.004723**	**0.006040**	**0.000391**	**0.002243**	**0.002633**
	NO_300_HL	0.012687	0.008243	0.020930	0.006286	0.001182	0.007467
	NO_300_HU	0.011851	0.004376	0.016227	0.004516	0.001782	0.006298
	Mean	***0.004953***	***0.004250***	***0.009203***	***0.002249***	***0.001389***	***0.003638***
J=30	HL_300_NO	0.002135	0.001369	0.003504	0.001110	0.001133	0.002243
	HL_300_HL	**0.000141**	**0.001158**	**0.001300**	**0.000082**	**0.000976**	**0.001058**
	HL_300_HU	0.001808	0.001403	0.003211	0.001174	0.001063	0.002238
	HU_300_NO	**0.000050**	**0.001208**	**0.001258**	**0.000311**	**0.001655**	**0.001966**
	HU_300_HL	0.006301	0.001296	0.007597	0.007964	0.000700	0.008664
	HU_300_HU	**0.000055**	**0.001197**	**0.001252**	**0.000170**	**0.001395**	**0.001565**
	NO_300_NO	**0.000025**	**0.001099**	**0.001125**	**0.000116**	**0.001260**	**0.001376**
	NO_300_HL	**0.012446**	**0.001867**	**0.014313**	**0.007170**	**0.001140**	**0.008310**
	NO_300_HU	0.011547	0.001530	0.013077	0.004283	0.001635	0.005917
	Mean	***0.003834***	***0.001347***	***0.005182***	***0.002487***	***0.001218***	***0.003704***

Table A17. Summary Statistics for the Slip Parameter of J by K for DINO(-H) when N= 1000

		K=6			K=8		
		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
J=12	HL_1000_NO	0.001861	0.000692	0.002553	0.000378	0.000265	0.000643
	HL_1000_HL	**0.000029**	**0.000655**	**0.000684**	**0.000004**	**0.000246**	**0.000249**
	HL_1000_HU	0.002604	0.000599	0.003204	0.000325	0.000264	0.000588
	HU_1000_NO	0.000234	0.001162	0.001396	0.000376	0.000434	0.000810
	HU_1000_HL	0.008830	0.001268	0.010098	0.006340	0.000238	0.006577
	HU_1000_HU	**0.000377**	**0.001053**	**0.001430**	**0.000106**	**0.000381**	**0.000487**
	NO_1000_NO	**0.000192**	**0.001125**	**0.001318**	**0.000186**	**0.000553**	**0.000740**
	NO_1000_HL	0.010002	0.004563	0.014565	0.007629	0.000305	0.007934
	NO_1000_HU	0.011111	0.001404	0.012515	0.002929	0.000580	0.003509
	Mean	***0.003916***	***0.001391***	***0.005307***	***0.002030***	***0.000363***	***0.002393***
J=30	HL_1000_NO	0.001264	0.000381	0.001645	0.000980	0.000286	0.001266
	HL_1000_HL	**0.000021**	**0.000333**	**0.000354**	**0.000035**	**0.000267**	**0.000302**
	HL_1000_HU	0.001093	0.000380	0.001472	0.000993	0.000288	0.001282
	HU_1000_NO	0.000022	0.000341	0.000364	0.000040	0.000436	0.000476
	HU_1000_HL	0.005608	0.000379	0.005987	0.007736	0.000207	0.007943
	HU_1000_HU	**0.000025**	**0.000341**	**0.000367**	**0.000015**	**0.000380**	**0.000395**
	NO_1000_NO	**0.000009**	**0.000313**	**0.000322**	**0.000009**	**0.000346**	**0.000355**
	NO_1000_HL	0.014653	0.000680	0.015333	0.009200	0.000251	0.009451
	NO_1000_HU	0.011660	0.000446	0.012106	0.003718	0.000431	0.004149
	Mean	***0.003817***	***0.000399***	***0.004217***	***0.002525***	***0.000321***	***0.002846***

Table A18. Summary Statistics for the Slip Parameter of J by K for DINO(-H) when N= 3000

