IDENTIFICATION, ESTIMATION, AND Q-MATRIX VALIDATION OF HIERARCHICALLY STRUCTURED ATTRIBUTES IN COGNITIVE DIAGNOSIS

BY LOKMAN AKBAY

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Education

Written under the direction of

Jimmy de la Torre

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2016

ABSTRACT OF THE DISSERTATION

Identification, Estimation, and Q-matrix Validation of Hierarchically Structured Attributes in Cognitive Diagnosis

by Lokman Akbay

Dissertation Director: Jimmy de la Torre

Many *cognitive diagnosis model* (CDM) examples assume independent cognitive skills; however, cognitive skills need not be investigated in isolation (Kuhn, 2011; Tatsuoka, 1995). Kuhn (2001) argues that some preliminary knowledge can be the foundation for more sophisticated knowledge or skills. When this type of hierarchical relationships among the attributes are not taken into account, estimation results of the conventional CDMs may be biased or less accurate. Hence, this dissertation investigates the change in the degree of accuracy and precision in the item parameter estimates and correct attribute classification rates of different estimation approaches based on modification of either the Q-matrix or prior distribution.

Modification of the prior distribution and the Q-matrix depend on the assumed hierarchical structure, as such, identifying the correct hierarchical structure is of the essence. To address the subjectivity in the conventional methods for attribute structure identification (i.e., expert opinions via content analysis and verbal data

analyses such as interviews and think-aloud protocols), this dissertation proposes a likelihood-ratio test based exhaustive empirical search for identifying hierarchical structures. It further suggests a likelihood-approach for selection of the most accurate hierarchical structure when multiple candidates are present.

Furthermore, implementation of the CDMs requires construction of a Q-matrix to indicate the associations between test items and attributes required for successful completion of the items (de la Torre, 2008; Chiu, 2013). Q-matrix construction heavily depends on content expert opinions, as such this subjective process may result in misspecifications in the Q-matrix. Up to date, several parametric and nonparametric Q-matrix validation methods have been proposed to address the misspecifications that may emerge due to fallible judgments of experts in Q-matrix construction (Chiu, 2013). Yet, although they have been examined under various conditions, none of these methods was tested under hierarchical attribute structures. Therefore, this dissertation further investigates the reciprocal impact of misspecified Q-matrix and hierarchical structure on hierarchy identification and Q-matrix validation.

The results showed that structured prior distribution led to the most accurate and precise item parameter estimation, and highest correct examinee classification. When an unstructured prior was employed, impact of structured Q-matrix was different for compensatory and noncompensatory CDMs. Furthermore, study results showed that likelihood-based exhaustive search was promising in identification/validation of hierarchical attribute structure. Lastly, results indicated that performance of Q-matrix validation methods might not be as high when they are used as is under hierarchical attribute structures.

Acknowledgements

Firstly, I wish to express my sincere thanks to my advisor Dr. Jimmy de la Torre for the continuous support, for his patience, motivation, and immense knowledge. I am grateful for having him as my advisor and mentor who was always helpful and encouraging. Besides my advisor, I would like to thank the rest of my dissertation committee: Dr. Chia-Yi Chiu, Dr. Youngsuk Suh, and Dr. Likun Hou, for their insightful comments.

My sincere thanks also goes to my fellow labmates: Charlie Iaconongelo, Nathan Minchen, Mehmet Kaplan, Ragip Terzi, Wenchao Ma, and others for the stimulating discussions, for peaceful working environment, and for all the fun we have had in the last five years. I also thank my dear friend Umit Atlamaz for his help from grammatical aspect of the study.

Last but not the least, I would like to thank my parents, my brothers, and my sisters, for their unceasing support and encouragement throughout my graduate studies and my life in general. I further take this opportunity to express gratitude to my wife, Yeliz, who was always there throughout this voyage.

Dedication

This dissertation is dedicated to my beloved son – Mehmet Rauf

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

Reasoning processes are generally assessed through complex tasks providing information on reasoning strategies and thinking processes. Significant role of cognitive theory in educational testing has been emphasized in the literature (e.g., Chipman, Nichols, & Brennan, 1995; Embretson, 1985). Embretson (1983) claimed that cognitive theory could improve psychometric practice by guiding the construct representation of test, which is defined by knowledge, mental process, and examinees response strategies. Cognitive requirements eliciting particular knowledge structures, processes, skills, and strategies could be assembled into cognitive models to develop test items (Leighton, Gierl, & Hunka,2004). A generic term *attribute* is used in psychometric literature to refer to cognitive processes, skills, knowledge representations, and problem solving steps that need to be assembled into cognitive models for test development (de la Torre, 2009; de la Torre & Lee, 2010).

After comprehensive examination, Leighton and Gierl (2007) concluded that, among the three types of educational tests (i.e., cognitive model of test specification, cognitive model of domain mastery, and cognitive model of task performance), only cognitive model of task performance was feasible for obtaining convincing evidence for diagnostic inferences on students' cognitive strengths and weaknesses. Assessments based on cognitive model of task performance are usually referred to as cognitively diagnostic assessment (CDA) in the psychometric literature (de la Torre & Minchen, 2014), and aim to identify the attribute mastery status of examinees. CDAs need to be purposefully developed to empirically confirm the examinees' thinking process in

problem solving. Hense, CDMs can also be used to validate specific models of human cognition (Corter, 1995).

Sample tasks administration with standard think-aloud procedure to a representative group of a target population can be useful in CDA development process (Chi, 1997; Taylor & Dionne, 2000). However, the role of cognitive theory needs to be well articulated in test design, only then CDA can prove to be useful in testing practice. Yet, until a few decades ago, the impact of cognitive theory on test design was minimal (Embretson, 1998; National Research Council, 2001), which was attributed by Embretson (1994) to lack of frameworks incorporating cognitive theory in test development. Thereafter, various testing approaches using cognitive theory in psychometric practice have been proposed. *The rule space methodology* [Tatsuoka, 1983], *the attribute hierarchy method* [Leighton et al., 2004], and *the generalized-DINA model framework* [de la Torre, 2011] are among these proposed approaches.

CDA, in general, aim to serve for formative assessment purposes so that feedback obtained from the analysis of the assessment results could be used to modify teaching and learning activities (DiBello & Stout, 2007). Therefore, interest in CDA rapidly increased as the need for formative assessment prompted by the No Child Left Behind Act (2001). This increased interest in CDA gave rise to the developments of statistical models to extract diagnostic information from CDA. These statistical models are restricted latent class models (Templin & Henson, 2006), and referred to as cognitive diagnosis models (CDMs) or diagnostic classification models (DCMs) (de la Torre & Minchen, 2014).

Attribute interaction in response construction and the attributes required for each item are among the features that need to be known to derive a CDM (Chiu, Douglas, & Li, 2009). Thus, a $J \times K$ item-by-attribute specifications matrix, referred to as Q-matrix (Tatsuoka, 1983), is used in CDM. The Q-matrix is usually a binary matrix of J rows and K columns where $j = 1, \ldots, J$ indicates the items and $k =$

$1, \ldots, K$ represents attributes measured by the test. In a Q-matrix, an element q_{jk} is coded as 1 if item j requires attribute k; otherwise, it is coded 0.

Moreover, when K attributes measured through a test, the test could partition examinees' latent ability space into 2^K latent classes. CDM classifies examinees into these latent classes based on examinees' attribute profile estimates. For example, for $K = 3$, an examinee is classified into latent class $\{110\}$ when the examinee has been inferred to have mastered the first two attributes out of three attributes.

It is not uncommon to see prerequisite relationships among attributes. In such cases, mastery of basic attributes is prerequisite for mastering more complex attributes (de la Torre, Hong, & Deng, 2010; Leighton et al., 2004; Templin & Bradshaw, 2014). When attributes have such hierarchical structure, CDMs need to take the hierarchy into account; otherwise, they may not be appropriate and useful (Templin & Bradshaw, 2014). Nevertheless, approaches to incorporate attribute hierarchy into CDM estimation have not been studied thoroughly. Furthermore, identification of hierarchical structures using statistical tests have not explored yet.

1.1 Motivation, Objectives, and Research Questions

Several studies suggest not to investigate cognitive skills in isolation (i.e., Kuhn, 2011; Tatsuoka, 1995). Some basic knowledge can be the foundation for more complex knowledge or skills (Kuhn, 2001). In that vein, attributes measured in educational and psychological assessments may hold a hierarchical structure (Gierl, Wang, & Zhou, 2008; Leighton et al., 2004; Templin & Bradshaw, 2014). Yet, assumption of independent cognitive skills is very common in CDM examples. Use of conventional CMDs with independent attributes assumption may yield biased or less accurate item parameter estimates that may eventually decrease attribute estimation accuracy. Thus, this dissertation investigates the change in the degree of accuracy

and precision in the item parameter estimation and correct attribute classification when either the Q-matrix or the prior distribution is modified by the hierarchical attribute structure.

Under the assumption of hierarchical attribute structure, multiple approaches can be employed in CDMs for item parameter estimation and examinee classification. This study discusses the approaches based on the constrained or unconstrained status of the Q-matrix component of a CDM and the prior distribution used in the estimation algorithm. When prior distribution is unstructured all prior probabilities for 2^K possible latent classes are nonzero. In structured prior distribution case, a prior probability of *zero* can be assigned to the latent classes that are theoretically impossible. Although Q-matrix can also be structured in accordance with the hierarchy, an unanswered question in CDM literature is whether it needs to be. Therefore, the first study in this dissertation presents different estimation approaches using structured and unstructured versions of the Q-matrix and the prior distribution.

The first study of this dissertation is designed to answer the following research questions;

1. Does employment of a structured prior distribution improve item parameter estimation in terms of accuracy and precision?

2. Does employment of a structured Q-matrix improve the item parameter estimation under the structured and unstructured prior distribution cases?

Prior distribution and Q-matrix are structured based on the assumed hierarchical attribute structure. Thus, misspecifications of the prerequisite relationships among the attributes can substantially degrade estimation accuracy. As such, correct hierarchical structure identification is of the essence. In practice, hierarchy derivation procedure relies on either expert opinions via content analysis or verbal data analyses such as interviews and think-aloud protocols (Cui & Leighton, 2009; Gierl

et al., 2008). Because either procedure is subjective, hierarchy derivation may result in disagreements over the prerequisite relationships that yield multiple hierarchies. Moreover, hierarchical structures obtained from *verbal analysis* and *expert opinion* may not be the same (Gierl et al., 2008).

Heretofore, no model based statistical tests were used in attribute hierarchy identification to address the subjectivity in the conventional hierarchy identification methods. To address this subjectivity, the second study of this dissertation proposes a model-fit based empirical exhaustive search method to identify prerequisite relationships among the attributes. Intended use of this method should complement rather than replace the current hierarchy derivation procedures that rely on domain experts' opinions. The second study of the dissertation also explores the viability of statistical model selection methods for hierarchy selection when multiple candidates are present.

The second study is designed to address the following research questions;

1. To what extent is the likelihood-ratio-test based statistical hypothesis testing useful for attribute hierarchy detection and hierarchical structure selection?

2. How viable are the AIC and BIC information criteria for hierarchy selection?

Recall that CDM implementations require construction of a Q-matrix, which indicates the associations between test items and attributes required for successful completion of the items (Chiu, 2013; de la Torre, 2008). Because the Q-matrix integrates cognitive specifications into test construction (Leighton et al., 2004), correct Q-matrix specification is essential to obtain maximum information on the attribute mastery patterns (de la Torre, 2008). However, misspecifications in a Q-matrix may occur due to subjective Q-matrix development procedures, such as expert opinions and verbal data analyses. Up to date, several parametric and nonparametric Q-matrix validation methods have been proposed to address the misspecifications that may emerge due to fallible judgments of experts (Chiu, 2013). Viability of these

Q-matrix validation methods has been tested in variety of conditions using either simulated or real data sets. In these studies, attributes are assumed to be either independent (e.g., 2^K attribute patterns are uniformly distributed) or dependent such that there is correlational or higher order relationships between the attributes. Because none of these methods was tested under truly hierarchical attribute conditions, this study aims to examine their performances under such conditions.

The second study of this dissertation implicitly assumes that Q-matrix used in hierarchy identification and/or selection was correctly specified. However, the Q-matrix used for structure selection may not be correct. Therefore, the third study complicates the structure selection by involving a misspecified Q-matrix. It also examines the impact of selected hierarchical structure on empirical Q-matrix validation procedures. Thus, the third study is concerned with two problems: (1) misspecifications in Q-matrix may adversely affect correct hierarchical structure selection, and (2) Q-matrix validation procedures may not yield accurate results if the hierarchical structure to begin with is misspecified. Therefore, examining and reporting the reciprocal impact of misspecified Q-matrix and hierarchical structure on hierarchy identification and Q-matrix validation are among the objectives of this dissertation.

The third study of this dissertation is designed to answer the following research questions;

1. To what extent does a misspecified Q-matrix degrade the hierarchical attribute structure detection and structure selection?

2. To what extent does an inaccurate attribute structure have impact on Q-matrix validation?

1.2 References

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The journal of the learning sciences, 6,* 271-315.

Chipman, S. F., Nichols, P. D., & Brennan, R. L. (1995). Introduction. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 1-18). Hillsdale, NJ: Erlbaum.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37,* 598-618.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74,* 633-665.

Corter, J. E. (1995). Using clustering methods to explore the structure of diagnostic tests. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 305-326). Hillsdale, NJ: Erlbaum.

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

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA Model: Development and applications. *Journal of educational measurement, 45,* 343-362.

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

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76,* 179-199.

de la Torre, J. (2014, June). Cognitive diagnosis modeling: A general framework approach. Workshop conducted on the 4th congress on Measurement and Evaluation in Education and Psychology, Ankara, Turkey.

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter

estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47,* 227-249.

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.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicologia Educative, 20,* 89-97.

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44,* 285-291.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107-135). New York: Plenum Press.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Embretson, S .E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics.* Orlando, FL: Academic Press.

Gierl, M. J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *Journal of Technology, Learning, and Assessment, 6* (6), 1-53.

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. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking

processes. *Educational Measurement: Issues and Practice, 26 (2),* 3-16.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuokas rule-space approach. *Journal of Educational Measurement, 41,* 205-237.

National Research Council (2001). *Knowign what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, and R. Glaser (Eds.). Board on Testing and Assessment, Center for Education. Washington, DC: National Academy Press.

No Child Left Behind Act of 2001, Pub. L. No. 1-7-110 (2001).

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345-354.

Tatsuoka, K. K. (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.

Taylor, K. L., & Dionne, J. P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92,* 413-425.

Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*, 317-339.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods, 11(3),* 287-305.

Chapter 2

Approaches to Estimating Hierarchical Attribute Structures

2.1 Introduction

Assessment of reasoning processes usually requires complex tasks that provide information about reasoning strategies and thinking processes. Educational measurement professionals have underscored the appreciable role of cognitive theory in educational testing (e.g., Chipman, Nichols, & Brennan, 1995; Embretson, 1985). Because knowledge, mental processes, and examinees' response strategies define construct representation, Embretson (1983) asserted that cognitive theory could improve psychometric practice by guiding the construct representation of a test. Leighton, Gierl, and Hunka (2004) argued that the cognitive requirements eliciting particular knowledge structures, processes, skills, and strategies could be assembled into cognitive models that are then used to develop test items. In the psychometric literature, the generic term *attributes* is used to refer to cognitive processes, skills, knowledge representations, and problem solving steps to be assembled into cognitive models for test development (de la Torre, 2009b; de la Torre & Lee, 2010).

Assessments designed and developed for identifying attribute mastery status of examinees to obtain convincing evidence for diagnostic inferences about students' cognitive strengths and weaknesses are referred to as cognitively diagnostic assessment (CDA: de la Torre & Minchen, 2014). CDAs need to be purposefully developed to

empirically confirm the examinees' thinking process in problem solving. These well-designed diagnostic tests can also help validating specific models of human cognition (Corter, 1995).

Administering sample tasks to a representative group of a target population with a standard think-aloud procedure can help the development process of a CDA (Chi, 1997; Taylor & Dionne, 2000). However, for CDA to impact testing practice, the role of cognitive theory needs to be well articulated in test design. Yet, until quite recently, the impact of cognitive theory on test design was minimal (Embretson, 1998; National Research Council, 2001). This minimal impact was attributed by Embretson (1994) to lack of frameworks that use cognitive theory in test development. Recently, various approaches integrating cognitive theory into psychometric practice have been proposed (e.g., *the rule space methodology* [Tatsuoka, 1983], *the attribute hierarchy method* [Leighton et al., 2004], and *the generalized-DINA model framework* [de la Torre, 2011]). Some of these approaches are purely psychometric in nature while others are not (de la Torre, 2014).

The CDA usually serve for formative assessment purposes and teaching and learning activities can be modified in accordance with the crucial feedback obtained from analysis of the assessment results (DiBello & Stout, 2007). Thus, the popularity of CDA rapidly increased as the need for formative assessment prompted by recent political changes including the No Child Left Behind Act (2001). Thereafter, quite a few statistical models that are used to extract diagnostic information from CDA, which are referred to as cognitive diagnosis models (CDMs) or diagnostic classification models (DCMs) (de la Torre & Minchen, 2014), were proposed. These models are the restricted latent class models (Templin & Henson, 2006).

Underlying assumptions to derive a CDM requires two features to be known: the attribute interaction in the response construction process and the attributes that are needed for each item (Chiu, Douglas, & Li, 2009). Therefore, in CDMs, a JxK

matrix, referred to as Q-matrix (Tatsuoka, 1983), is used to set item-by-attribute specifications. The Q-matrix is a binary matrix of J rows and K columns where $j = 1, \ldots, J$ indicates the items and $k = 1, \ldots, K$ represents attributes measured by the test. Item j requires examinees to possess attribute k for success if q_{jk} element of the matrix is coded as 1. When q_{jk} is 0, it means that kth attribute is not necessary for solving item j.

In some cases, attributes may have hierarchical structure such that mastery of basic attributes is prerequisite for mastering more complex attributes (de la Torre, Hong, & Deng, 2010; Leighton et al., 2004; Templin & Bradshaw, 2014). In such cases, CDMs need to take the hierarchical structure into account; otherwise, they may not be appropriate and useful (Templin & Bradshaw, 2014). Nevertheless, many CDM examples assume independent cognitive skills. Hence, this dissertation investigates the change in the degree of accuracy and precision in the item calibration and correct attribute classification rate when either the Q-matrix or the prior distribution is modified in accordance with the hierarchical attribute structure.

When attributes are hierarchical, several approaches under CDMs can be employed for model parameter estimation and attribute classification. The approaches discussed in this study are based on the constraint or unconstraint status of the Q-matrix component of a CDM and the prior distribution to be used in the estimation algorithm. For an unstructured prior distribution all prior probabilities for 2^K numbers of possible latent classes are nonzero. In other situations, a prior probability of zero can be assigned in a structured prior distribution for latent classes that are theoretically impossible while nonzero prior probabilities are assigned to the permissible classes. The Q-matrix can also be structured in accordance with the hierarchy. Therefore, an unanswered question in CDM literature is concerned with whether the Q-matrix needs to be structured in accordance with the hierarchical structure of the attributes. From this point of view, this study presents different estimation

approaches based on structured or unstructured status of the Q-matrix and the prior distribution.

2.2 Background

Within the cognitive diagnosis modeling (CDM) literature, various specific and general models with different underlying assumptions about the relationships between the attributes and test performances have been developed. De la Torre (2011) shows that the commonly used specific CDMs are special cases of the general models. For example, the *generalized deterministic inputs, noisy "and" gate* (G-DINA; de la Torre, 2011) model is one of the general cognitive diagnosis models, from which the *deterministic input, noisy "and" gate* (DINA; de la Torre, 2009b, Junker and Sijtsma, 2001), *deterministic input, noisy "or" gate* (DINO; Templin and Henson, 2006), and *additive-CDM* (A-CDM; de la Torre, 2011), among others, can be derived.

The IRF of the *generalized-DINA model* (G-DINA; de la Torre, 2011) under the *identity link* is

$$P(\alpha_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \qquad (2.1)$$

where K_j^* represents the number of required attributes for the j^{th} item (notice that K_j^* is item specific and does not represents the total number of attributes measured by a test); l represents a particular attribute pattern out of $2^{K_j^*}$ possible attribute patterns; δ_{j0} is the intercept for the item j; δ_{jk} is the main effect due to α_k; $\delta_{jkk'}$ represents interaction effect due to α_k and $\alpha_{k'}$; and $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$ (de la Torre, 2011). Therefore, the G-DINA model splits examinees into $2^{K_j^*}$ latent groups for item j based on the probability of answering item j correctly.

DINA Model: In the psychometric literature, due to containing only two item parameters (i.e., guessing and slip), the DINA model is known as one of the most parsimonious and interpretable CDMs (de la Torre, 2009b). The DINA model is also known as a conjunctive model (de la Torre, 2011; de la Torre & Douglas, 2004), which assumes that missing one of the several required attributes for an item is the same as having none of the required attributes (de la Torre, 2009b; Rupp & Templin, 2008). This assumption can be statistically represented by the *conjunctive condensation function* (Maris, 1995, 1999). Given an examinee is in a particular latent class, $\boldsymbol{\alpha}_l$, and the j^{th} row of the Q-matrix (i.e., attribute specification of j^{th} item) the conjunctive condensation rule generates a group-specific deterministic response ($\eta_{lj} = 1$ or 0) through the function

$$\eta_{lj} = \prod_{k=1}^{K} \alpha_{lk}^{q_{jk}}. \tag{2.2}$$

Moreover, the probabilistic component of the item response function (IRF) of the DINA model allows the possibility of *slipping* on an item when an examinee possesses all the required attributes for it. Likewise, The IRF also allows the possibility that an examinee lacking at least one of the required attributes can *guess* the item. The probabilities of slipping and guessing for item j are denoted as $s_j = P(X_{ij} = 0|\eta_{ij} = 1)$ and $g_j = P(X_{ij} = 1|\eta_{ij} = 0)$, respectively, where X_{ij} is the observed response of examinee i to item j. Given s_j and g_j, the IRF of the DINA model is written as

$$P(X_j = 1|\boldsymbol{\alpha}_l) = P(X_j = 1|\eta_{jl}) = g_j^{(1-\eta_{jl})}(1 - s_j)^{\eta_{jl}} \tag{2.3}$$

where $\boldsymbol{\alpha}_l$ is attribute pattern l among 2^K possible attributes patterns; η_{jl} is the expected response of an examinee to item j who possesses attribute pattern l; and g_j and s_j are guessing and slip parameters, respectively (de la Torre, 2009a). Notice

that g_j and $(1-s_j)$ correspond to δ_{j0} and $\delta_{j12K_j^*}$, respectively, in the G-DINA model representation. Thus, the G-DINA reduces to the DINA model by setting all the parameters but δ_{j0} and $\delta_{j12\ldots K_j^*}$ to zero.

DINO Model: The DINO model is the disjunctive counterpart of DINA model with the assumption that having one of the several required attributes is the same as having more than one or all required attributes to answer an item successfully (Rupp & Templin, 2008; Templin & Rupp, 2006). Due to disjunctive nature of the model, given an examinee's latent class, $\boldsymbol{\alpha}_l$, and the j^{th} row of the Q-matrix, the group-specific deterministic response (i.e., $\omega_{lj} = 1$ or 0) for the model is obtained by the function

$$\omega_{lj} = 1 - \prod_{k=1}^{K}(1-\alpha_{lk})^{q_{jk}}. \tag{2.4}$$

As such, the DINO model also splits examinees into two groups: One group consists of examinees possessing at least one of the required attributes for the item, and another group consists of examinees who mastered none of the required attributes.

Similar to the DINA model, the DINO model also has two item parameters; $s_j^* = P(X_{ij}=0|\omega_{ij}=1)$ and $g_j^* = P(X_{ij}=1|\omega_{ij}=0)$, where $1-s_j^*$ is the probability that examinee i correctly answers item j given that the examinee has mastered at least one of the required attributes, and g_j^* is the probability that examinee i correctly answers item j when the examinee has not mastered any required attribute. The item response function of the DINO model is

$$P(X_j=1|\boldsymbol{\alpha}_l) = P(X_j=1|\omega_{jl}) = {g_j}^{(1-\omega_{jl})}(1-s_j)^{\omega_{jl}} \tag{2.5}$$

where $\boldsymbol{\alpha}_l$ is attribute pattern l; ω_{jl} is the expected response of an examinee to item j who possesses attribute pattern l; and g_j^* and s_j^* are guessing and slip parameters for item j, respectively (Templin & Rupp, 2006).

As mentioned earlier, the DINO model can also be derived from the G-DINA model. This is attained by setting $\delta_{jk} = -\delta_{jk'k''} = \ldots = (-1)^{K_j^*+1}\delta_{j12\ldots K_j^*}$ in the G-DINA model (de la Torre, 2011). In words, the G-DINA model is reduced to the DINO by constraining the main and the interaction effects to be equal with alternating sign, thus, that allowing only two probabilities: $\delta_{j0} = g_j^*$ and $\delta_{j0} + \delta_{jk} = 1 - s_j^*$.

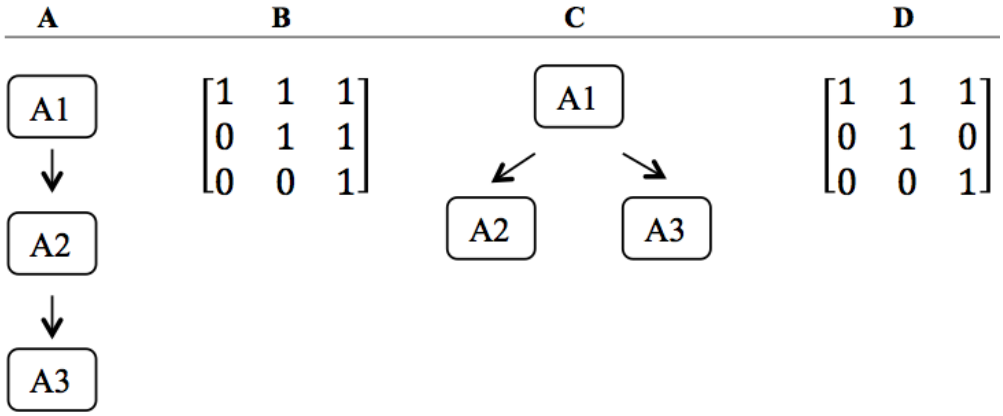
2.3 Estimation Approaches

2.3.1 Permissible Latent Classes and Structured Prior Distribution

When all attributes are independent (i.e., mastery of one attribute is not prerequisite for mastering another attribute) the latent classes are unstructured and all of the 2^K latent classes are permissible. Conversely, when the attributes are dependent with respect to some hierarchical relations, the latent classes are referred to as (*hierarchically*) *structured*, where some constraints defining impossible latent classes exist (de la Torre et al., 2010; Leighton et al., 2004). For example, based on the linear hierarchical structure among three attributes in Figure 2.1A, the latent classes 000, 100, 110, and 111 are permissible, whereas 010, 001, 101, and 011 patterns are not. That is, attribute patterns having an attribute without possessing the prerequisite(s) are not allowed. Likewise, in the divergent structure depicted in Figure 2.1C, the latent classes 000, 100, 110, 101, and 111 exist; yet, 010, 001, and 011 patterns do not. In other words, because mastering the second and third attributes requires mastery of the first attribute, no examinee could have an attribute profile of 010, 001, or 011 in the divergent structure.

De la Torre et al. (2010) found that empirical Bayes estimation method, in general, outperforms fully Bayes estimation method. It should be noted here that the difference between the fully and empirically Bayes methods lies in the prior weights

Figure 2.1: Hierarchies with Corresponding R-Matrices



which are updated in the empirical Bayes method after each iteration, whereas they remain fixed in the fully Bayes method. Although we cannot precisely know the distribution of the permissible latent classes, we can assign a prior probability of *zero* to theoretically non-existent classes by careful consideration of the hierarchical structure of the attributes. Therefore, imposing *zero prior probabilities* for the latent classes that are not permissible within a particular hierarchy yields a structured prior distribution.

2.3.2 Structured and Unstructured Q-matrix

In cases where attributes are ordered, test items may or may not *explicitly* require more basic attributes (i.e., prerequisites) along with a complex one for successful completion (de la Torre et al., 2010; Leighton et al., 2004). One example is given in de la Torre et al. (2010) where they argue that 'taking the derivative' presupposes 'knowledge of basic arithmetic operation', yet, an item can be constructed such that it solely requires ability to differentiate without the need for basic arithmetic operations. Thus, a Q-matrix for noncompensatory models (e.g., DINA) can be designed such that it follows the hierarchical structure of the attributes where, all more basic attributes, even though they are not explicitly probed by the item, are represented by

1. In contrast, a Q-matrix can also be designed such that only attributes explicitly needed for successful completion are specified. In this manuscript, these two types of Q-matrices are referred to as *structured* or *implicit Q-matrix*, and *unstructured* or *explicit Q-matrix*, respectively.

To demonstrate the differences in the two Q-matrices, an item requiring only the third attribute out of the three in Figure 2.1A can be represented as 001 in an explicit Q-matrix, whereas it is specified as 111 in an implicit counterpart. Note that because the latent classes 001, 101, and 011 are among the nonexistent classes when the linear hierarchy applies, regardless of the Q-matrix types, the item can only be correctly answered by the examinees possessing all three attributes.

Conversely, for compensatory models, the Q-matrix is also structured such that only the most basic attribute is specified even though the item also probes the more complex ones. For instance, when three attributes are linearly hierarchical, where A1 is the most basic attribute and A3 is the most complex attribute, an item probing all three attributes can be represented as either 100 by an implicit Q-matrix, or 111 by an explicit Q-matrix. Likewise an item requiring the second and third attributes is specified as 010 in an implicit Q-matrix, whereas an explicit Q-matrix specifies both of the second and third attributes (i.e., 011). Notice that, following the DINO model, examinees must have mastered at least the second attribute to successfully complete this item. Then, regardless of the type of the Q-matrix, this item would be answered correctly by the examinees that are either in the 110 class or in the 111 class among the permissible latent classes.

Although examinee responses are not affected by how a Q-matrix is constructed, impact of item-attribute specification (i.e., implicit vs. explicit) on estimation of hierarchically ordered attributes has not been studied in the CDM literature. Hence, to fill this gap, the current study aims to investigate whether designing a Q-matrix corresponding to hierarchical structure (i.e., implicit Q-matrix) improves the

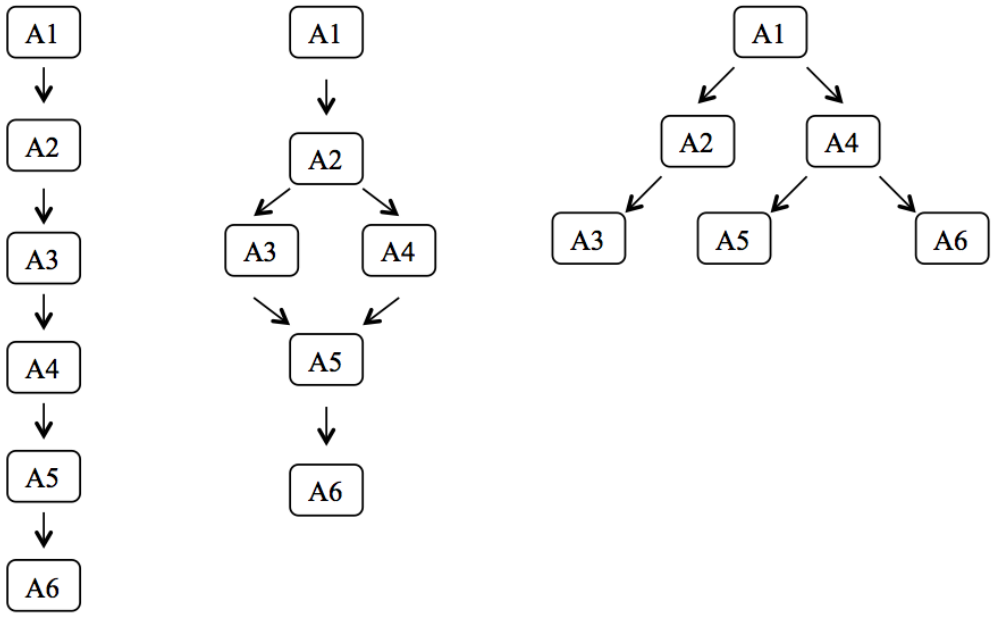
parameter estimation and attribute classification.

2.4 Simulation Study

A simulation study was designed to understand the impact of *structured Q-matrix*, if any, on item parameter estimation and attribute classification when attributes are hierarchical. To accomplish this, the unstructured and structured versions of the Q-matrices were crossed with the unstructured and structured forms of latent classes. This resulted in four different approaches that can be employed in CDM estimations. Three general attribute hierarchy types (i.e., linear, convergent, and divergent) consisted of six attributes, as defined by Leighton et al. (2004), were considered. These structures are illustrated in Figure 3.1 and the permissible latent classes under each hierarchy are given in Appendix 2A. For instance, in the linear case, instead of 2^6 attribute patterns, only seven attribute patterns (i.e., 000000, 100000, 110000, 111000, 111100, 111110, and 111111) are permissible. Likewise, 12 and 16 latent classes exist when six attributes follow the given convergent and divergent structures, respectively.

The unstructured Q-matrix and its structured counterparts used throughout the study are given in Appendix 2B. Unstructured Q-matrix consisted of items requiring one, two, and three attributes. Although the first 18 items designed to measure each of the six attributes equally, two more items (i.e., item 19 and item 20) were added to have at least two items differentiating adjacent latent classes (e.g., 000000 and 100000, and 111110 and 111111) when the Q-matrix is structured. Ideal response patterns corresponding to the unstructured Q-matrix are given in Appendices 2C and 2D, which shows that all permissible latent classes are identifiable. Furthermore, the impact of the estimation approaches were studied in different hierarchical conditions where three levels of item quality and two generating models (i.e., DINA and DINO) were employed. The item quality was defined by the item discrimination (i.e., $1-s-g$)

Figure 2.2: Linear, Convergent, and Divergent Hierarchies Defined by Leighton et al. (2004)



as higher, lower, and mixed item qualities.

Three-levels of item quality, two CDMs (i.e., the DINA and DINO models), three general hierarchy types, and two different sample sizes were crossed to form the simulation conditions. For the higher-quality (HQ) items, the lowest and highest success probabilities (i.e., $P(\mathbf{0})$ and $P(\mathbf{1})$) were generated from $U(0.05, 0.20)$ and $U(0.80, 0.95)$, respectively. For the lower-quality (LQ) items, the lowest and highest success probabilities were drawn from $U(0.15, 0.30)$ and $U(0.70, 0.85)$, respectively. In other words, the slip and guessing parameters to generate the data were drawn from $U(0.05, 0.20)$ and $U(0.15, 0.30)$ for higher and lower item quality conditions, respectively. Additionally, for mixed item quality conditions, lowest and highest success probabilities were drawn from $U(0.05, 0.30)$ and $U(0.70, 0.95)$, respectively. Attributes were generated following the linear, convergent, and divergent hierarchies. The sample size also had two levels (i.e., $N = 500$ and $N = 1000$ examinees). In all conditions, the test length and number of attributes measured were fixed to twenty and six, respectively. Moreover, the number of replication for each condition was

Table 2.1: Simulation Factors in the Study

Type of CDM	Prior Distribution	Type of Q-matrix	Type of Hierarchy	Item Quality	Sample Size
DINA	Structured	Explicit	Linear	Higher Quality	500
DINO	Unstructured	Implicit	Convergent	Mixed Quality	1000
			Divergent	Lower Quality	

Note. CDM = cognitive diagnosis model; DINA = deterministic input, noisy "and" gate model; DINO = deterministic input, noisy "or" gate model.

fixed to 100. Throughout the study data generation and model estimation performed using the OxMetrics programming language (Doornik, 2011). All the factors with varying levels are summarized in Table 2.1. Item parameter estimation was carried out with marginal maximum likelihood estimator (MMLE) via expectation-maximization (EM) algorithm and attribute estimation was based on expected a posteriori (EAP) estimator.

2.5 Simulation Results

To determine the impact of the Q-matrix design on item parameter estimation accuracy and precision, the *bias* and *the root mean squared error (RMSE)* of the estimates across 100 replications were computed. The bias and RMSE for guessing are defined as

$$bias_{g_j} = \frac{1}{R}\sum_{r=1}^{R}(\hat{g}_{jr} - g_{jr}),$$

$$RMSE_{g_j} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{g}_{jr} - g_{jr})^2},$$

respectively, where R is the number of replications in each condition, $\hat{g}_{jr}$ is the guessing parameter estimate for item j in replication r, g_{jr} is the generating guessing parameter for item j in replication r. Notice that the same formulas can be used for slip parameter, where g is replaced by s.

The correct attribute classification rates at the individual-attribute level (i.e., *correct attribute classification rate; CAC*) and at the attribute-vector level (i.e., *correct vector classification rate; CVC*) were also investigated. The CAC and CVC can be computed using the formulae

$$CAC_k = \sum_{r=1}^{R}\sum_{i=1}^{N}\frac{I|\hat{\alpha}_{ik}^r = \alpha_{ik}^r|}{NR}, \tag{2.6}$$

and

$$CVC = \sum_{r=1}^{R}\sum_{i=1}^{N}\frac{I|\hat{\boldsymbol{\alpha}}_{i}^r = \boldsymbol{\alpha}_{i}^r|}{NR}, \tag{2.7}$$

respectively, where N is the total number of examinees, R is the total number of replications, I is the indicator function, α_{ik}^r is true mastery status of examinee i for attribute k in replication r, and $\hat{\alpha}_{ik}^r$ is the expected a posteriori (EAP) estimate of examinee i for attribute k in replication r, $\boldsymbol{\alpha}_i^r$ is generating attribute pattern of examinee i in replication r, and $\widehat{\boldsymbol{\alpha}}_i^r$ is the estimated attribute pattern for the same examinee in the same replication. Moreover, the false-positive and false-negative classification rates resulting from two distinct ways of Q-matrix specification under structured and unstructured versions of prior distribution were reported.

2.5.1 DINA Model Results

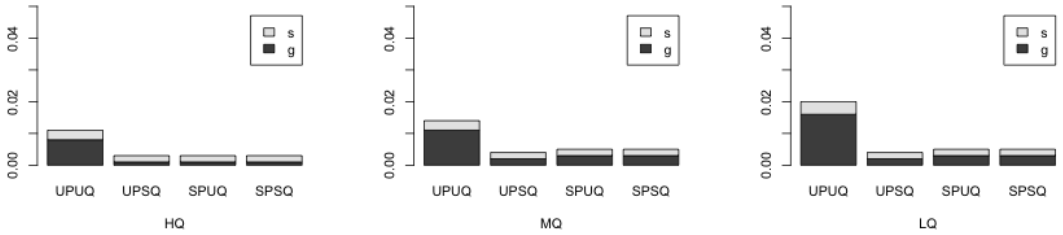
It should be noted here that, due to space constraint, the simulation results based on the DINA model will be discussed in detail, and only DINO model results that depart significantly from the DINA model counterpart will be emphasized.