		K=6			K=8		
		ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
J=12	HL_3000_NO	0.001732	0.000239	0.001971	0.000299	0.000091	0.000390
	HL_3000_HL	**0.000036**	**0.000218**	**0.000254**	**0.000002**	**0.000084**	**0.000086**
	HL_3000_HU	0.002714	0.000207	0.002921	0.000254	0.000092	0.000345
	HU_3000_NO	0.000279	0.000363	0.000643	0.000434	0.000116	0.000551
	HU_3000_HL	0.008989	0.000433	0.009422	0.006503	0.000070	0.006572
	HU_3000_HU	**0.000245**	**0.000288**	**0.000533**	**0.000290**	**0.000100**	**0.000389**
	NO_3000_NO	**0.000138**	**0.000269**	**0.000407**	**0.000209**	**0.000165**	**0.000374**
	NO_3000_HL	0.010720	0.000725	0.011445	0.008498	0.000104	0.008602
	NO_3000_HU	0.010307	0.000540	0.010848	0.003798	0.000177	0.003975
	Mean	***0.003907***	***0.000365***	***0.004271***	***0.002254***	***0.000111***	***0.002365***
J=30	HL_3000_NO	0.001223	0.000125	0.001347	0.000841	0.000093	0.000934
	HL_3000_HL	**0.000005**	**0.000114**	**0.000118**	**0.000022**	**0.000087**	**0.000109**
	HL_3000_HU	0.001063	0.000123	0.001186	0.000882	0.000091	0.000973
	HU_3000_NO	0.000006	0.000116	0.000122	0.000010	0.000123	0.000133
	HU_3000_HL	0.006359	0.000120	0.006479	0.007751	0.000064	0.007814
	HU_3000_HU	**0.000006**	**0.000116**	**0.000122**	**0.000007**	**0.000113**	**0.000119**
	NO_3000_NO	**0.000006**	**0.000109**	**0.000115**	**0.000004**	**0.000104**	**0.000108**
	NO_3000_HL	0.014463	0.000252	0.014714	0.008934	0.000099	0.009033
	NO_3000_HU	0.011338	0.000161	0.011498	0.003977	0.000154	0.004130
	Mean	***0.003830***	***0.000137***	***0.003967***	***0.002492***	***0.000103***	***0.002595***

Table A19. Differences of Summary Statistics in the Guessing Parameter between the DINA(-H) and DINO(-H) Models for the Interaction Effect of J by K

		K=6			K=8		
	DINO-DINA	ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
J=12	HL_HL - HL_HL	-0.00009	0.00065	0.00055	-0.00018	0.00133	0.00115
	HU_HU -HU_HU	0.00018	0.00396	0.00413	0.00047	0.01717	0.01764
	NO_NO - NA_NA	0.00043	0.00273	0.00316	0.00067	0.00478	0.00545
	Mean	**0.00017**	**0.00245**	**0.00261**	**0.00032**	**0.00776**	**0.00808**
J=30	HL_HL - HL_HL	-0.00612	0.00030	-0.00582	0.00001	0.00096	0.00097
	HU_HU -HU_HU	-0.00609	0.00203	-0.00405	0.00107	0.02664	0.02770
	NO_NO - NA_NA	-0.00606	0.00125	-0.00481	0.00069	0.00315	0.00385
	Mean	**-0.00609**	**0.00119**	**-0.00489**	**0.00059**	**0.01025**	**0.01084**

Table A20. Differences of Summary Statistics in the Slip Parameter between the DINA(-H) and DINO(-H) Models for the Interaction Effect of J by K

		K=6			K=8		
	DINO-DINA	ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
J=12	HL_HL - HL_HL	0.00001	-0.00072	-0.00070	-0.00003	-0.00149	-0.00152
	HU_HU -HU_HU	0.00014	-0.00114	-0.00100	-0.00059	-0.00386	-0.00445
	NO_NO - NA_NA	-0.00035	-0.00199	-0.00235	-0.00077	-0.00475	-0.00553
	Mean	**-0.00007**	**-0.00128**	**-0.00135**	**-0.00046**	**-0.00337**	**-0.00383**
J=30	HL_HL - HL_HL	-0.00646	-0.00050	-0.00696	0.00002	-0.00083	-0.00080
	HU_HU -HU_HU	-0.00631	-0.00081	-0.00712	-0.00010	-0.00216	-0.00226
	NO_NO - NA_NA	-0.00638	-0.00113	-0.00751	-0.00042	-0.00316	-0.00358
	Mean	**-0.00638**	**-0.00081**	**-0.00720**	**-0.00017**	**-0.00205**	**-0.00221**