Effect of Implicit Q-matrix on Item Parameter Estimation

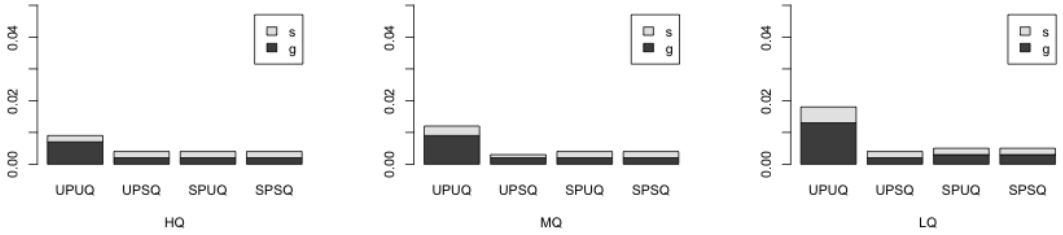
Figure 2.3 and Figure 2.4 depict the mean absolute item parameter bias observed by employment of different estimation approaches when samples consist of

1000 and 500 examinees, respectively. Similarly, mean RMSEs obtained via four estimation approaches when the sample sizes are 1000 and 500 are given in Figures 2.5 and 2.6, respectively. In all four figures, the upper panels indicate the mean absolute bias and RMSE obtained when attributes are linearly hierarchical. Likewise, the middle and lower panels show the bias and RMSE results for the convergent and

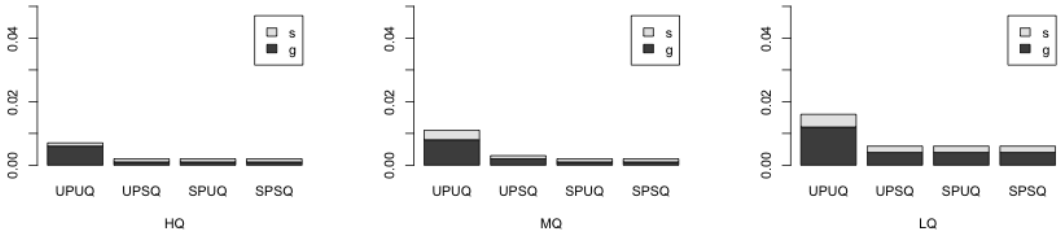
Figure 2.3: Mean Absolute Item Bias of the Estimation Approaches: $N = 1000$



(a) Linear



(b) Convergent

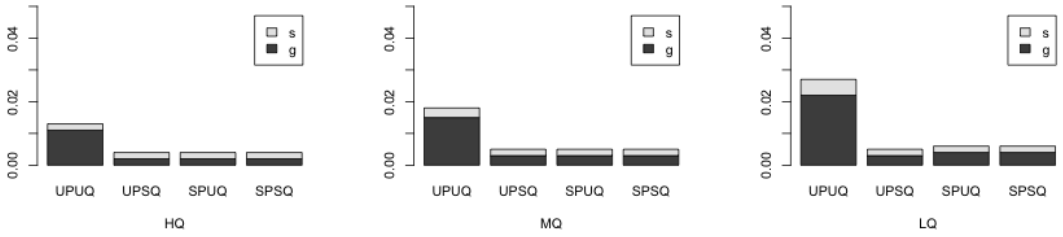


(c) Divergent

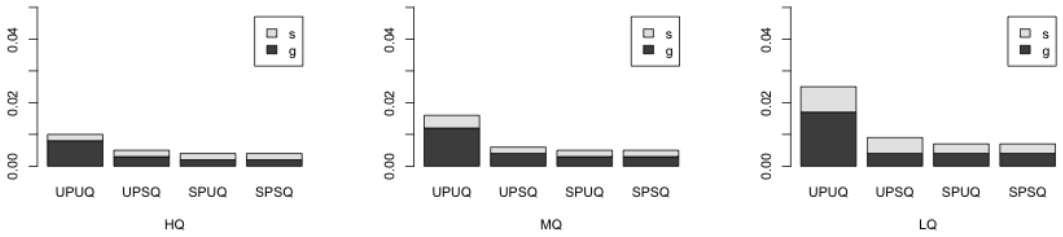
Note. HQ = higher quality; MQ = mixed quality; LQ = lower quality; UPUQ = unstructured prior with unstructured Q-matrix; UPSQ = unstructured prior with structured Q-matrix; SPUQ = structured prior with unstructured Q-matrix; and SPSQ = structured prior with structured Q-matrix.

divergent hierarchical structures, respectively. Furthermore, within each panel there are three horizontally located bar-plots, which were produced by three levels of item qualities. Thus, scrolling across the columns of a row shows the effect of item quality on item parameter bias and RMSE. It can clearly be seen from these four figures that by keeping the sample size, hierarchical structure, and estimation approach constant,

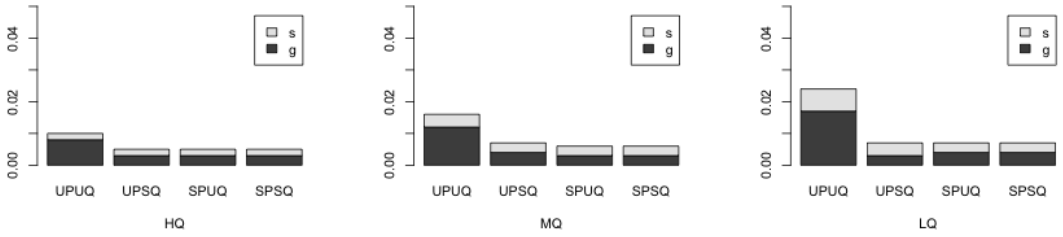
Figure 2.4: Mean Absolute Item Bias of the Estimation Approaches: $N = 500$



(a) Linear



(b) Convergent

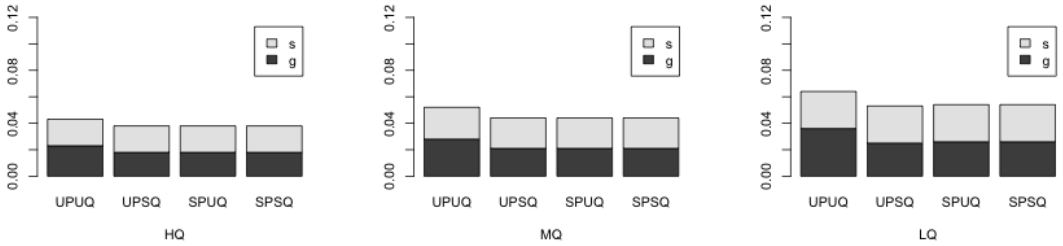


(c) Divergent

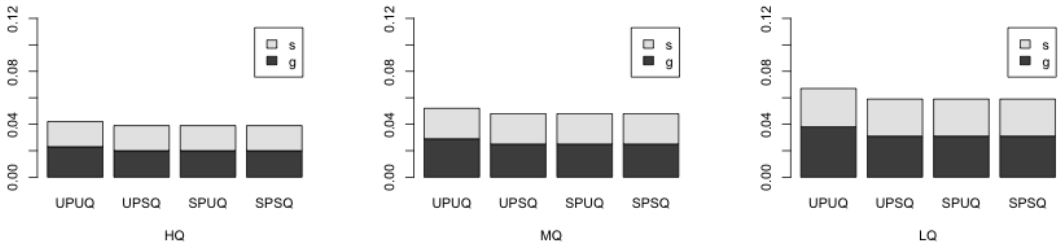
Note. HQ = higher quality; MQ = mixed quality; LQ = lower quality; UPUQ = unstructured prior with unstructured Q-matrix; UPSQ = unstructured prior with structured Q-matrix; SPUQ = structured prior with unstructured Q-matrix; and SPSQ = structured prior with structured Q-matrix.

both the mean absolute bias and RMSE increased as the item quality decreased. These figures also demonstrate that regardless of hierarchical structure and item quality, mean absolute bias and RMSE decreased as the sample increased. One important observation is that mean absolute bias and RMSE significantly decreased when either the Q-matrix or the prior distribution was structured. When neither the

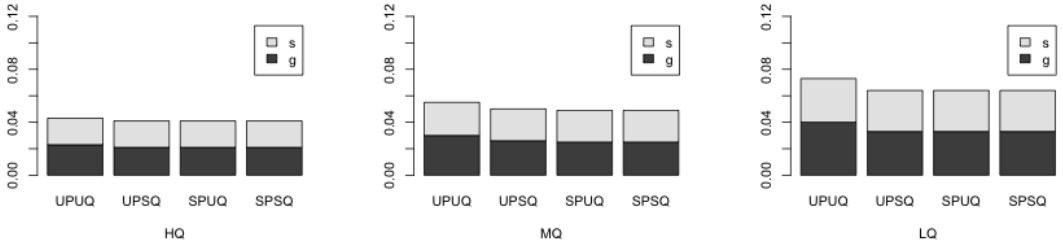
Figure 2.5: Mean Item RMSE of the Estimation Approaches: $N = 1000$



(a) Linear



(b) Convergent



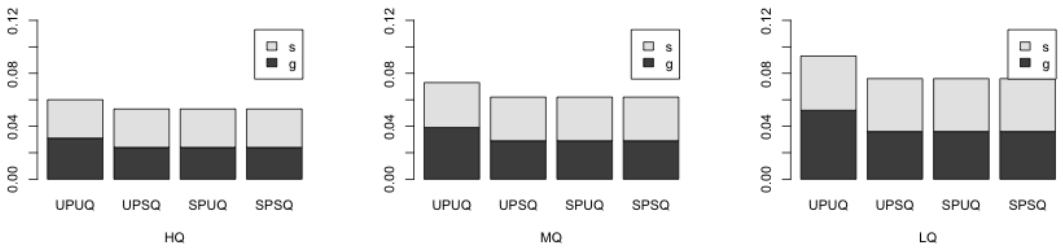
(c) Divergent

Note. HQ = higher quality; MQ = mixed quality; LQ = lower quality; UPUQ = unstructured prior with unstructured Q-matrix; UPSQ = unstructured prior with structured Q-matrix; SPUQ = structured prior with unstructured Q-matrix; and SPSQ = structured prior with structured Q-matrix.

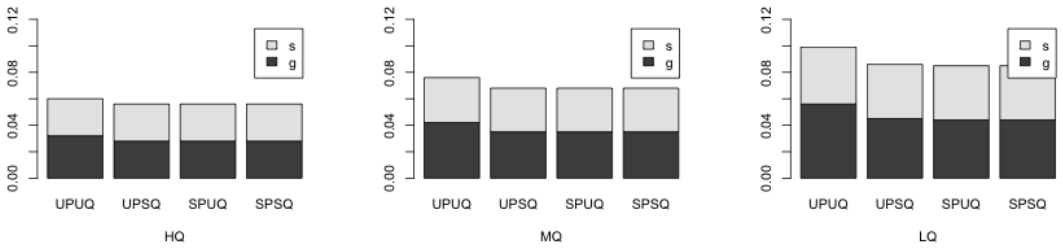
Q-matrix nor the prior distribution was structured, the bias, especially of the guessing parameter, was inflated.

Although it was clear from Figures 2.3, 2.4, 2.5, and 2.6 that item parameter estimation bias and RMSE became larger if neither the Q-matrix nor prior distribution is structured, we investigated the obtained parameter biases and RMSEs for

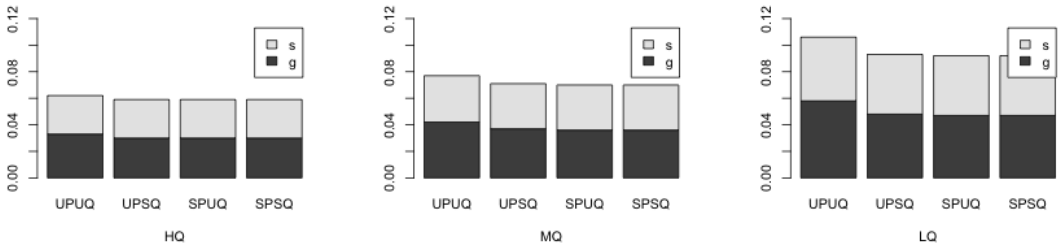
Figure 2.6: Mean Item RMSE of the Estimation Approaches: $N = 500$



(a) Linear



(b) Convergent



(c) Divergent

Note. HQ = higher quality; MQ = mixed quality; LQ = lower quality; UPUQ = unstructured prior with unstructured Q-matrix; UPSQ = unstructured prior with structured Q-matrix; SPUQ = structured prior with unstructured Q-matrix; and SPSQ = structured prior with structured Q-matrix.

all 20-items to provide more accurate information on the remaining three estimation approaches. In examining the parameter estimates for each of 20 items in various conditions, we saw that the parameter estimates using the unstructured and structured versions of the Q-matrix were *identical* when the prior distribution was structured to match the latent class structure. Therefore the results for cases in which prior distribution was structured (i.e., SPUQ and SPSQ) will be reported together as SP_Q condition. This result implies that use of either Q-matrix results in the same item parameter estimates when the prior distribution is properly structured. In other respects, these two types of Q-matrices yielded different item parameter estimates when the prior distribution was unstructured. Therefore, the results based on the two types of Q-matrices needed to be reported separately for the unstructured prior distribution cases. Because the results obtained from each of the estimation approaches presented similar patterns under three levels of generating item parameters, only the results obtained from mixed item quality are discussed in detail. The item parameter biases and corresponding RMSEs by the distinct estimation approaches are given in Tables 2.2 and 2.3, respectively. It should be noted here that although only simulated responses of 1000 examinees were used to obtain the results given in Tables 2.2 and 2.3, sample size of 500 resulted in slightly higher biases and RMSEs, but the pattern was similar.

Results given in Table 2.2 indicate that regardless of the types of hierarchies involved, the guessing and slip parameters can be estimated with the maximum absolute biases of 0.050 and 0.012 (i.e., items 5 in linear and 15 in convergent), respectively, when neither the Q-matrix nor the prior distribution was structured (i.e., UPUQ). These two values reduce to 0.011 and 0.06 (i.e., items 11 in divergent and 6 in linear) when only the Q-matrix was structured (i.e., UPSQ). Furthermore, when we compare the biases obtained from structured Q-matrix with unstructured prior distribution to the biases observed when estimation involved structured prior distribution (SP_Q),

Table 2.2: Item Parameter Bias ($N = 1000$, Mixed Quality Items)

		Linear			Convergent			Divergent		
Par.	Items	UPUQ	UPSQ	SP_Q	UPUQ	UPSQ	SP_Q	UPUQ	UPSQ	SP_Q
g	1	0.001	-0.006	-0.012	0.012	-0.002	-0.008	0.037	0.003	-0.006
	2	-0.043	-0.002	-0.003	-0.049	-0.003	-0.005	-0.018	-0.001	-0.002
	3	-0.042	-0.001	-0.001	-0.017	-0.001	-0.001	-0.013	-0.001	-0.001
	4	-0.043	0.001	0.001	-0.010	0.004	0.003	-0.018	0.011	0.000
	5	-0.050	0.001	0.001	-0.049	-0.007	0.002	-0.020	0.000	0.000
	6	-0.003	0.001	0.001	-0.003	0.001	0.001	-0.003	0.000	0.000
	7	-0.001	-0.001	-0.002	0.006	0.001	-0.001	0.005	0.003	0.003
	8	-0.007	-0.005	-0.005	-0.002	-0.001	-0.002	-0.004	-0.003	-0.003
	9	-0.003	-0.001	-0.001	-0.002	0.000	-0.001	-0.002	0.001	0.001
	10	-0.005	-0.002	-0.002	-0.004	-0.002	-0.003	-0.001	-0.001	-0.001
	11	-0.001	0.000	0.000	-0.001	-0.001	0.000	-0.002	-0.001	-0.001
	12	-0.002	-0.001	-0.001	-0.002	-0.001	-0.001	0.000	0.000	0.000
	13	-0.001	-0.001	-0.001	0.000	-0.001	-0.002	-0.002	-0.002	-0.002
	14	-0.001	0.000	0.000	-0.001	0.000	-0.001	-0.001	-0.001	-0.001
	15	-0.004	-0.003	-0.003	-0.004	-0.003	-0.003	-0.004	-0.003	-0.003
	16	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	17	-0.001	-0.001	-0.001	-0.001	-0.002	-0.001	-0.001	-0.001	-0.001
	18	-0.001	0.000	0.000	-0.001	0.000	0.000	-0.001	0.000	0.000
	19	-0.003	-0.008	-0.014	0.014	0.003	-0.002	0.035	0.008	0.000
	20	-0.003	0.001	0.001	-0.003	0.001	0.001	-0.001	0.002	0.002
s	1	-0.002	0.000	0.000	-0.003	-0.001	0.000	-0.005	-0.001	0.000
	2	-0.001	0.000	0.001	-0.003	0.000	0.001	-0.003	-0.001	-0.001
	3	-0.002	-0.001	-0.001	-0.001	-0.001	0.001	0.003	0.003	0.003
	4	-0.001	0.000	0.000	-0.002	-0.001	0.001	-0.011	-0.001	-0.001
	5	0.001	0.002	0.002	-0.002	0.000	0.000	-0.001	0.000	0.000
	6	-0.005	-0.006	-0.006	-0.002	-0.003	-0.003	-0.001	-0.002	-0.002
	7	0.003	0.002	0.002	-0.001	0.001	0.002	0.004	0.004	0.004
	8	0.002	0.000	0.000	0.000	-0.001	0.001	0.001	0.001	0.001
	9	0.002	-0.001	-0.001	0.004	0.000	0.001	-0.001	-0.001	0.000
	10	0.007	0.001	0.001	0.008	0.000	0.002	0.001	0.000	0.000
	11	0.003	0.003	0.003	-0.002	0.000	-0.001	0.002	0.000	0.000
	12	-0.002	-0.003	-0.003	0.001	0.001	0.001	-0.001	0.000	0.000
	13	0.002	0.002	0.002	-0.001	0.000	0.002	0.000	0.001	0.001
	14	0.005	0.002	0.002	0.000	-0.002	-0.001	-0.003	-0.003	-0.002
	15	0.010	0.004	0.004	0.012	0.004	0.004	0.009	0.005	0.005
	16	0.005	0.005	0.005	0.004	0.003	0.006	-0.001	0.000	0.000
	17	-0.004	-0.004	-0.004	-0.005	-0.003	-0.003	0.000	-0.001	-0.001
	18	0.005	0.005	0.005	-0.001	0.001	0.001	0.003	0.001	0.001
	19	-0.001	0.001	0.001	-0.003	0.000	0.000	-0.004	0.000	0.000
	20	0.003	0.001	0.001	-0.001	-0.002	-0.002	0.003	0.002	0.002

Note. N = sample size; Par. = item parameter; UPUQ = unstructured prior and unstructured Q-matrix; UPSQ = unstructured prior and structured Q-matrix; and SP_ Q = structured prior with either Q-matrix.

Table 2.3: Item Parameter RMSE ($N = 1000$, Mixed Quality Items)

		Linear			Convergent			Divergent		
Par.	Items	UPUQ	UPSQ	SP_Q	UPUQ	UPSQ	SP_Q	UPUQ	UPSQ	SP_Q
g	1	0.054	0.051	0.051	0.074	0.068	0.068	0.100	0.084	0.082
	2	0.060	0.031	0.031	0.075	0.043	0.043	0.038	0.026	0.026
	3	0.051	0.021	0.021	0.029	0.022	0.022	0.022	0.016	0.016
	4	0.051	0.018	0.018	0.023	0.020	0.019	0.060	0.050	0.045
	5	0.057	0.013	0.013	0.058	0.022	0.020	0.032	0.019	0.019
	6	0.015	0.015	0.015	0.016	0.016	0.016	0.018	0.017	0.017
	7	0.031	0.030	0.030	0.043	0.041	0.041	0.031	0.031	0.031
	8	0.022	0.021	0.021	0.019	0.020	0.020	0.016	0.016	0.016
	9	0.019	0.018	0.018	0.015	0.015	0.015	0.017	0.016	0.016
	10	0.014	0.014	0.014	0.017	0.017	0.017	0.018	0.018	0.018
	11	0.013	0.013	0.013	0.015	0.015	0.015	0.014	0.014	0.014
	12	0.014	0.014	0.014	0.016	0.016	0.016	0.017	0.017	0.017
	13	0.020	0.020	0.020	0.019	0.019	0.019	0.016	0.016	0.016
	14	0.016	0.015	0.015	0.014	0.014	0.014	0.013	0.013	0.013
	15	0.015	0.015	0.015	0.014	0.014	0.014	0.013	0.013	0.013
	16	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013
	17	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013
	18	0.013	0.013	0.013	0.014	0.014	0.014	0.014	0.014	0.014
	19	0.060	0.056	0.056	0.086	0.079	0.078	0.111	0.094	0.093
	20	0.013	0.012	0.012	0.014	0.013	0.013	0.016	0.016	0.016
s	1	0.015	0.015	0.015	0.013	0.013	0.013	0.015	0.013	0.013
	2	0.014	0.014	0.014	0.014	0.013	0.013	0.020	0.020	0.020
	3	0.015	0.015	0.015	0.020	0.020	0.020	0.022	0.022	0.022
	4	0.019	0.019	0.019	0.022	0.022	0.021	0.021	0.017	0.017
	5	0.022	0.022	0.022	0.018	0.018	0.018	0.023	0.023	0.023
	6	0.036	0.035	0.035	0.029	0.029	0.029	0.023	0.023	0.023
	7	0.018	0.017	0.017	0.016	0.015	0.015	0.023	0.022	0.022
	8	0.017	0.017	0.017	0.018	0.019	0.018	0.027	0.026	0.026
	9	0.025	0.024	0.024	0.031	0.032	0.032	0.035	0.034	0.034
	10	0.028	0.027	0.027	0.030	0.028	0.028	0.029	0.028	0.028
	11	0.028	0.028	0.028	0.023	0.023	0.023	0.027	0.026	0.026
	12	0.032	0.032	0.032	0.025	0.025	0.025	0.022	0.022	0.022
	13	0.017	0.017	0.017	0.022	0.022	0.022	0.025	0.024	0.024
	14	0.018	0.017	0.017	0.028	0.028	0.028	0.026	0.026	0.026
	15	0.022	0.019	0.019	0.036	0.033	0.033	0.033	0.033	0.032
	16	0.036	0.036	0.036	0.034	0.033	0.034	0.030	0.029	0.029
	17	0.028	0.028	0.028	0.025	0.024	0.024	0.028	0.027	0.027
	18	0.033	0.033	0.033	0.025	0.025	0.025	0.027	0.026	0.026
	19	0.014	0.014	0.014	0.012	0.012	0.012	0.013	0.012	0.012
	20	0.034	0.033	0.033	0.024	0.023	0.023	0.024	0.024	0.024

Note. N = sample size; Par. = item parameter; UPUQ = unstructured prior and unstructured Q-matrix; UPSQ = unstructured prior and structured Q-matrix; and SP_ Q = structured prior with either Q-matrix.

it can be seen that they were very similar such that differences can only be observed in the third decimal points. Therefore, more accurate item parameter estimation can be achieved by structuring the Q-matrix and/or prior distribution in concordance with the hierarchical structure among measured attributes.

Similar to the bias case, RMSEs obtained from UPUQ were slightly higher in comparison to the RMSEs obtained based on UPSQ and SP_Q. This can easily be observed for the guessing parameter RMSEs by looking at the results corresponding to items 3, 4, and 5. When RMSEs of UPSQ and SP_Q are compared, it can be observed that RMSEs of these approaches are almost identical, where the maximum difference is in the third decimal place (see items 4, 8, 16 and 19 in convergent and items 1, 4, 15, and 19 in divergent cases). Thus, these results based on RMSEs indicate that more precise item parameter estimation can be obtained by structuring the Q-matrix and/or prior distribution in accordance with the attribute hierarchy.

As a whole, in all types of hierarchies, the parameter estimates obtained from UPUQ were relatively poor. When the prior distribution was unstructured, more accurate and precise item parameter estimates are obtained when an implicit Q-matrix was used. Because estimations with implicit and explicit Q-matrices yielded identical solutions when the prior distribution was structured, use of implicit Q-matrix does not harm the item parameter estimation procedure. However, the most accurate and precise item parameter estimates are produced with SP_Q, the estimates based on UPSQ are comparably accurate and precise.

Effect of Implicit Q-matrix on Attribute Estimation

The CAC and CVC rates are given in Table 2.4. The CAC and CVC results reported in the upper panel of the table are based on the simulated response vectors of 500 examinees, whereas the result on the lower panel are obtained based on a sample size of 1000 examinees. Scrolling across the columns of a row of the table shows

the classification results based on the linear, convergent, and divergent hierarchical structures. The first six columns under each hierarchy indicate the correct individual attribute classification. The seventh column (i.e., $\bar{A}$) reports the average correct individual attribute classification. The next column labeled as CVC provides correct attribute vector classification rates. Scrolling down the columns enable us to compare the CAC and CVC rates that can be obtained based on the four estimation approaches (i.e., UPUQ, UPSQ, SPUQ, and SPSQ), in which SPUQ and SPSQ are reported together as SP_Q. Lastly, both the upper and lower panels of the table provide results under three levels of item qualities.

Regardless of the hierarchical structures, both CAC and CVC rates were higher for larger sample size conditions (i.e., $N = 1000$) and these classification rates increased as the item quality increased. Similar to the item parameter estimation case, CACs and CVCs resulted from implicit and explicit Q-matrices were identical when prior distribution was structured. Again, this result implies that once the prior distribution is structured, structuring the Q-matrix does not provide additional information for examinee classification. Both classification rates were higher than the ones obtained from estimation approaches involving an unstructured prior distribution. One surprising result was that, unlike in item parameter estimation, structuring the Q-matrix resulted in slight reduction in both CAC and CVC rates when unstructured prior distribution was used. In other words, individual attribute and attribute vector estimation accuracy of UPSQ was slightly less accurate than UPUQ.

Attribute misclassifications of the implicit and explicit Q-matrix use were also investigated in terms of false-negative and false-positive classification rates. A false-negative is said to have occurred when an examinee who mastered an attribute is classified as a non-master. Similarly, a false-positive occurs when an examinee is considered as master of an attribute s/he has not mastered. Because the estimation approaches resulted in similar patterns across different item qualities, and for the two

Table 2.4: Correct Attribute and Vector Classification Rates

			Linear								Convergent								Divergent							
			CAC								CAC								CAC							
N	IQ	EA	A1	A2	A3	A4	A5	A6	$\bar{A}$	CVC	A1	A2	A3	A4	A5	A6	$\bar{A}$	CVC	A1	A2	A3	A4	A5	A6	$\bar{A}$	CVC
500	HQ	UPUQ	.96	.95	.97	.96	.95	1.00	.97	.83	.98	.97	.97	.95	.93	.99	.96	.82	.98	.94	.98	.94	.95	.98	.96	.80
		UPSQ	.97	.95	.97	.96	.97	.99	.97	.84	.98	.97	.97	.93	.93	.99	.96	.82	.98	.94	.98	.94	.94	.97	.96	.80
		SP_Q	.97	.96	.98	.98	.99	1.00	.98	.90	.98	.97	.98	.95	.94	1.00	.97	.85	.98	.95	.98	.95	.96	.98	.97	.83
	MQ	UPUQ	.95	.93	.94	.94	.92	.99	.95	.74	.97	.95	.94	.91	.89	.98	.94	.73	.96	.91	.95	.91	.91	.95	.93	.69
		UPSQ	.95	.93	.94	.93	.94	.98	.95	.74	.97	.94	.94	.89	.90	.98	.94	.71	.97	.91	.95	.91	.91	.94	.93	.68
		SP_Q	.95	.94	.97	.97	.97	.99	.97	.83	.97	.96	.95	.91	.91	.98	.95	.76	.97	.91	.96	.92	.93	.95	.94	.73
	LQ	UPUQ	.91	.89	.90	.89	.87	.97	.91	.61	.94	.92	.89	.84	.84	.95	.90	.58	.93	.85	.91	.85	.85	.90	.88	.53
		UPSQ	.92	.89	.90	.88	.88	.95	.90	.61	.95	.91	.89	.82	.84	.94	.89	.57	.96	.85	.90	.86	.84	.89	.88	.52
		SP_Q	.93	.91	.93	.94	.95	.98	.94	.72	.95	.93	.90	.85	.86	.96	.91	.63	.96	.86	.93	.87	.87	.91	.90	.57
1000	HQ	UPUQ	.97	.96	.97	.97	.97	1.00	.97	.86	.98	.97	.97	.95	.94	.99	.97	.84	.98	.94	.98	.95	.95	.98	.96	.82
		UPSQ	.97	.95	.97	.96	.97	.99	.97	.84	.98	.97	.97	.93	.94	.99	.96	.82	.98	.94	.98	.94	.94	.97	.96	.80
		SP_Q	.97	.96	.98	.98	.99	1.00	.98	.90	.98	.97	.98	.95	.95	1.00	.97	.86	.98	.95	.98	.95	.96	.98	.97	.83
	MQ	UPUQ	.95	.94	.95	.95	.95	.99	.95	.78	.97	.95	.95	.91	.90	.98	.94	.75	.97	.91	.96	.91	.92	.95	.94	.71
		UPSQ	.95	.93	.95	.93	.94	.98	.95	.75	.97	.95	.94	.89	.91	.98	.94	.73	.97	.91	.95	.91	.91	.94	.93	.69
		SP_Q	.95	.94	.96	.97	.98	.99	.97	.83	.97	.96	.95	.92	.92	.99	.95	.77	.97	.92	.96	.92	.93	.95	.94	.73
	LQ	UPUQ	.92	.90	.92	.91	.90	.98	.92	.66	.95	.92	.90	.86	.85	.96	.91	.61	.95	.86	.92	.86	.87	.91	.89	.56
		UPSQ	.93	.89	.90	.88	.88	.95	.90	.61	.95	.92	.89	.83	.85	.94	.90	.59	.96	.86	.90	.86	.85	.89	.89	.54
		SP_Q	.93	.91	.93	.94	.95	.98	.94	.72	.95	.93	.90	.86	.87	.96	.91	.64	.96	.86	.92	.87	.88	.91	.90	.58

Note. N = sample size; IQ = item quality; EA = Estimation Approach; CAC = correct attribute classification; CVC = correct vector classification; A1-A6 = measured attributes; $\bar{A}$ = mean CAC; HQ = higher quality; MQ = mixed quality; LQ = lower quality; UPUQ = unstructured prior and unstructured Q-matrix; UPSQ = unstructured prior and structured Q-matrix; and SP_Q = structured prior with either Q-matrix.

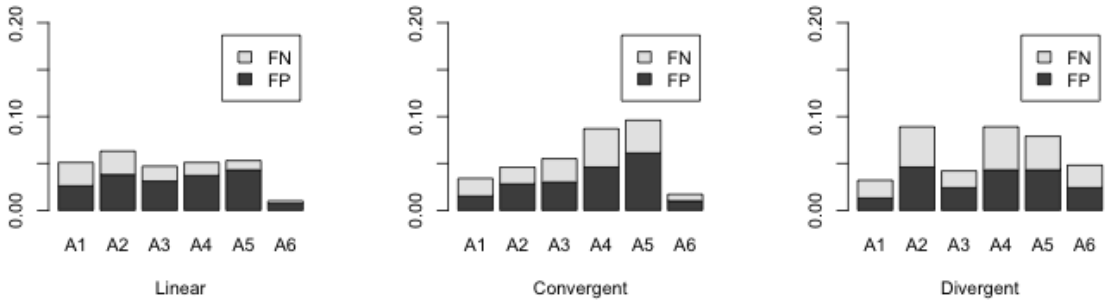
levels of sample sizes, only the results based on mixed item quality under the sample size of 1000 examinees were given in Figure 2.7. In the figure the false-negative rates are stacked up on top of false positives. As can be seen, there are nine bar-plots where scrolling across the columns shows the three hierarchies, whereas scrolling down enables comparison of the misclassifications by the estimation approaches. Here the upper panel, middle and lower panels represent the misclassification results of UPUQ, UPSQ, and SP_Q, respectively.

These misclassification bars show that misclassification occured less often for linear hierarchy, and more often in divergent hierarchy. This result may have arisen from the fact that linear, convergent, and divergent hierarchies involve 7, 12, and 16 latent classes, respectively, which means that each examinee needs to be assigned to one of the 7, 12, and 16 latent classes. One interesting observation is that false-negative classification rates of UPUQ, UPSQ, and SP_Q were much more similar than the corresponding false-positive classification rates. The figure also shows that false-positive rates were higher under estimation approaches involving unstructured prior distribution, whereas false-positive rates were in the same level with the false-negative rates when estimation involved structured priors. It should also be noted here that the smallest misclassification rates were observed under structured prior conditions, which result in approximately equal false-negative and false-positive rates.

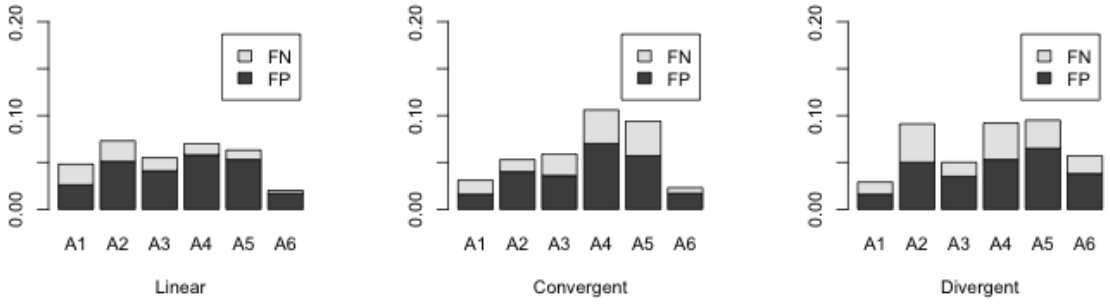
2.5.2 DINO Model Results

The DINO model item parameter estimates were similar to those obtained from the DINA model in terms of size of bias and RMSE, as well as the accuracy and precision produced by the estimation approaches. However, the attribute estimation results were quite different. Individual attribute estimation accuracy and attribute pattern estimation accuracy of the estimation approaches for the DINO model are given in Table 2.5. Note that these results are based the condition where item quality

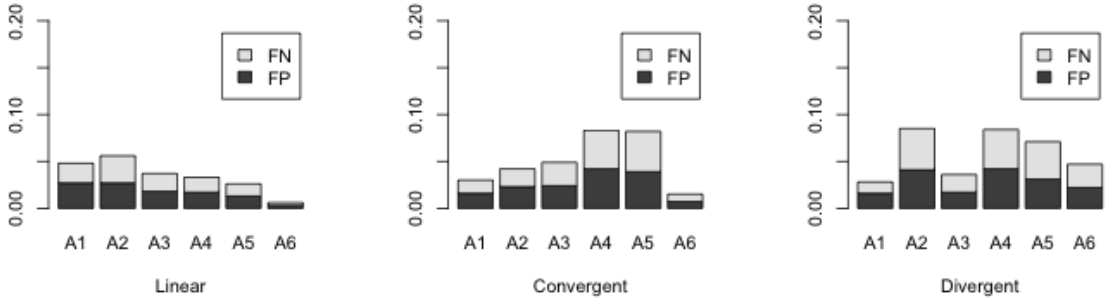
Figure 2.7: False-Negative and False-Positive Attribute Classificatin Rates (MQ Items and $N = 1000$)



(a) UPUQ



(b) UPSQ



(c) SP_Q

Note. FN = false-negative; FP = false-positive; and A1-A6 = measured attributes.

Table 2.5: Attribute and Vector Classification Rates: DINO (MQ and $N = 1000$)

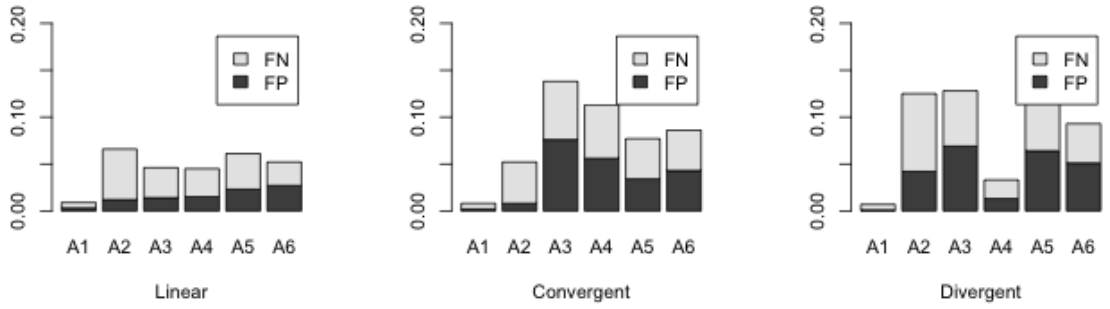
		CAC							
Hierarchy	Est. appr.	A1	A2	A3	A4	A5	A6	$\bar{A}$	CVC
Linear	UPUQ	.99	.93	.95	.95	.94	.95	.95	.78
	UPSQ	.99	.97	.97	.96	.94	.94	.96	.82
	SP_Q	.99	.97	.97	.97	.95	.95	.97	.83
Convergent	UPUQ	.99	.95	.86	.89	.92	.91	.92	.65
	UPSQ	1.00	.98	.87	.89	.93	.91	.93	.67
	SP_Q	1.00	.98	.87	.89	.93	.92	.93	.68
Divergent	UPUQ	.99	.88	.87	.97	.87	.91	.91	.61
	UPSQ	1.00	.90	.87	.97	.87	.91	.92	.62
	SP_Q	1.00	.90	.88	.97	.87	.91	.92	.64

Note. CAC = correct attribute classification; CVC = correct vector classification; Est. Appr. = estimation approach; A1-A6 = measured attributes; $\bar{A}$ = mean CAC; UPUQ = unstructured prior with unstructured Q-matrix; UPSQ = unstructured prior with structured Q-matrix; and SP_Q = structured prior with unstructured Q-matrix.

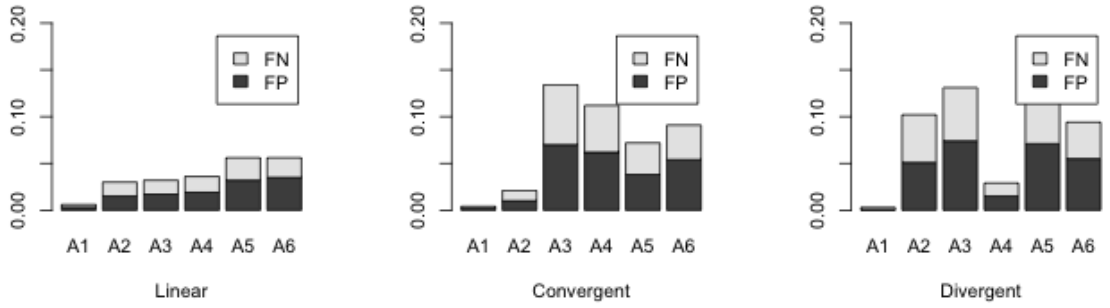
was mixed, and sample size was 1000. These results are given to exemplify how the DINO results can differ from the DINA results. Again, SPUQ and SPSQ classification rates were identical, and the highest as they also were in the DINA case. Yet, now CAC and CVC rates of structured Q-matrix were higher than the ones produced by unstructured Q-matrix under unstructured prior distribution conditions. This result may imply that, in general, holding all other factors constant, items measuring less attributes are more informative in both the DINA and DINO models.