Table A21. Differences of Summary Statistics in the Guessing Parameter between the DINA(-H) and DINO(-H) Models for the Interaction Effect of N by K

		K=6			K=8		
	DINO-DINA	ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
N=300	HL_HL - HL_HL	0.00002	0.00100	0.00102	0.00000	0.00240	0.00239
	HU_HU -HU_HU	0.00025	0.00699	0.00724	0.00167	0.05050	0.05218
	NO_NO - NA_NA	0.00065	0.00442	0.00507	0.00145	0.00815	0.00961
	Mean	**0.00031**	**0.00414**	**0.00444**	**0.00104**	**0.02035**	**0.02139**
N=1000	HL_HL - HL_HL	-0.00926	0.00030	-0.00896	-0.00009	0.00075	0.00066
	HU_HU -HU_HU	-0.00916	0.00152	-0.00764	0.00039	0.01043	0.01082
	NO_NO - NA_NA	-0.00917	0.00123	-0.00794	0.00035	0.00276	0.00310
	Mean	**-0.00920**	**0.00102**	**-0.00818**	**0.00022**	**0.00465**	**0.00486**
N=3000	HL_HL - HL_HL	-0.00008	0.00012	0.00003	-0.00015	0.00027	0.00012
	HU_HU -HU_HU	0.00005	0.00048	0.00051	0.00023	0.00478	0.00502
	NO_NO - NA_NA	0.00008	0.00031	0.00039	0.00026	0.00098	0.00124
	Mean	**0.00002**	**0.00030**	**0.00031**	**0.00011**	**0.00201**	**0.00213**

Table A22. Differences of Summary Statistics in the Slip Parameter between the DINA(-H) and DINO(-H) Models for the Interaction Effect of N by K

		K=6			K=8		
	DINO-DINA	ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
N=300	HL_HL - HL_HL	0.00005	-0.00129	-0.00123	-0.00002	-0.00259	-0.00261
	HU_HU -HU_HU	0.00018	-0.00228	-0.00210	-0.00107	-0.00641	-0.00748
	NO_NO - NA_NA	-0.00033	-0.00310	-0.00343	-0.00148	-0.00811	-0.00959
	Mean	**-0.00003**	**-0.00222**	**-0.00225**	**-0.00086**	**-0.00570**	**-0.00656**
N=1000	HL_HL - HL_HL	-0.00975	-0.00041	-0.01016	0.00000	-0.00067	-0.00067
	HU_HU -HU_HU	-0.00940	-0.00046	-0.00986	-0.00001	-0.00195	-0.00195
	NO_NO - NA_NA	-0.00967	-0.00123	-0.01090	-0.00017	-0.00273	-0.00290
	Mean	**-0.00961**	**-0.00070**	**-0.01030**	**-0.00006**	**-0.00178**	**-0.00184**
N=3000	HL_HL - HL_HL	0.00001	-0.00012	-0.00011	0.00000	-0.00021	-0.00020
	HU_HU -HU_HU	-0.00004	-0.00019	-0.00022	0.00004	-0.00067	-0.00063
	NO_NO - NA_NA	-0.00010	-0.00036	-0.00047	-0.00013	-0.00104	-0.00117
	Mean	**-0.00004**	**-0.00022**	**-0.00026**	**-0.00003**	**-0.00064**	**-0.00067**

Table A23. Differences of Summary Statistics in the Guessing Parameter between the DINA(-H) and DINO(-H) Models for the Interaction Effect of N by J

		J=12			J=30		
	DINO-DINA	ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
N=300	HL_HL - HL_HL	-0.00001	0.00207	0.00205	0.00003	0.00134	0.00137
	HU_HU -HU_HU	0.00046	0.01986	0.02031	0.00148	0.03762	0.03910
	NO_NO - NA_NA	0.00117	0.00796	0.00912	0.00093	0.00463	0.00556
	Mean	**0.00054**	**0.00996**	**0.01049**	**0.00081**	**0.01453**	**0.01534**
N=1000	HL_HL - HL_HL	-0.00016	0.00065	0.00049	-0.00920	0.00040	-0.00879
	HU_HU -HU_HU	0.00028	0.00845	0.00873	-0.00905	0.00351	-0.00554
	NO_NO - NA_NA	0.00027	0.00260	0.00287	-0.00910	0.00140	-0.00770
	Mean	**0.00013**	**0.00390**	**0.00403**	**-0.00912**	**0.00177**	**-0.00734**
N=3000	HL_HL - HL_HL	-0.00023	0.00024	0.00000	0.00000	0.00015	0.00014
	HU_HU -HU_HU	0.00024	0.00338	0.00362	0.00004	0.00188	0.00192
	NO_NO - NA_NA	0.00022	0.00071	0.00093	0.00011	0.00059	0.00070
	Mean	**0.00008**	**0.00144**	**0.00152**	**0.00005**	**0.00087**	**0.00092**