The false-positive and false-negative classification rates of the DINO model estimations under mixed item quality and 1000 examinees were given in Figure 2.8. The false-negative and false-positive rates of the all estimation approaches were higher for the DINO model; however, this may be an artifact of the Q-matrix we began with. Furthermore, recall that in DINA model estimation, false-positive rates under UPUQ and UPSQ estimations were significantly higher than false-negatives. Here in DINO, false-positive and false-negative rates were much more balanced.

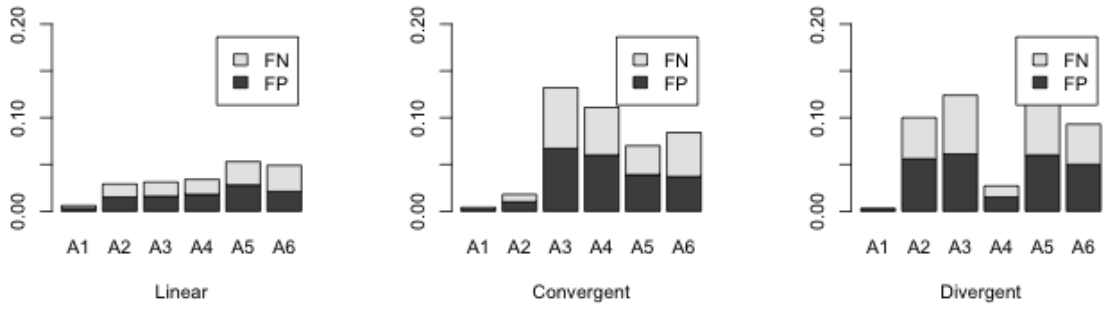
Figure 2.8: False-Negative and False-Positive Attribute Classificatin Rates for DINO (MQ Items and $N = 1000$)



(a) UPUQ



(b) UPSQ



(c) SP_Q

Note. FN = false negative; FP = false positive; and A1-A6 = measured attributes.

2.6 Real Data Analysis

The Simulation study was followed by a numerical example. The analyzed data consisted of 2922 exaimnees' binary responses to the 28 items in the grammar section of the Examination for the Certificate of Proficiency in English (ECPE), which was developed and administered by the University of Michigan English Language Institute in 2003. The dataset and the Q-matrix are available in and obtained from the 'CDM' package (Robitzsch, Kiefer, George, & Uenlue, 2014) in R software environment. The dataset have been analyzed in several studies (e.g., Chiu, Douglas, & Li, 2009; Henson & Templin, 2007; Templin & Bradshaw, 2014), and Templin and Bradshaw (2014) reported that the three attributes measured by the ECPE grammar test had a linear hierarchy among the three attributes (i.e., A1 = Lexical rules is a prerequisite to A2 = Cohesive rules, which is a prerequisite to A3 = Morphosyntactic rules).

The DINA model was fitted to the data with different estimation approaches defined by structured and unstructured versions of the prior distribution and Q-matrix. In the estimations, Q-matrix and the prior distribution were structured such that *Lexical rules* was a prerequisite for *Cohesive rules*, which in turn was a prerequisite attribute for *Morphosyntactic rules*. The obtained item parameter estimates are given in Table 2.6. Based on the simulation results, we expected the item parameter estimates to be similar, especially for those produced by UPSQ and SP_Q. The table indicates that the maximum absolute difference in parameter estimates between UPUQ and UPSQ was 0.043 (i.e., item 24). The maximum difference between the item parameters obtained from UPSQ and SP_Q was 0.001 (i.e., items 2, 8, 25, and 26). These results support the simulation result that implicit Q-matrix provides item parameter estimates as accurate as the calibrations with a structured prior distribution.

Table 2.6: ECPE Test Item Parameter Estimates

	UPUQ		UPSQ		SP_Q	
Items	Guessing	Slip	Guessing	Slip	Guessing	Slip
1	0.705	0.085	0.712	0.096	0.712	0.096
2	0.723	0.101	0.747	0.107	0.746	0.108
3	0.438	0.266	0.438	0.263	0.438	0.263
4	0.481	0.162	0.472	0.165	0.472	0.165
5	0.764	0.040	0.758	0.042	0.758	0.042
6	0.718	0.067	0.712	0.068	0.712	0.068
7	0.544	0.085	0.546	0.083	0.545	0.083
8	0.801	0.040	0.824	0.047	0.824	0.047
9	0.535	0.199	0.532	0.204	0.532	0.204
10	0.481	0.163	0.497	0.161	0.497	0.161
11	0.556	0.099	0.558	0.098	0.558	0.098
12	0.194	0.306	0.199	0.305	0.199	0.305
13	0.632	0.122	0.643	0.121	0.643	0.121
14	0.515	0.212	0.526	0.208	0.526	0.209
15	0.749	0.040	0.743	0.041	0.743	0.041
16	0.549	0.126	0.551	0.124	0.551	0.124
17	0.816	0.058	0.814	0.061	0.813	0.061
18	0.729	0.086	0.723	0.087	0.723	0.087
19	0.473	0.150	0.465	0.154	0.465	0.154
20	0.239	0.296	0.242	0.294	0.242	0.294
21	0.621	0.097	0.623	0.096	0.623	0.096
22	0.322	0.188	0.311	0.193	0.311	0.193
23	0.635	0.076	0.668	0.081	0.667	0.081
24	0.311	0.322	0.354	0.330	0.353	0.331
25	0.511	0.272	0.517	0.267	0.517	0.267
26	0.555	0.211	0.550	0.213	0.550	0.213
27	0.263	0.369	0.279	0.366	0.279	0.366
28	0.659	0.086	0.652	0.088	0.652	0.088

Note. UPUQ = unstructured prior and unstructured Q-matrix; UPSQ = unstructured prior and structured Q-matrix; and SP_Q = structured prior with either Q-matrix.

To demonstrate the individual-attribute and attribute-vector estimation consistency, agreement of UPUQ, UPSQ, and SP_Q are given in Table 2.7. As can be seen from the table, the highest agreement was on the estimation of the first attribute (i.e., Lexical rules), which was the most basic attribute among the three. The lowest agreement was on the second attribute (i.e., Cohesive rules). In general, the highest attribute-wise and pattern-wise agreement were observed among the UPUQ and SP_Q. It should be recalled here that, in the simulation results, the CAC and CVC rates of UPUQ and SP_Q were higher than the UPSQ attribute estimation rates.

Table 2.7: Classification Agreements

	UPSQ				SP_Q			
	A1	A2	A3	Pattern	A1	A2	A3	Pattern
UPUQ	.99	.80	.90	.75	.99	.96	.98	.94
UPSQ	...	...	...	...	1.00	.77	.89	.73

Note. UPUQ = unstructured prior and unstructured Q-matrix; UPSQ = unstructured prior and structured Q-matrix; and SP_Q = structured prior with either Q-matrix.

2.7 Conclusion and Discussion

CDMs are useful tools that provide fine-grained information on examinees' strengths and weaknesses. This type of specific information can then be used to inform classroom instructions and learning. To obtain diagnostic information on examinees' mastery status of a set of attributes, CDAs are developed and administered. The response data are then analyzed by the CDMs to provide diagnostic information. In some cases, attributes may hold a hierarchical structure, such that more basic attributes must be mastered before mastering more complex attributes. In such cases, more accurate item parameter estimation and examinee classification can be achieved by structuring either the Q-matrix or prior distribution in the model estimation procedure.

This study was designed to understand the impact of a structured Q-matrix on item parameter estimation and examinee classification when attributes were hierarchical. Study results indicated that structuring the Q-matrix provides more accurate and precise item parameter estimates in both DINA and DINO models. Although structured Q-matrix resulted in higher attribute and vector-level attribute estimation in the DINO case, it yielded lower attribute and vector level attribute estimation under the DINA model. Results also indicated that both the structured and unstructured versions of the Q-matrix yielded identical item parameter estimates and examinee classification when prior distribution was structured. Furthermore, the highest

attribute and vector-level correct classification rates were obtained when prior distributions were structured so that only the latent classes allowed by the hierarchy were involved in estimation.

Although CDAs are primarily designed to be used as formative assessmens in low-stakes contexts, we cannot discount their potential use in high-stakes testing situations. In such cases, use of an estimation approach that produces more accurate and precise estimates, even if the improvement is slight, might be vital. Furthermore, in practice, the same level of attribute estimation accuracy might be accomplished with shorter tests when hierarchical attribute structure is taken into account.

Several limitations of the current study need to be mentioned. This study considered several factors for model estimation; however, test length and number of attributes were fixed. Moreover, only one explicit Q-matrix was used, which was well-balanced (i.e., all latent classes are identifiable and all attributes are measured approximately equal number of times). It would be interesting to see the impact of unbalanced and/or incomplete (i.e., not all possible single-attribute items included) Q-matrices on model estimation approaches. Lastly, Impact of misspecified Q-matrix on estimation approaches could also be investigated.

When the prior distribution and/or Q-matrix was structured, hierarchy was assumed to be known. However, in practice, hierarchical structure among the attributes may not always be well established. Thus, incorrect specification of the hierarchical relationships among the attributes can be expected to adversely impact the model estimation. In such a case, item parameter estimates and examinee classifications may be adversely affected by structured prior distribution and/or Q-matrix. Therefore, correctly identifying hierarchical relationships among the attributes is of vital importance. Then, development of statistical methods to validate expert-based hierarchical structures can be a potential future research direction.

2.8 References

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The journal of the learning sciences, 6,* 271-315.

Chipman, S. F., Nichols, P. D., & Brennan, R. L. (1995). Introduction. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 1-18). Hillsdale, NJ: Erlbaum.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74,* 633-665.

Corter, J. E. (1995). Using clustering methods to explore the structure of diagnostic tests. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 305-326). Hillsdale, NJ: Erlbaum.

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33,* 163-183.

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

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76,* 179-199.

de la Torre, J. (2014, June). Cognitive diagnosis modeling: A general framework approach. Workshop conducted on the 4th congress on Measurement and Evaluation in Education and Psychology, Ankara. Turkey.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69,* 333-353.

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47,* 227-249.

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.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicologia Educative, 20,* 89-97.

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44,* 285-291.

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

Doornik, J. A. (2011). *Object-oriented matrix programming using Ox (Version 6.20).* London: Timberlake Consultants Press.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Embretson, S .E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics.* Orlando, FL: Academic Press.

Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107-135). New York: Plenum Press.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

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.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuokas rule-space approach. *Journal of Educational Measurement, 41,* 205-237.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60,* 523-547.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64,* 187-212.

National Research Council (2001). *Knowign what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, and R. Glaser (Eds.). Board on Testing and Assessment, Center for Education. Washington, DC: National Academy Press.

No Child Left Behind Act of 2001, Pub. L. No. 1-7-110 (2001).

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

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345-354.

Taylor, K. L., & Dionne, J. P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92,* 413-425.

Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*, 317-339.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods, 11(3),* 287-305.

2.9 Appendices

Appendix 2A: Permissible Latent Classes by General Hierarchy Types

	Attributes						Hierarchies				Attributes						Hierarchies		
LC	A1	A2	A3	A4	A5	A6	Lin.	Con.	Div.	LC	A1	A2	A3	A4	A5	A6	Lin.	Con.	Div.
$\boldsymbol{\alpha}_1$	0	0	0	0	0	0	✓	✓	✓	$\boldsymbol{\alpha}_{33}$	1	0	0	0	0	0	✓	✓	✓
$\boldsymbol{\alpha}_2$	0	0	0	0	0	1				$\boldsymbol{\alpha}_{34}$	1	0	0	0	0	1			
$\boldsymbol{\alpha}_3$	0	0	0	0	1	0				$\boldsymbol{\alpha}_{35}$	1	0	0	0	1	0			
$\boldsymbol{\alpha}_4$	0	0	0	0	1	1				$\boldsymbol{\alpha}_{36}$	1	0	0	0	1	1			
$\boldsymbol{\alpha}_5$	0	0	0	1	0	0				$\boldsymbol{\alpha}_{37}$	1	0	0	1	0	0			✓
$\boldsymbol{\alpha}_6$	0	0	0	1	0	1				$\boldsymbol{\alpha}_{38}$	1	0	0	1	0	1			✓
$\boldsymbol{\alpha}_7$	0	0	0	1	1	0				$\boldsymbol{\alpha}_{39}$	1	0	0	1	1	0			✓
$\boldsymbol{\alpha}_8$	0	0	0	1	1	1				$\boldsymbol{\alpha}_{40}$	1	0	0	1	1	1			✓
$\boldsymbol{\alpha}_9$	0	0	1	0	0	0				$\boldsymbol{\alpha}_{41}$	1	0	1	0	0	0			
$\boldsymbol{\alpha}_{10}$	0	0	1	0	0	1				$\boldsymbol{\alpha}_{42}$	1	0	1	0	0	1			
$\boldsymbol{\alpha}_{11}$	0	0	1	0	1	0				$\boldsymbol{\alpha}_{43}$	1	0	1	0	1	0			
$\boldsymbol{\alpha}_{12}$	0	0	1	0	1	1				$\boldsymbol{\alpha}_{44}$	1	0	1	0	1	1			
$\boldsymbol{\alpha}_{13}$	0	0	1	1	0	0				$\boldsymbol{\alpha}_{45}$	1	0	1	1	0	0			
$\boldsymbol{\alpha}_{14}$	0	0	1	1	0	1				$\boldsymbol{\alpha}_{46}$	1	0	1	1	0	1			
$\boldsymbol{\alpha}_{15}$	0	0	1	1	1	0				$\boldsymbol{\alpha}_{47}$	1	0	1	1	1	0			
$\boldsymbol{\alpha}_{16}$	0	0	1	1	1	1				$\boldsymbol{\alpha}_{48}$	1	0	1	1	1	1			
$\boldsymbol{\alpha}_{17}$	0	1	0	0	0	0				$\boldsymbol{\alpha}_{49}$	1	1	0	0	0	0	✓	✓	✓
$\boldsymbol{\alpha}_{18}$	0	1	0	0	0	1				$\boldsymbol{\alpha}_{50}$	1	1	0	0	0	1			
$\boldsymbol{\alpha}_{19}$	0	1	0	0	1	0				$\boldsymbol{\alpha}_{51}$	1	1	0	0	1	0			
$\boldsymbol{\alpha}_{20}$	0	1	0	0	1	1				$\boldsymbol{\alpha}_{52}$	1	1	0	0	1	1			
$\boldsymbol{\alpha}_{21}$	0	1	0	1	0	0				$\boldsymbol{\alpha}_{53}$	1	1	0	1	0	0		✓	✓
$\boldsymbol{\alpha}_{22}$	0	1	0	1	0	1				$\boldsymbol{\alpha}_{54}$	1	1	0	1	0	1			✓
$\boldsymbol{\alpha}_{23}$	0	1	0	1	1	0				$\boldsymbol{\alpha}_{55}$	1	1	0	1	1	0		✓	✓
$\boldsymbol{\alpha}_{24}$	0	1	0	1	1	1				$\boldsymbol{\alpha}_{56}$	1	1	0	1	1	1		✓	✓
$\boldsymbol{\alpha}_{25}$	0	1	1	0	0	0				$\boldsymbol{\alpha}_{57}$	1	1	1	0	0	0	✓	✓	✓
$\boldsymbol{\alpha}_{26}$	0	1	1	0	0	1				$\boldsymbol{\alpha}_{58}$	1	1	1	0	0	1			
$\boldsymbol{\alpha}_{27}$	0	1	1	0	1	0				$\boldsymbol{\alpha}_{59}$	1	1	1	0	1	0		✓	
$\boldsymbol{\alpha}_{28}$	0	1	1	0	1	1				$\boldsymbol{\alpha}_{60}$	1	1	1	0	1	1		✓	
$\boldsymbol{\alpha}_{29}$	0	1	1	1	0	0				$\boldsymbol{\alpha}_{61}$	1	1	1	1	0	0	✓	✓	✓
$\boldsymbol{\alpha}_{30}$	0	1	1	1	0	1				$\boldsymbol{\alpha}_{62}$	1	1	1	1	0	1			✓
$\boldsymbol{\alpha}_{31}$	0	1	1	1	1	0				$\boldsymbol{\alpha}_{63}$	1	1	1	1	1	0	✓	✓	✓
$\boldsymbol{\alpha}_{32}$	0	1	1	1	1	1				$\boldsymbol{\alpha}_{64}$	1	1	1	1	1	1	✓	✓	✓

Note. LC represents the possible latent classes; ✓ shows the permissible latent classes; A1 through A6 indicate the six attributes; Lin. is the linear hierarchy; Con. is the convergent hierarchy; Div. is the divergent hierarchy.

Appendix 2B: Explicit and Implicit Q-matrices

		Explicit Q-matrix						Implicit Q-matrices																	
								Linear						Convergent						Divergent					
	j	A1	A2	A3	A4	A5	A6	A1	A2	A3	A4	A5	A6	A1	A2	A3	A4	A5	A6	A1	A2	A3	A4	A5	A6
DINA	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	2	0	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
	3	0	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	0
	4	0	0	0	1	0	0	1	1	1	1	0	0	1	1	0	1	0	0	1	0	0	1	0	0
	5	0	0	0	0	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	0	0	1	1	0
	6	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1
	7	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
	8	0	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	0
	9	0	0	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0
	10	0	0	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	0	0	1	1	0
	11	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1
	12	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1
	13	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	0
	14	0	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0
	15	0	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0
	16	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1
	17	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1
	18	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1
	19	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	20	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1
DINO	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	2	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
	3	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
	4	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0
	5	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0
	6	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
	7	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	8	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
	9	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	1	0	0
	10	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0
	11	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1
	12	1	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	13	1	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	14	0	1	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0
	15	0	0	1	1	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	1	0	0
	16	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0
	17	1	0	0	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	18	1	1	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	19	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	20	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1

Note. DINA = deterministic input, noisy "and" gate model; DINO = deterministic input, noisy "or" gate model; A1 through A6 are the measured attributes; and j = item.

Appendix 2C: Ideal Response Patterns: DINA

	Latent classes						Ideal response patterns																	
Structure	α_1	α_2	α_3	α_4	α_5	α_6	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Linear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	0
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Convergent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	0	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	1	0	1	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	1	1
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
	1	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	1
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	0
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Divergent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	0	0	1	1	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
	1	0	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1
	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	1	1
	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	0	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	1	0	1	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	1	1	1
	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	1	1
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	0
	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note. DINA = deterministic input, noisy "and" gate model; and α_1 through α_6 are the attributes.

Appendix 2D: Ideal Response Patterns: DINO

	Latent classes						Ideal response patterns																			
Structure	α_1	α_2	α_3	α_4	α_5	α_6	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Linear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	1	0
	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	0
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Convergent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	1	0
	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0
	1	1	0	1	0	0	1	1	0	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	0
	1	1	1	0	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Divergent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	1	0
	1	0	0	1	0	0	1	0	0	1	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	0
	1	0	0	1	1	0	1	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	0
	1	0	0	1	0	1	1	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	1	0	0	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0
	1	1	0	1	0	0	1	1	0	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	0	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	0
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note. DINO = deterministic input, noisy "or" gate model; and α_1 through α_6 are the attributes.