Table A24. Differences of Summary Statistics in the Slip Parameter between the DINA(-H) and DINO(-H) Models for the Interaction Effect of N by J

		J=12			J=30		
	DINO-DINA	ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
N=300	HL_HL - HL_HL	-0.00005	-0.0024	-0.00244	0.00007	-0.00147	-0.00141
	HU_HU -HU_HU	-0.00079	-0.00557	-0.00637	-0.00011	-0.00311	-0.00321
	NO_NO - NA_NA	-0.00127	-0.0068	-0.00806	-0.00054	-0.00441	-0.00495
	Mean	**-0.00070**	**-0.00492**	**-0.00562**	**-0.00019**	**-0.00300**	**-0.00319**
N=1000	HL_HL - HL_HL	0.00000	-0.0007	-0.0007	-0.00974	-0.00037	-0.01011
	HU_HU -HU_HU	0.00008	-0.00138	-0.0013	-0.00949	-0.00102	-0.01051
	NO_NO - NA_NA	-0.00023	-0.00247	-0.00269	-0.00961	-0.00149	-0.0111
	Mean	**-0.00005**	**-0.00152**	**-0.00156**	**-0.00961**	**-0.00096**	**-0.01057**
N=3000	HL_HL - HL_HL	0.00001	-0.00019	-0.00018	0.00000	-0.00013	-0.00013
	HU_HU -HU_HU	0.00002	-0.00053	-0.00051	-0.00001	-0.00033	-0.00034
	NO_NO - NA_NA	-0.0002	-0.00086	-0.00106	-0.00004	-0.00053	-0.00057
	Mean	**-0.00006**	**-0.00053**	**-0.00058**	**-0.00002**	**-0.00033**	**-0.00035**

Table A25. Differences of Summary Statistics in the Guessing Parameter between the DINA(-H) and DINO(-H) Models for the Three-Way Interaction Effect of N by J by K

		K=6			K=8		
	DINO-DINA	ASB(g)	AVAR(g)	AMSE(g)	ASB(g)	AVAR(g)	AMSE(g)
N=300	J=12	0.00055	0.00558	0.00612	0.00053	0.01435	0.01487
	J=30	0.00007	0.00270	0.00277	0.00156	0.02636	0.02791
N=1000	J=12	-0.00005	0.00145	0.00140	0.00032	0.00635	0.00666
	J=30	-0.01834	0.00059	-0.01775	0.00011	0.00295	0.00306
N=3000	J=12	0.00003	0.00031	0.00033	0.00013	0.00258	0.00270
	J=30	0.00000	0.00030	0.00030	0.00010	0.00144	0.00154

Table A26. Differences of Summary Statistics in the Slip Parameter between the DINA(-H) and DINO(-H) Models for the Three-Way Interaction Effect of N by J by K

		K=6			K=8		
	DINO-DINA	ASB(s)	AVAR(s)	AMSE(s)	ASB(s)	AVAR(s)	AMSE(s)
N=300	J=12	-0.00007	0.00369	-0.00190	0.00024	-0.00569	0.00025
	J=30	-0.00125	0.00146	0.00197	0.00282	-0.00911	-0.01379
N=1000	J=12	-0.00006	-0.00180	-0.00112	-0.00013	-0.00143	-0.00260
	J=30	-0.01915	-0.00086	-0.01970	-0.00010	-0.00137	-0.00180
N=3000	J=12	-0.00008	-0.00028	-0.00035	-0.00004	-0.00078	-0.00082
	J=30	-0.00001	-0.00017	-0.00018	-0.00002	-0.00050	-0.00052

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 & 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

Battista, M. T., & Clements, D. H. (1996). Students understanding of three-dimensional rectangular arrays of cubes. *Journal of Research in Mathematics Education, 27*, 258-292. doi:10.2307/749365

Battista, M. T., & Larson, C. N. (1994). The role of the *Journal of Research in Mathematics Education* in advancing the learning and teaching of elementary school mathematics. *Teaching Children Mathematics, 1*, 178-182.

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

Bruner, J. S. (1966). *Toward a theory of instruction*. Cambridge, MA: Harvard University Press.

Carlin, B., & Louis, A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Chapman & Hall.

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

Choi, K. M. (2011). *What make changes in US 8th graders' mathematics literacy? Using TIMSS 2003 and 2007 via cognitive diagnostic modeling*. Unpublished manuscript. The University of Iowa, IA.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46, 429-449.* doi:10.1111/j.1745-3984.2009.00091.x

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

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35,* 8-24. doi:10.1177/0146621610377081

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Volume 26, Psychometrics* (pp. 979–1030). Amsterdam, The Netherlands: Elsevier. doi: 10.1016/S0169-7161(06)26031-0

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Mahwah, NJ: Erlbaum.

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

Duschl, R. A., Schweingruber, H. A., & Shouse, A. (Eds.). (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8.* Washington, D.C.: National Academy Press.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

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

Gitomer, D. H., & Yamamoto, K. (1991). Performance modeling that integrates latent trait and class theory. *Journal of Educational Measurement, 28,* 173–189. doi:10.1111/j.1745-3984.1991.tb00352.x

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., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*, 262–277. doi:10.1177/0146621604272623

Henson, R. A., Templin, J. L., & Douglas, J. (2007). Using efficient model based sum-scores for conducting skills diagnoses. *Journal of Educational Measurement, 44*, 361-376. doi:10.1111/j.1745-3984.2007.00044.x

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. (2009). *Implementation of the DINA Cognitive Diagnostic Model.* Unpublished manuscript. Iowa city, IA: ACT, Inc.

Huebner, A. (2010). An overview of recent development in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research & Evaluation, 15*, 1-7. http://pareonline.net/getvn.asp?v=15&n=3

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

Kass, R. E. (1993). Bayes factor in practice. *The Statistician*, *42*, 551–560. doi:10.2307/2348679

Kass, R. E., & Raftery, A.E. (1995). Bayes factor. *Journal of the American Statistical Association*, *430*, 773–795. doi:10.2307/2291091

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. (1995). Psychometric latent response models. *Psychometrika*, *60*, 523–547. doi:10.1007/BF02294327

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.

National Mathematics Advisory Panel. (2008). *Foundations for Success: The Final Report of the National Mathematics Advisory Panel*, U.S. Department of Education: Washington, DC. http://www2.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf

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.

Park, Y. S., Lee, Y.-S., & Choi, K. (2010, April). *Cognitive diagnostic analysis of timss 2007 using the DINA model: A multilevel comparison of high, average, and low performance*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

R Development Core Team. (2011). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.Rproject.org

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).

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44,* 293-311. doi:10.1111/j.1745-3984.2007.00040.x

Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response theory models. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment in education: Theory and practice* (pp. 205–241). Cambridge: Cambridge University Press.

Rupp, A. A., & Templin, J. L. (2008a). 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. (2008b). 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., & 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

Shin, N., Stevens, S. Y., Short, H., & Krajcik. J. (2009). *Learning Progressions to Support Coherence Curriculum in Instructional Material, Instruction, and Assessment Design.* Presented at the Learning Progressions in Science Conference. Iowa City, IA.

Smith, C. L., Wiser, M., Anderson, C. W. & Krajcik, J., (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*, 1-98. doi:10.1080/15366367.2006.9678570

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.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

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

Templin, J. L., Henson, R. A., & Douglas, J. (2007). *General theory and estimation of cognitive diagnosis models: Using Mplus to derive model estimates*. National Council on Measurement in Education training session, Chicago, Illinois.

Templin, J. L., Henson, R. A., Templin, S. E., & Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement, 32*, 559-574. doi: 10.1177/0146621607300286

Virginia Department of Education. Licensure Regulations for Mathematics Specialists for elementary and middle education. Retrieved from http://www.doe.virginia.gov/VDOE/Compliance/TeacherED/nulicvr.pdf

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report RR-05-16). Princeton, NJ: Educational Testing Service.

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. (2006). *Cognitive diagnosis for NAEP proficiency data* (Research Report No. RR-06–08). Princeton, NJ: Educational Testing Service.

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.