Chapter 3

Likelihood Ratio Approach for Attribute Hierarchy Identification and Selection

3.1 Introduction

In many educational and psychological tests, examinees are required to use their knowledge, skills, strategies, and cognitive competencies to successfully complete the assessment tasks. These types of categorical latent variables representing the knowledge states of examinees are referred to as *attributes*, which may have a hierarchical structure (Templin & Bradshaw, 2014). Acquisition of domain related attributes may proceed sequentially as the cognitive and educational research suggest that building conceptual understanding requires connecting novel knowledge to preliminary or more basic knowledge (Linn, Eylon, & Davis, 2004; Smith, Wiser, Anderson, & Krajcik, 2006; Vosniadou & Brewer, 1992). Therefore, curriculum need to be designed and developed coherently such that disciplinary and interdisciplinary ideas need to form a meaningful structure that allows teaching steps build upon one another (Schmidt, Wang, & McKnight, 2005).

Gierl, Zheng, & Cui (2008) argued that cognitive processes function within a larger network of inter-related skills so that they share dependencies. This type of dependencies may form hierarchical structures among the attributes. Leighton, Gierl, & Hunka, (2004) asserted that the ordering among the attributes may be derived by empirical or theoretical considerations. For example, Leighton et al. (2004) discussed and explained a hierarchical structure among the seven syllogistic reasoning

attributes. Similarly, Based on a task analysis on SAT algebra I and II items, Gierl, Wang, & Zhou (2008) showed hierarchical relationships among nine ratio and algebra attributes. Likewise, Templin & Bradshaw (2014) reported a linear hierarchical relationship among the three attributes measured on an English proficiency certification test.

When attributes follow a hierarchical structure, the Q-matrix and the prior distribution employed in the model estimation can be modified so that more accurate and precise item and person parameters are obtained (Akbay & de la Torre, 2015). Modification of the prior distribution and the Q-matrix depend on the assumed hierarchical structure so that identifying the correct hierarchical structure is of the essence. Specification of an incorrect hierarchical relationship between any two attributes can substantially degrade estimation accuracy. As such, the importance of correctly identifying the hierarchical structure among attributes cannot be overemphasized.

In current applications, attribute hierarchy is derived from either expert opinions via content analysis or verbal data analyses such as interviews and think-aloud protocols (Cui & Leighton, 2009; Gierl et al., 2008). These hierarchy derivation procedures may result in disagreements over the prerequisite relationships, which may consequently yield more than one hierarchical structure. Furthermore, emerging hierarchical structures from *verbal analysis* and *expert opinion* approaches may not be the same (Gierl et al., 2008). In the literature, there is no model based statistical tests that can address the subjectivity in the conventional methods for attribute structure identification. Therefore, to address this subjectivity, this study proposes a model-fit based empirical exhaustive search method that can be used for identifying the hierarchical relationships among the predefined set of attributes. The proposed method is intended to complement rather than replace the current procedures that rely on experts' decisions.

3.2 Background

The *deterministic input, noisy "and" gate* (DINA; de la Torre, 2009b, Junker and Sijtsma, 2001) model has two item parameters (i.e., guessing and slip). This property makes the DINA model one of the most parsimonious and interpretable CDMs (de la Torre, 2009b). The DINA is known to be a conjunctive model (de la Torre, 2011; de la Torre & Douglas, 2004) as it assumes that missing one of the required attributes result in the baseline probability that is equal to the probability of answering an item when none of the required attributes is mastered (de la Torre, 2009b; Rupp & Templin, 2008). For a given examinee latent group, $\boldsymbol{\alpha}_l$, and the j^{th} q-vector; an ideal response ($\eta_{lj} = 1$ or 0) for the latent group is produced by the *conjunctive condensation function* (Maris, 1995, 1999),

$$\eta_{lj} = \prod_{k=1}^{K} \alpha_{lk}^{q_{jk}}. \tag{3.1}$$

Hence, examinees are splitted into two groups by the the DINA model. The first group can be referred to as mastery group that involves examinees who mastered all required attributes for the item, and the second group, which can be called nonmastery group, consists of examinees who lack at least one of the required attributes.

Possibility of *slipping* on an item for mastery group members, and *guessing* on the item for nonmastery group members are allowed by the probabilistic component of the item response function (IRF) of the model. The probabilities of slipping and guessing on item j are denoted as $s_j = P(X_{ij} = 0|\eta_{ij} = 1)$ and $g_j = P(X_{ij} = 1|\eta_{ij} = 0)$, respectively, where X_{ij} is the observed response of examinee i to item j. Given s_j and g_j, the IRF of the DINA model is written as

$$P(X_j = 1|\boldsymbol{\alpha}_l) = P(X_j = 1|\eta_{jl}) = g_j^{(1-\eta_{jl})}(1 - s_j)^{\eta_{jl}}, \tag{3.2}$$

where $\boldsymbol{\alpha}_l$ is attribute pattern l among 2^K possible attributes patterns; η_{jl} is the expected response of an examinee to item j who possesses attribute pattern l; and g_j and s_j are guessing and slip parameters, respectively (de la Torre, 2009a).

The *deterministic input, noisy "or" gate* (DINO; Templin and Henson, 2006) model is the disjunctive counterpart of the DINA model. It assumes that having at least one of the required attributes and having all required attributes produce the same probability of success on answering an item correctly (Rupp & Templin, 2008; Templin & Rupp, 2006). Because of the disjunctive nature of the model, for a known q-vector of item-j, ideal response of an examinee (i.e., $\omega_{ij} = 1$ or 0) in latent group $\boldsymbol{\alpha}_l$ is produced by the function

$$\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{lk})^{q_{jk}}. \tag{3.3}$$

Although the DINO also splits exaimnees into mastery and nonmastery group, now, the nonmastery group comprises examinees who lack all the required attributes, and the rests are classified into the mastery group.

The DINO model's item parameters are defined as $s_j^* = P(X_{ij} = 0|\omega_{ij} = 1)$ and $g_j^* = P(X_{ij} = 1|\omega_{ij} = 0)$. Therefore, $1 - s_j^*$ becomes the success probability of examinees in the mastery group on item j, and g_j^* becomes the success probability of examinees in the nonmastery group. The item response function of the DINO model can be written as

$$P(X_j = 1|\boldsymbol{\alpha}_l) = P(X_j = 1|\omega_{jl}) = {g_j}^{(1-\omega_{jl})}(1 - s_j)^{\omega_{jl}}, \tag{3.4}$$

where $\boldsymbol{\alpha}_l$ is attribute pattern l; ω_{jl} is the expected response of an examinee to item j who possesses attribute pattern l; and g_j^* and s_j^* are guessing and slip parameters for item j, respectively (Templin & Rupp, 2006).

3.3 An Empirical Exhaustive Search for Identifying Hierarchical Attribute Structure

Specification of an incorrect hierarchical relationship between any two attributes can substantially degrade classification accuracy due to the imposed constraints on the prior distribution. Thus, the importance of correctly identifying the hierarchy among the attributes cannot be overemphasized. As mentioned earlier, attribute hierarchy may be derived from either empirical or theoretical considerations (Leighton et al., 2004). In such a process, content experts define hierarchical relationships via protocol (or verbal) analysis through a sample of items (Gierl et al., 2008), which may result in disagreement among experts over the hierarchical relations. Consequently, this process may yield more than one hierarchical structure. In this study, a model-fit based empirical exhaustive search method for attribute structure identification is proposed to address this subjectivity.

When attribute k is prerequisite to attribute k', not all of the 2^K attribute patterns are permissible. An example is provided in Table 3.1 involving three attributes where A1 is prerequisite for A2, whereas A3 is independent from A1 and A2. In this case, a model can be estimated by one of the two ways depending on how the prior distribution is treated. The first way is to use an unstructured prior distribution in which all attribute patterns are permissible; the second way is to employ a properly structured prior distribution that does not permit certain attribute patterns. By treating the latter as the *null model* and the former as the *alternative model*, we can apply a likelihood ratio test (LRT) with an expectation of rejecting the null hypothesis when attribute k is *not* a prerequisite attribute for attribute k'. Moreover, Akaike information criterion (AIC) and Bayesian information criterion (BIC) may also be used to evaluate the fit of the two competing models.

Table 3.1: Status of Possible Attribute Patterns when A1 is Prerequisite for A2

Latent Classes	Attributes			Status
	A1	A2	A3	
$\boldsymbol{\alpha}_1$	0	0	0	✓
$\boldsymbol{\alpha}_2$	0	0	1	✓
$\boldsymbol{\alpha}_3$	0	1	0	✗
$\boldsymbol{\alpha}_4$	0	1	1	✗
$\boldsymbol{\alpha}_5$	1	0	0	✓
$\boldsymbol{\alpha}_6$	1	0	1	✓
$\boldsymbol{\alpha}_7$	1	1	0	✓
$\boldsymbol{\alpha}_8$	1	1	1	✓

Note. ✓ = the permissible latent classes; ✗ = impermissible latent classes.

Rationale and Search Algorithm: In circumstances where attribute k is prerequisite to attribute k', $3(2^{K-2})$ latent classes are permissible. For example, when six attributes are measured and one attribute, say A1, is prerequisite for another attribute, say A2, then the subset of attribute patterns conforming to this hierarchical relationship (i.e., 00****, 10****, and 11****) becomes permissible, whereas the attribute patterns not implied by this prerequisite relationship (i.e., 01****) will not be allowed. Here * stands for either 0 or 1 allowing 16 different classes. Thus, 1/4 of 2^K latent classes would not be allowed when one attribute is assumed to be prerequisite for another one.

Furthermore, when items are sufficiently discriminating, using a structured prior distribution in the estimation is expected to yield a model-fit statistic that is not too different from the model-data fit obtained from model estimation using an unstructured prior distribution. In the circumstances, we obtain a null model by constraining the prior distribution, whereas the alternative model puts no constraint on the prior distribution. For instance, DINA or DINO model estimates $2J + 2^K - 1$ parameters when prior distribution is unstructured. When the prior distribution is structured such that impermissible latent classes are assigned zero probabilities, the number of parameters to be estimated reduces to $2J + L - 1$ where L is the number

of permissible latent classes.

Due to the nested relationship between the constrained and unconstrained models, we can apply a likelihood ratio test based hypothesis testing with an expectation of retaining the null hypothesis (i.e., the null model fits the data equally well) when attribute k is a prerequisite for attribute k'. Therefore, an empirical exhaustive search based on the LRT can be implemented to identify hierarchical structure of attributes. To attain this, ${}_KP_2 = K(K-1)$ reduced models need to be specified such that each of these reduced models assumes a distinct prerequisite relationship between attributes k and k' (i.e., all possible pairwise prerequisite relationships between K attributes need to be specified as reduced models). Then, LRT based hypothesis testing can be carried out between each of the reduced models and the full model. The hierarchy identification procedure can be carried out in six steps;

- Step 1: Estimate the model parameters with an *unstructured* prior distribution (i.e., alternative model) and record the $-2LL$.
- Step 2: Estimate the model parameters with a *structured* prior distribution conforming to the assumption that attribute k is prerequisite to attribute k' and record the $-2LL$.
- Step 3: Repeat Step 2 for all possible (i.e., ${}_KP_2$) pairwise presumptive relationship.
- Step 4: Compare the fit of the alternative model against the fit of each of ${}_KP_2$ null models.
- Step 5: Report the test results as binary outcomes where 0 and 1 stand for rejection and retained null hypothesis, respectively.
- Step 6: Let these binary outcomes fill in the off-diagonals of an KxK identity matrix, which becomes an R-matrix representing all direct and indirect prerequisite relationships.

Table 3.2: Demonstration of the Implementation of Search Algorithm

	Permissible latent classes											
Hypotheses	000	100	010	001	110	101	011	111	-2LL	Deviance	p-val.	Rej.
A1 → A2	✓	✓		✓	✓	✓		✓	28699.59	0.36	0.547	✗
A1 → A3	✓	✓	✓		✓	✓		✓	28699.23	0.01	0.966	✗
A2 → A1	✓		✓	✓	✓		✓	✓	29306.97	607.74	0.000	✓
A2 → A3	✓	✓	✓		✓		✓	✓	28700.21	0.98	0.322	✗
A3 → A1	✓		✓	✓		✓	✓	✓	29985.20	1285.96	0.000	✓
A3 → A2	✓	✓		✓		✓	✓	✓	29493.34	794.11	0.000	✓
Full Model	✓	✓	✓	✓	✓	✓	✓	✓	28699.23			

Note.p-val.= p-value obtained from the chi-square test with two degrees of freedom; Rej.= rejection decision.

Implementation of the Exhaustive Search Algorithm To illustrate the implementation of the algorithm, consider three linearly hierarchical attributes. Draw the DINA guessing and slip parameters from U(0.05, 0.30), and generate response data for 1000 examinees who respond to 28 items. Then, specify $3(2^{3-2}) = 6$ reduced models such that each time a pair of attributes will have prerequisite relationships, whereas the third attribute will be independent. Then, estimate each of the reduced model and the full model by setting prior distributions compatible with the hypothesized hierarchies.

Once the full and each of the reduced models have been fitted to the data, we can test each of the hypothesis, and make decision on rejection or retaining the null hypotheses as demonstrated in Table 3.2. The results in the table show that the first, second, and fourth reduced models fitted to the data as good as the full model. Therefore, these hypotheses were retained, whereas the rest of the hypotheses were rejected. One can create a KxK identity matrix in the same order that the hypotheses constructed and fill its off-diagonals with the binary decision outcomes row-by-row as shown in Table 3.3. It is then becomes an R-matrix that defines hierarchical attribute structure. In this case, the resulting R-matrix indicates that A1 is prerequisite for both A2 and A3; and A2 is also prerequisite for A3 such that they all together define

a linear structure among the three attributes.

Table 3.3: Incorporation of the Hypothesis Testing Results into R-Matrix

	Identity Matrix				R-Matrix		
	A1	A2	A3		A1	A2	A3
A1	1	0	0	A1	1	**1**	**1**
A2	0	1	0	A2	**0**	1	**1**
A3	0	0	1	A3	**0**	**0**	1

Notice that the exhaustive search is computationally intensive and more efficient algorithms may also be developed. One way to accomplish this may require fixing the hierarchical relationship, when found, for the rest of the search. So that, this can reduce the number of remaining possible pairwise hierarchical relationships to be looked for. Furthermore, the method can also be used iteratively by constraining the prior distribution after each cycle. The algorithm may run until all the hypothesis regarding possible pairwise hierarchical relationships are rejected in one cycle (i.e., akin to refinement techniques used in differential item functioning framework).

3.3.1 Hierarchical Structure Selection

Domain experts can identify the attributes as well as the hierarchical structure among them via two common approaches. In the first approach, experts base their decision on the literature and existing theories about cognitive process of human performance (Embretson, 1998; Leighton et al., 2004). In the second approach, experts identify the attribute and hierarchical structure among them by analyzing the examinee response data, which are directly collected via interview and think-aloud procedures (Chi, 1997; Leighton et al., 2004). A desirable way involves both approaches iteratively where the former approach is used to identify the attributes and the latter is used to validate them (Tjoe, & de la Torre, 2014); however, time and cost arising from conducting both approaches may restrain this option.

Although these two approaches are commonly used, they may result in different hierarchical structures among the attributes (see Gierl et al., 2008). Moreover, even within the same approach, experts may not have a consensus, and two or more hierarchical structures can be proposed. When this is the case, the optimum structure must be selected to provide the most accurate information regarding the examinees' attribute-mastery level. Therefore, this study proposes a likelihood-ratio approach for hierarchical structure selection when multiple structures are proposed.

Recall that LRT is only useful when the models being compared are nested such that the null model can be attained by constraining the alternative model. Thus, in this manuscript, the null and alternative models are based on the hierarchical structures S^0 and S^A, respectively, where S^0 subsumes S^A such that all direct and indirect prerequisite relationships specified in S^A are also specified in S^0. Thus, it further implies that all permissible latent classes defined by S^0 are also in the set of permissible latent classes defined by S^A (i.e., $\mathbf{L}^0 \subset \mathbf{L}^A$). For example, the linear structure among six attributes defined by Leighton et al., (2004) contains all the prerequisite relationships among any pair of attributes that are defined in a convergent structure. Thus, the convergent structure can be regarded as S^A whereas the linear structure can be considered as S^0. Likewise, all seven permissible latent classes defined by linear hierarchy are also members of the set of latent classes allowed by convergent structure. Consequently, a model allowing $\mathbf{L}^0$ in the estimation becomes the null model whereas the model allowing $\mathbf{L}^A$ can be regarded as the alternative model. Consequently, LRT can be useful for selecting one of the competing structures owing to their nested set-up.

Comparing the two structures through the LRT is straightforward when the candidate structures are nested. In other words, selection of the most appropriate structure can be carried out directly using LRT when $\mathbf{L}^0 \subset \mathbf{L}^A$. However, LRT cannot be directly used when two candidate structures are not nested. In such circumstances,

a unified structure (i.e., S^U) can be defined such that it allows union of the permissible latent class sets allowed by S^1 and S^2 (i.e., $\mathbf{L}^1 \subset \mathbf{L}^U$ and $\mathbf{L}^2 \subset \mathbf{L}^U$). Therefore, S^1 and S^2 can be *indirectly* compared through S^U. Alternatively, AIC and BIC model selection criteria can also be used to compare two hierarchical structures that are not nested.

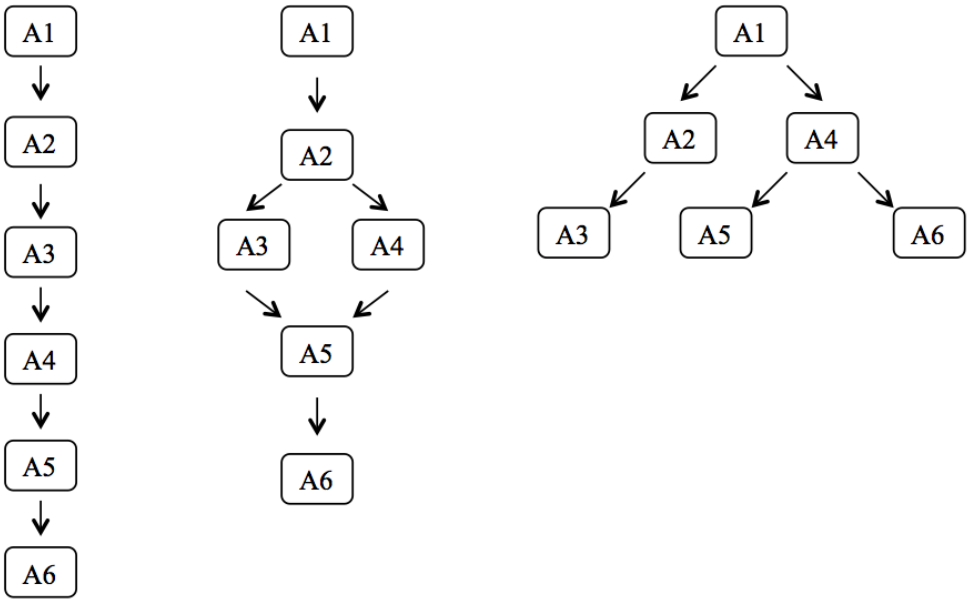
3.4 Simulation Studies

3.4.1 Design

Two simulation studies were designed to assess the viability of the likelihood-ratio based exhaustive search for attribute structure identification and the likelihood approach for hierarchical structure selection. For the first simulation study, the three general attribute hierarchy types (i.e., linear, convergent, and divergent) consisting of six attributes, as defined by Leighton et al. (2004), were considered. These structures are illustrated in Figure 3.1 and their associated permissible latent classes are given in Appendix 3A. An unstructured attributes condition was also included, where all possible latent classes were allowed. Two CDMs (i.e., the DINA and DINO) and two different sample sizes (i.e., $N = 500$ and $N = 1000$) were employed to assess the viability of the search algorithm. Impact of item quality and significance-levels (i.e., α-levels) were also among the considered factors. On top of the LRT, AIC and BIC model selection criteria were employed.

The second simulation study was conducted based on four hypothetical hierarchical structures (i.e., S^1, S^2, S^3, and S^4). These four structures are demonstrated in Figure 3.2, and their corresponding permissible latent classes (i.e., $\mathbf{L}^1$, $\mathbf{L}^2$, $\mathbf{L}^3$, and $\mathbf{L}^4$) are given in Appendix 3B. As can be seen from Appendix 3B, $\mathbf{L}^1$ is a subset of other sets of permissible latent classes. Similarly, $\mathbf{L}^2$ and $\mathbf{L}^3$ are subsets of $\mathbf{L}^4$, whereas $\mathbf{L}^2$ and $\mathbf{L}^3$ are not subsets of one another. This simulation study aimed to assess the

Figure 3.1: General Hierarchy Types in Leighton et al., 2004

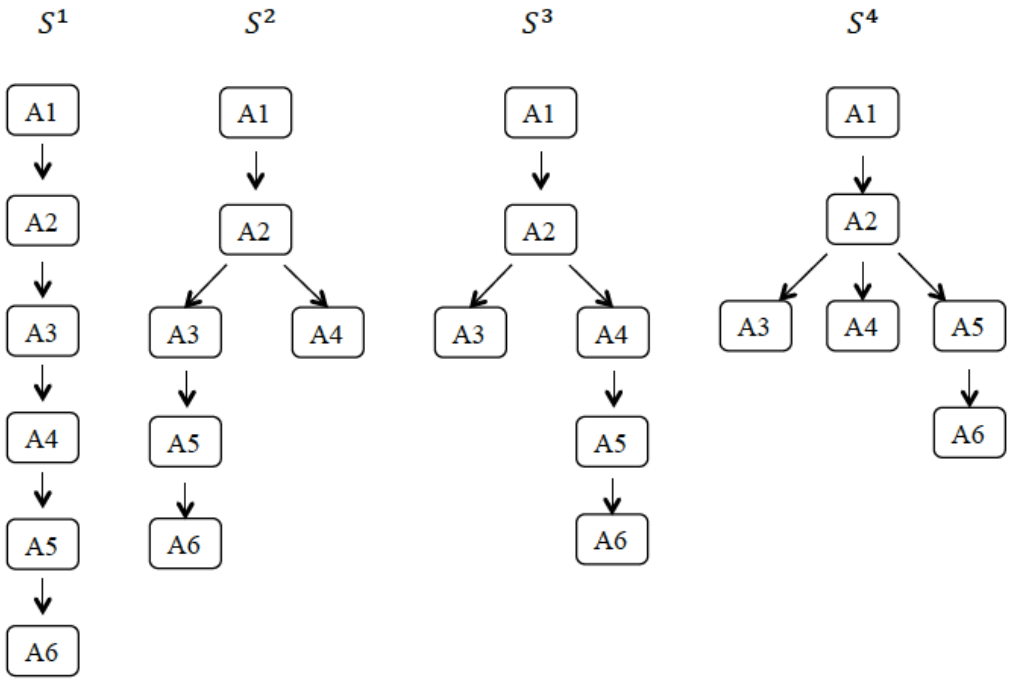


viability of the likelihood ratio approach for hierarchical structure selection as well as investigating impact of some factors that may have on selection of the correct or the most accurate structure. These factors included the size of the difference in the sets of permissible latent classes, item quality, generating CDM, and sample size. Lastly, the AIC and BIC information criteria were also evaluated for structure selection. The Q-matrix that was used for both simulation studies is given in Table 3.4. Ideal response patterns of permissible latent classes by CDM are given in Appendices 3D and 3E to show that all permissible latent classes are identifiable given the Q-matrix.

3.4.2 Data Generation and Model Estimation

The three-levels of item quality were combined with the other factors that were taken into account in this research. For the higher-quality (HQ) items, the lowest and highest success probabilities (i.e., $P(\mathbf{0})$ and $P(\mathbf{1})$) were generated from $U(0.05, 0.20)$ and $U(0.80, 0.95)$. In other words, slip and guessing parameters were drawn from $U(.05, 0.20)$ in the data generation. For the lower-quality (LQ) items, the lowest

Figure 3.2: Four Hypothetical Hierarchical Structures for Six Attributes



and highest success probabilities were drawn from $U(0.15, 0.30)$ and $U(0.70, 0.85)$, respectively, so that the slip and guessing parameters were drawn from $U(0.15, 0.30)$. A third level of item quality was referred to as mixed-quality (MQ), for which the generating parameters were drawn from $U(0.05, 0.30)$.

In all conditions, the test length and number of attributes were fixed to 20-items and six-attributes. Examinees' attribute profiles followed a uniform distribution of permissible latent classes. For the first simulation study, the attributes were generated following the linear, convergent, and divergent hierarchies on top of an unstructured attribute condition. In the second simulation study, examinee attribute profiles followed the four hypothetical hierarchies given in Figure 3.2. In both studies, 100 datasets were generated and analyzed for each condition. Item parameters estimated using marginal maximum likelihood (MML) estimator via expectation-maximization (EM) algorithm. Then, attribute estimation was accomplished using expected a posteriori (EAP) estimator. Data generation and model estimation were carried out by

Table 3.4: The Q-Matrix

	Attributes							Attributes					
Item	A1	A2	A3	A4	A5	A6	*Item*	A1	A2	A3	A4	A5	A6
1	1	0	0	0	0	0	11	0	0	0	0	1	1
2	0	1	0	0	0	0	12	1	0	0	0	0	1
3	0	0	1	0	0	0	13	1	1	1	0	0	0
4	0	0	0	1	0	0	14	0	1	1	1	0	0
5	0	0	0	0	1	0	15	0	0	1	1	1	0
6	0	0	0	0	0	1	16	0	0	0	1	1	1
7	1	1	0	0	0	0	17	1	0	0	0	1	1
8	0	1	1	0	0	0	18	1	1	0	0	0	1
9	0	0	1	1	0	0	19	1	0	0	0	0	0
10	0	0	0	1	1	0	20	0	0	0	0	0	1

Note. A1 through A6 are the measured attributes.

the Oxmetrics programming language (Doornik, 2011). In the analysis, prior distribution that was used in estimation process was structured in concordance with the assumed hierarchy. Table 3.5 summarizes all factors considered in these two simulation studies.

3.5 Results

3.5.1 Results of Simulation Study I

To assess the viability of the search algorithm, false-positive (FP) and false-negative (FN) results are reported in Table 3.6. These results were obtained based on LRT under various significance levels. FP and FN rates indicate the empirical Type-I and Type-II error rates (i.e., α and β), respectively. Complement of FP is the true-negative (i.e., $1 - \alpha$), which can be referred to as *sensitivity*. Sensitivity indicates the proportion of retained null hypothesis when it is true. Likewise, the complement of FN is the true-positive (i.e., $1 - \beta$), which may also be referred to as *specificity*, and it reports the proportion of rejected null hypothesis when it is wrong. It should be noted here that the proportions given in Table 3.6 are averaged across the

Table 3.5: Simulation Factors

Simulation I					
CDM	Sample size	True str.	Item quality	Selection criterion	LRT α-level
DINA	500	Linear	High Quality	LRT	α=0.01
DINO	1000	Convergent	Mixed Quality	AIC	α=0.05
		Divergent	Low Quality	BIC	α=0.10
		Unstructured			α=0.20
Simulation II					
CDM	Sample size	True str.	Item quality	Selection criterion	Candidate str.
DINA	500	S^2	High Quality	LRT	S^1
DINO	1000		Mixed Quality	AIC	S^2
			Low Quality	BIC	S^3
					S^4

Note. DINA = deterministic input, noisy "and" gate model; DINO = deterministic input, noisy "or" gate model; True str. = true structure; LRT = likelihood ratio test; AIC = Akaike information criterion; BIC = Bayesian information criterion.

true and false hypothesized pairwise relationships in accordance with the hierarchical structure. For example, because all the hypotheses we test in unstructured conditions are false, we cannot have FPs, therefore, proportions are obtained by averaging the FNs observed across all $K(K-1) = 30$ hypotheses test results. Similarly, in linear structure, half of the hypotheses we test are true whereas the others are wrong, so FP and FN rates become average of 15 hypothesis test results for each replication.

FPs are, in general, close to zero for linear and divergent hierarchies. The largest FPs (i.e., .003 and .004 for the DINA and DINO) were observed under the significance level of .01 when the sample size was 500 and item quality was low. These FP rates were even smaller in larger sample and higher item quality conditions. Yet, elevated FP was obtained for the convergent hierarchy. Observed FPs in convergent hierarchy case in the DINA were .286; whereas FP rates varied from .001 to .286 in DINO model conditions. These high FP rates may be due to the fact that examinees can master A5 when they master either A3 or A4. Careful review of the analysis results (although not shown in this manuscript) indicated that, for the convergent

Table 3.6: Hypotheses Testing Results: LRT

			DINA model								DINO model							
			Linear		Convergent		Divergent		Unstructured		Linear		Convergent		Divergent		Unstructured	
N	IQ	Sig.	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
500	HQ	α=.01	.000	.007	.286	.039	.000	.000	NA	.000	.000	.013	.082	.029	.000	.031	NA	.000
		α=.05	.001	.002	.286	.021	.000	.000	NA	.000	.000	.009	.126	.018	.000	.021	NA	.000
		α=.10	.001	.001	.286	.018	.000	.000	NA	.000	.000	.005	.158	.013	.000	.017	NA	.000
		α=.20	.002	.001	.286	.012	.000	.000	NA	.000	.000	.003	.189	.006	.001	.012	NA	.000
	MQ	α=.01	.000	.052	.286	.057	.000	.002	NA	.001	.000	.052	.015	.063	.000	.077	NA	.000
		α=.05	.000	.043	.286	.051	.000	.001	NA	.000	.000	.040	.038	.047	.000	.059	NA	.000
		α=.10	.000	.031	.286	.048	.000	.001	NA	.000	.000	.033	.050	.040	.002	.048	NA	.000
		α=.20	.003	.024	.286	.042	.000	.000	NA	.000	.000	.023	.079	.032	.004	.038	NA	.000
	LQ	α=.01	.000	.133	.285	.094	.000	.061	NA	.033	.001	.126	.001	.211	.000	.304	NA	.034
		α=.05	.000	.097	.286	.077	.000	.038	NA	.014	.001	.094	.007	.147	.001	.204	NA	.016
		α=.10	.001	.082	.286	.074	.000	.030	NA	.007	.001	.079	.010	.120	.002	.162	NA	.008
		α=.20	.003	.069	.288	.066	.002	.018	NA	.005	.003	.068	.016	.093	.004	.120	NA	.005
1000	HQ	α=.01	.000	.000	.286	.003	.000	.000	NA	.000	.000	.000	.237	.004	.000	.005	NA	.000
		α=.05	.000	.000	.286	.001	.000	.000	NA	.000	.000	.000	.257	.001	.000	.004	NA	.000
		α=.10	.000	.000	.286	.001	.000	.000	NA	.000	.000	.000	.270	.001	.000	.003	NA	.000
		α=.20	.000	.000	.286	.001	.000	.000	NA	.000	.000	.000	.276	.001	.000	.001	NA	.000
	MQ	α=.01	.000	.010	.286	.036	.000	.000	NA	.000	.000	.017	.106	.031	.000	.028	NA	.000
		α=.05	.000	.006	.286	.030	.001	.000	NA	.000	.000	.009	.137	.025	.000	.024	NA	.000
		α=.10	.000	.003	.286	.024	.001	.000	NA	.000	.000	.007	.156	.019	.000	.018	NA	.000
		α=.20	.000	.001	.286	.016	.001	.000	NA	.000	.000	.004	.184	.014	.000	.014	NA	.000
	LQ	α=.01	.000	.059	.286	.060	.000	.008	NA	.001	.000	.061	.004	.073	.000	.085	NA	.000
		α=.05	.000	.041	.286	.058	.000	.005	NA	.000	.000	.044	.014	.061	.000	.060	NA	.000
		α=.10	.000	.035	.286	.051	.000	.002	NA	.000	.000	.037	.027	.053	.000	.046	NA	.000
		α=.20	.000	.026	.286	.046	.000	.001	NA	.000	.000	.029	.053	.040	.000	.036	NA	.000

Note. N = sample size; IQ = item quality; Sig. = significance level; FP = false-positive; FN = false-negative; HQ = higher quality; MQ = mixed quality; LQ = lower quality; and NA = not applicable.

hierarchy, the search algorithm resulted in a hierarchical structure where A3 and A4 are not prerequisite to A5; however, A1 and A2 are still prerequisites for A5. So, this resulting hierarchy allows some additional latent classes (i.e., 110010 and 110011) in the estimation process along with 12 latent classes conforming the convergent hierarchy.

Notice that FP results in additional latent classes to be allowed in the permissible latent class set, whereas FN yields in discarding some of the latent classes that conform to the true hierarchical structure. Therefore, adverse impact of FN on model estimation may be much stronger in comparison to the negative impact of FP. As shown in the table, FN rates varied across conditions. When sample size was 1000, the largest FN observed for DINA and DINO models were .060 and .085, respectively. The corresponding values for the small sample size were .133 and .304, respectively. For mixed item quality conditions, the FNs decreased significantly and they approached to zero under the high quality item conditions.

When sample size was high, the reported specificity (i.e., $1 - \beta$) for all conditions were about and over .940, .960, and .995 when lower, mixed, and higher quality items were used. For the smaller samples, .870, .940, and .960 were observed, respectively, when the generating and estimating model was DINA. When the DINO model was fitted, these values were comparable in the linear and convergent hierarchy, but larger in the divergent hierarchy. This reduction may be due to the Q-matrix used in this study. The impact of significance level should also be mentioned here. In all conditions, except the convergent hierarchical structure, the FPs were much smaller than the nominal alpha levels. However, considering the fact that FNs significantly reduced by employment of larger significance levels, $\alpha = .20$ may be employed to minimize FN decisions.

Table 3.7 presents the FN and FP results when the AIC and BIC criteria were used. Comparison of the results of the LRT, AIC, and BIC showed that the results of

Table 3.7: Hypotheses Testing Results: AIC and BIC

			AIC				BIC			
			DINA		DINO		DINA		DINO	
N	Hierarchy	IQ	FP	FN	FP	FN	FP	FN	FP	FN
500	Linear	HQ	.000	.008	.000	.013	.011	.001	.004	.000
		MQ	.000	.053	.000	.055	.017	.008	.015	.012
		LQ	.000	.133	.000	.128	.038	.031	.032	.035
	Convergent	HQ	.286	.039	.087	.024	.289	.004	.280	.001
		MQ	.286	.057	.023	.064	.292	.023	.203	.008
		LQ	.285	.094	.003	.209	.305	.038	.109	.045
	Divergent	HQ	.000	.000	.000	.029	.008	.000	.005	.004
		MQ	.000	.002	.000	.074	.024	.000	.035	.015
		LQ	.000	.062	.000	.286	.034	.003	.035	.040
	Unstructured	HQ	NA	.000	NA	.000	NA	.000	NA	.000
		MQ	NA	.001	NA	.000	NA	.000	NA	.000
		LQ	NA	.034	NA	.034	NA	.000	NA	.000
1000	Linear	HQ	.000	.000	.000	.000	.006	.000	.001	.000
		MQ	.000	.011	.000	.017	.008	.000	.008	.000
		LQ	.000	.060	.000	.061	.008	.006	.012	.009
	Convergent	HQ	.286	.004	.237	.004	.289	.000	.287	.000
		MQ	.286	.036	.104	.033	.290	.004	.250	.008
		LQ	.286	.060	.004	.072	.289	.029	.133	.026
	Divergent	HQ	.000	.000	.000	.005	.004	.000	.010	.001
		MQ	.000	.000	.000	.028	.015	.000	.010	.006
		LQ	.000	.008	.000	.085	.009	.000	.020	.019
	Unstructured	HQ	NA	.000	NA	.000	NA	.000	NA	.000
		MQ	NA	.000	NA	.000	NA	.000	NA	.000
		LQ	NA	.001	NA	.000	NA	.000	NA	.000

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; N = sample size; IQ = item quality; FP = false-positive; FN = false-negative; HQ = higher quality; MQ = mixed quality; and LQ = lower quality.

the AIC criterion were almost identical to the results of the LRT under significance level of .01. Due to the fact that FN rates were higher under significance level of .01 than FN rates under larger significance level (i.e., .20); it can be concluded that use of LRT with more liberal significance levels may be preferable over use of the AIC criterion. FP rates of AIC were also comparable to the ones obtained from LRT under the significant level of .01. Moreover, in comparison of BIC and LRT, it was seen that the BIC significantly decreased the FN rates with a largest FN rate of .045 (500 examinees, convergent hierarchy, and lower item quality condition) across all

conditions. Yet, in return, FP rates slightly increased (up to .05). Given the fact that. under the BIC criterion, both FP and FN rates were under .05 for all conditions of linear, divergent, and unstructured hierarchies, BIC could be used to identify direct prerequisite relationships among the attributes.

3.5.2 Results of Simulation Study II

Table 3.8 summarizes the simulation results for structure selection when generating model was the DINA. The table has two main panels where the upper panel provides the results of 500 examinees, whereas the lower panel is created based on sample size of 1000 examinees. In both panels, the results were organized such that LRT, AIC, and BIC results can be seen by scrolling across the columns. Likewise, scrolling across rows presents the results obtained under higher, mixed, and lower item quality conditions. The structure selection results were given in terms of null hypothesis rejection rates where the null hypotheses specify that the more parsimonious model fits the data as good as the more general model. The generating and the candidate hierarchical structures are given as the column and row labels, respectively.

For example when generating structure was S^1, fit of the model allowing the latent classes that are permissible by S^1 was compared with the model-fits that are obtained by the structured-DINA models allowing the latent classes specified by S^2, S^3, and S^4. Across 100 replications when the sample size was 1000 and items were higher quality, rejection rates of S^1 were .00, .01, and .00 in favor of S^2, S^3, and S^4, respectively. In other words, the null hypotheses that structured-DINA model based on S^1 fits the data as well as the structured-DINA models consistent with S^2, S^3, and S^4, were retained 100%, 99%, and 100% of the time, respectively.

Similarly, when the generating hierarchy was S^2, fit of the structured-DINA model based on S^2 was compared against the structured-DINA models based on S^1, S^3, and S^4. Fit of the model based on S^1 was rejected 100% of the time in favor

Table 3.8: Structure Selection Results: DINA

			Higher Quality			Mixed Quality			Lower Quality		
N	SM	Structure	S^1	S^2	S^4	S^1	S^2	S^4	S^1	S^2	S^4
500	LRT	S^1	—	1.00	1.00	—	1.00	1.00	—	1.00	1.00
		S^2	.00	—	1.00	.00	—	1.00	.00	—	1.00
		S^3	.00	—	1.00	.00	—	1.00	.03	—	1.00
		S^4	.00	.01	—	.00	.01	—	.00	.00	—
		(S^3 vs S^4)		1.00			1.00			1.00	
	AIC	S^1	—	1.00	1.00	—	1.00	1.00	—	1.00	1.00
		S^2	.01	—	1.00	.03	—	1.00	.01	—	1.00
		S^3	.02	.00	1.00	.03	.00	1.00	.06	.00	1.00
		S^4	.00	.01	—	.00	.01	—	.00	.03	—
	BIC	S^1	—	1.00	1.00	—	1.00	1.00	—	1.00	1.00
		S^2	.00	—	1.00	.00	—	1.00	.00	—	1.00
		S^3	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00
		S^4	.00	.00	—	.00	.00	—	.00	.00	—
1000	LRT	S^1	—	1.00	1.00	—	1.00	1.00	—	1.00	1.00
		S^2	.00	—	1.00	.01	—	1.00	.00	—	1.00
		S^3	.01	—	1.00	.01	—	1.00	.00	—	1.00
		S^4	.00	.00	—	.00	.00	—	.00	.00	—
		(S^3 vs S^4)		1.00			1.00			1.00	
	AIC	S^1	—	1.00	1.00	—	1.00	1.00	—	1.00	1.00
		S^2	.00	—	1.00	.02	—	1.00	.00	—	1.00
		S^3	.01	.00	1.00	.01	.00	1.00	.01	.00	1.00
		S^4	.00	.00	—	.00	.00	—	.00	.01	—
	BIC	S^1	—	1.00	1.00	—	1.00	1.00	—	1.00	1.00
		S^2	.00	—	1.00	.00	—	1.00	.00	—	1.00
		S^3	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00
		S^4	.00	.00	—	.00	.00	—	.00	.00	—

Note. N = sample size; SM = selection method; LRT = likelihood ratio test; AIC = Akaika information criterion; BIC = Bayesian information criterion; and S^1-S^4 are the hypothetical hierarchical structures.

of the model consistent with S^2; and the model by S^2 was rejected 0% of the time in favor of the model by S^4. It should also be noted here that, because structured models based on S^2 and S^3 are not nested, their comparison by LRT was obtained through S^4. The model fit comparisons of S^3 to S^4 are given as the fifth row in LRT model selection results. As can be seen from the table, when compared against S^4, model by S^2 was retained 100% of the time and model by S^3 was retained 0% of the time.

Table 3.9: Structure Selection Results: DINO

			Higher Quality			Mixed Quality			Lower Quality		
N	SM	Structure	S^1	S^2	S^4	S^1	S^2	S^4	S^1	S^2	S^4
500	LRT	SS^1	—	1.00	.96	—	1.00	1.00	—	.95	.48
		S^2	.02	—	1.00	.03	—	1.00	.03	—	.95
		S^3	.05	—	1.00	.03	—	1.00	.01	—	.99
		S^4	.00	.00	—	.00	.01	—	.02	.02	—
		(S^3 vs S^4)		.97			.86			.51	
	AIC	S^1	—	1.00	1.00	—	1.00	1.00	—	.97	.99
		S^2	.06	—	1.00	.08	—	1.00	.06	—	.97
		S^3	.08	.00	1.00	.08	.00	1.00	.07	.03	.99
		S^4	.00	.02	—	.00	.02	—	.02	.03	—
	BIC	S^1	—	1.00	1.00	—	.96	.86	—	.58	.23
		S^2	.00	—	1.00	.00	—	.93	.00	—	.51
		S^3	.00	.00	1.00	.00	.00	.97	.00	.03	.58
		S^4	.00	.00	—	.00	.00	—	.00	.00	—
1000	LRT	S^1	—	1.00	1.00	—	1.00	.96	—	1.00	1.00
		S^2	.02	—	1.00	.01	—	1.00	.00	—	1.00
		S^3	.02	—	1.00	.01	—	1.00	.00	—	1.00
		S^4	.02	.02	—	.00	.01	—	.00	.00	—
		(S^3 vs S^4)		1.00			1.00			.88	
	AIC	S^1	—	1.00	1.00	—	1.00	1.00	—	1.00	1.00
		S^2	.04	—	1.00	.04	—	1.00	.03	—	1.00
		S^3	.04	.00	1.00	.01	.00	1.00	.03	.02	1.00
		S^4	.01	.03	—	.00	.02	—	.00	.03	—
	BIC	S^1	—	1.00	1.00	—	1.00	.99	—	.92	.74
		S^2	.00	—	1.00	.00	—	1.00	.00	—	.87
		S^3	.00	.02	1.00	.00	.00	1.00	.00	.02	.93
		S^4	.00	.00	—	.00	.00	—	.00	.00	—

Note. N = sample size; SM = selection method; LRT = likelihood ratio test; AIC = Akaika information criterion; BIC = Bayesian information criterion; and S^1-S^4 are the hypothetical hierarchical structures.

Table 3.8 shows that when sample size was 1000, regardless of the item quality levels, the LRT and AIC selected the generating hierarchy at least 99% and 98% of the time, respectively. The BIC selected the true hierarchy 100% of the time. When the sample size was reduced to 500, BIC still performed perfectly on selecting the generating structured-DINA model among the four candidates. Yet, the true model (i.e., DINA model constrained by the generating hierarchy) selection rates of the LRT and AIC decreased down to 97% and 94%, respectively, for the lower quality item

conditions. These two selection rates were to 99% and 98%, respectively, when the generating item parameters were of higher quality. Thus, in general, we can say that all three model selection methods were able to accurately identify the DINA model structured by the generating hierarchy.

When the generating model was the DINO and sample size was 1000, all three model selection methods identified the generating structure at and above 95% of the time under higher and mixed item quality conditions. However, their performance, particularly that of the BIC, significantly decreased when item quality was lower. Similar results with further reduction in identification of the true structure were observed under the sample size of 500. Under the lower item quality, the model selection methods tended to select the more parsimonious structures when the generating hierarchical structure was more liberal. Again, true model identification ability of the BIC was relatively poor in comparison to the LRT and AIC especially under the lower item quality conditions.

Results given in tables 3.8 and 3.9 suggest that true hierarchical structure can be identified accurately by all three model selection methods when items are at least mixed item quality and the sample has more than 500 examinees. Results indicated that when the generating model is DINA, the model selection methods can select the true hierarchical structure even under the lower item quality conditions. Although the observed results show that structure selection in DINO model would not be as accurate under lower item quality conditions, the observed differences under the DINA and DINO models might be due to the Q-matrix used in this study as the information it provides for the two models might not be in the same level.

Table 3.10: Attribute Hierarchy Search on ECPE Data

Null	A1		A2		A3		Full
hypothesis	A2	A3	A1	A3	A1	A2	model
-2LL	85689.47	85689.76	85721.90	85701.65	85912.15	85745.69	85686.52
Deviance	2.95	3.24	35.37	15.12	225.63	59.16	–
p-values	. 22	.20	.00	.00	.00	.00	–
AIC	85705.47	85705.76	85737.90	85717.65	85928.15	85761.69	85706.52
BIC	85753.31	85753.60	85785.74	85765.49	85975.99	85809.53	85766.32

Note. A1 = lexical rules; A2 = cohesive rules; A3 = morphosyntactic rules; AIC = Akaika information criterion; and BIC = Bayesian information criterion.

3.6 Real Data Analysis

The simulation studies were followed by a numerical example for hierarchy identification and selection. The dataset to be analyzed consists of 2922 exaimnees' binary responses to the 28 items in the grammar section of the Examination for the Certificate of Proficiency in English (ECPE). The test was originally developed and administered by the University of Michigan English Language Institute in 2003. The response data and the Q-matrix, which was developed later for CDM analysis, are available in and obtained from the 'CDM' package (Robitzsch, Kiefer, George, & Uenlue, 2014) in R software environment.

First, possible hierarchical structure was identified by the exhaustive search algorithm. Model selections based on LRT, AIC, and BIC were summarized in Table 3.10. Based on the LRT and AIC, null models assuming A1 was prerequisite to A2, and A1 was prerequisite to A3 fitted the data as well as the full model. Further, BIC indicated that, on top of the above two null models, the model assuming A2 was prerequisite for A3 also fitted the data as well as the full model. Therefore, exhaustive search employing LRT and AIC resulted in a divergent hierarchy where A1 was prerequisite for both A2 and A3; whereas BIC resulted in a linear hierarchy where acquisition of A2 required A1; and A3 required both A1 and A2. In the resulting linear and divergent structures, four and five latent classes are permissible,

Table 3.11: Comparison of Attribute Profile Proportions

Attribute Profile		000	100	010	001	110	101	011	111
LCDM	unstructured	.301	.129	.012	.009	.175	.018	.011	.346
DINA	unstructured	.309	.049	.041	.008	.103	.027	.012	.451
	linear	.357	.068	—	—	.102	—	—	.472
	divergent	.357	.073	—	—	.091	.044	—	.434

Note. Attributes are lexical rules, cohesive rules, and morphosyntactic rules, respectively.

respectively, rather than all eight latent classes. Then, the final structure could be selected by comparing the likelihoods of the models with four and five latent classes. When the fit of these two resulting structures (i.e., linear and divergent) were compared with 1 degrees of freedom, the p-value was .000. Therefore, the the linear hierarchy was rejected in favor of the divergent hierarchy.

The same data were analyzed using the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009) in Templin & Bradshaw (2014). They argued that attributes had a linear hierarchy. Although the LCDM and DINA outputs cannot always be compared due to differences in their structural parameters, Table 3.11 show that the LCDM and DINA yielded similar attribute estimations for this particular data when all attributes are assumed to be independent. Thus, it can be argued that the results obtained by these two models are at least comparable. The last two rows of the table show that few examinees were estimated to have attribute profile 101 (4.4 % of the sample), whereas rest of the sample were estimated to have one of the four remaining permissible latent classes. Given the different CDMs employed (i.e., DINA and LCDM), and the small proportion of examinees being classified in the latent class 101 under the divergent hierarchical structure, it is acceptable to arrive at a hierarchy that is not linear.

As the final step, all possible hierarchical structures were compared in terms of model fit. There were a total of 16 hierarchical structures that could be specified. These possible structures and corresponding permissible latent classes are given in

Appendix 3C. When each of the reduced models based on possible hierarchical structures were compared to the full model, only three reduced models (i.e., A1→A2, A3; A1→A3, A2; and A1→{A2, A3}) fitted to the data as well as the full model. Therefore, the final model chosen was A1→{A2, A3}) as it was the most parsimonious one among the three.

3.7 Conclusion and Discussion

When attributes hold a hierarchical structure, CDM model estimation can be improved by taking this hierarchical relationship into account in the estimation process. However, attribute hierarchy must be correct, otherwise, incorrect assumptions on the hierarchy may degrade the model estimation. In this study, an empirical exhaustive search algorithm to identify hierarchical relationships among the attributes was proposed and the viability of the algorithm was investigated in various conditions. In this search algorithm using likelihood based model selection methods each of all possible direct prerequisite relationships among attributes are statistically tested. In this statistical tests, the null hypothesis states that structured-CDM based on a direct prerequisite relationship among two attributes fits the data as well as its unstructured counterpart.

Based on the results we can conclude that the likelihood ratio test based exhaustive search yields an R-matrix that specifies all the prerequisite relationships among all attributes for linear, divergent, and unstructured conditions. However, in the convergent structure case, it fails to identify that both A3 and A4 are prerequisites to A5; rather it specifies a direct prerequisite relationship between A2 and A5. Consequently, permissible latent classes obtained from the exhaustive search includes two additional latent classes on top of the ones defined by the convergent hierarchical structure. Even in this circumstance, the exhaustive search eliminates many of the non-existing latent classes. Thus, overall, the results indicate that we

can, most of the time, recover the generating hierarchical structure successfully. It was also emphasized that the LRT was superior to the AIC in determining the hierarchical structure. Among all three, the BIC was the most successful in recovering the true direct prerequisite relationships among the attributes. It should be noted here that although the method is promising and can be used for exploratory purposes in attribute hierarchy identification; the intent of the method is to complement rather than replace the current subjective procedures.

Moreover, in this study, model-fit based hierarchical structure selection was also investigated. Results indicated that generating hierarchical structure can be accurately selected among the several candidates when model selection criteria such as LRT, AIC, and BIC are used. Correct hierarchy selection rates of all three criteria are 95% or even higher when response data are generated by at least mixed quality items. In small sample conditions, correct hierarchy selection rates were smaller, especially under DINO models. Incorrect hierarchy selection rates were higher with lower quality items.

In practice, multiple nonnested hierarchical structures (e.g., $S2$ and $S3$) could be selected when they are compared against the more liberal candidate (e.g., $S4$). In such tie conditions, practitioner can always consider the p-values in LRT, and AIC and BIC results to break the tie. Another option can be to proceed with the more liberal hierarchical structure (e.g., $S4$) when it does not significantly increase the number of permissible latent classes. Furthermore, practitioner may also consider to compare the selected hierarchical structure to the unstructured model when the selected hierarchy is the most liberal one among the candidates. However, it should be noted that practitioners need to work with domain experts to make final judgments on the attribute structure.

Although several factors that may have impact on the viability of the search algorithm and hierarchy selection were studied in the study, more conditions (i.e.,

varying test lengths, number of attributes, and hierarchies) may be needed to obtain more information on the practicability of the search algorithm and usefulness of likelihood based hierarchy selection. Thus, the use of fixed test length, attributes, and Q-matrix is among the limitations of the study. Because the AIC and BIC take the sample size into account, performance of these two criteria may need to be further studied under different sample sizes. Furthermore, investigation of impact of misspecifications in Q-matrix on the viability of search algorithm and hierarchy selection may also be instructive. Therefore, considering misspecified Q-matricies may be a next step.

3.8 References

Akbay, L., & de la Torre, J. (2015, April). Effect of Q-matrix design under hierarchical attribute structures. Poster presented at the annual meeting of National Council on Measurement in Education, Chicago, IL - USA.

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The journal of the learning sciences, 6,* 271-315.

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

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33,* 163-183.

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

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76,* 179-199.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69,* 333-353.

Doornik, J. A. (2011). *Object-oriented matrix programming using Ox (Version 6.20).* London: Timberlake Consultants Press.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

Gierl, M. J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *Journal of Technology, Learning, and Assessment, 6*(6), 1-50.

Gierl, M. J., Zheng, Y., & Cui, Y. (2008). Using the attribute hierarchy method to identify and interpret cognitive skills that produce group differences. *Journal of Educational Measurement, 45*, 65-89.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74,* 191-210.

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.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuokas rule-space approach. *Journal of Educational Measurement, 41,* 205-237.

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). Mahvah, NJ: Lawrence Erlbaum Associates.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60,* 523-547.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64,* 187-212.

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

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.

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 & Perspective, 4, 1-98.

Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*, 317-339.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods, 11(3),* 287-305.

Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal, 26,* 237-255.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. Cognitive psychology, 24, 535-585.

3.9 Appendices

Appendix 3A: Permissible Latent Classes by General Hierarchy Types

	Attributes						Hierarchies				Attributes						Hierarchies		
LC	A1	A2	A3	A4	A5	A6	Lin.	Con.	Div.	LC	A1	A2	A3	A4	A5	A6	Lin.	Con.	Div.
$\boldsymbol{\alpha}_1$	0	0	0	0	0	0	✓	✓	✓	$\boldsymbol{\alpha}_{33}$	1	0	0	0	0	0	✓	✓	✓
$\boldsymbol{\alpha}_2$	0	0	0	0	0	1				$\boldsymbol{\alpha}_{34}$	1	0	0	0	0	1			
$\boldsymbol{\alpha}_3$	0	0	0	0	1	0				$\boldsymbol{\alpha}_{35}$	1	0	0	0	1	0			
$\boldsymbol{\alpha}_4$	0	0	0	0	1	1				$\boldsymbol{\alpha}_{36}$	1	0	0	0	1	1			
$\boldsymbol{\alpha}_5$	0	0	0	1	0	0				$\boldsymbol{\alpha}_{37}$	1	0	0	1	0	0			✓
$\boldsymbol{\alpha}_6$	0	0	0	1	0	1				$\boldsymbol{\alpha}_{38}$	1	0	0	1	0	1			✓
$\boldsymbol{\alpha}_7$	0	0	0	1	1	0				$\boldsymbol{\alpha}_{39}$	1	0	0	1	1	0			✓
$\boldsymbol{\alpha}_8$	0	0	0	1	1	1				$\boldsymbol{\alpha}_{40}$	1	0	0	1	1	1			✓
$\boldsymbol{\alpha}_9$	0	0	1	0	0	0				$\boldsymbol{\alpha}_{41}$	1	0	1	0	0	0			
$\boldsymbol{\alpha}_{10}$	0	0	1	0	0	1				$\boldsymbol{\alpha}_{42}$	1	0	1	0	0	1			
$\boldsymbol{\alpha}_{11}$	0	0	1	0	1	0				$\boldsymbol{\alpha}_{43}$	1	0	1	0	1	0			
$\boldsymbol{\alpha}_{12}$	0	0	1	0	1	1				$\boldsymbol{\alpha}_{44}$	1	0	1	0	1	1			
$\boldsymbol{\alpha}_{13}$	0	0	1	1	0	0				$\boldsymbol{\alpha}_{45}$	1	0	1	1	0	0			
$\boldsymbol{\alpha}_{14}$	0	0	1	1	0	1				$\boldsymbol{\alpha}_{46}$	1	0	1	1	0	1			
$\boldsymbol{\alpha}_{15}$	0	0	1	1	1	0				$\boldsymbol{\alpha}_{47}$	1	0	1	1	1	0			
$\boldsymbol{\alpha}_{16}$	0	0	1	1	1	1				$\boldsymbol{\alpha}_{48}$	1	0	1	1	1	1			
$\boldsymbol{\alpha}_{17}$	0	1	0	0	0	0				$\boldsymbol{\alpha}_{49}$	1	1	0	0	0	0	✓	✓	✓
$\boldsymbol{\alpha}_{18}$	0	1	0	0	0	1				$\boldsymbol{\alpha}_{50}$	1	1	0	0	0	1			
$\boldsymbol{\alpha}_{19}$	0	1	0	0	1	0				$\boldsymbol{\alpha}_{51}$	1	1	0	0	1	0			
$\boldsymbol{\alpha}_{20}$	0	1	0	0	1	1				$\boldsymbol{\alpha}_{52}$	1	1	0	0	1	1			
$\boldsymbol{\alpha}_{21}$	0	1	0	1	0	0				$\boldsymbol{\alpha}_{53}$	1	1	0	1	0	0		✓	✓
$\boldsymbol{\alpha}_{22}$	0	1	0	1	0	1				$\boldsymbol{\alpha}_{54}$	1	1	0	1	0	1			✓
$\boldsymbol{\alpha}_{23}$	0	1	0	1	1	0				$\boldsymbol{\alpha}_{55}$	1	1	0	1	1	0		✓	✓
$\boldsymbol{\alpha}_{24}$	0	1	0	1	1	1				$\boldsymbol{\alpha}_{56}$	1	1	0	1	1	1		✓	✓
$\boldsymbol{\alpha}_{25}$	0	1	1	0	0	0				$\boldsymbol{\alpha}_{57}$	1	1	1	0	0	0	✓	✓	✓
$\boldsymbol{\alpha}_{26}$	0	1	1	0	0	1				$\boldsymbol{\alpha}_{58}$	1	1	1	0	0	1			
$\boldsymbol{\alpha}_{27}$	0	1	1	0	1	0				$\boldsymbol{\alpha}_{59}$	1	1	1	0	1	0		✓	
$\boldsymbol{\alpha}_{28}$	0	1	1	0	1	1				$\boldsymbol{\alpha}_{60}$	1	1	1	0	1	1		✓	
$\boldsymbol{\alpha}_{29}$	0	1	1	1	0	0				$\boldsymbol{\alpha}_{61}$	1	1	1	1	0	0	✓	✓	✓
$\boldsymbol{\alpha}_{30}$	0	1	1	1	0	1				$\boldsymbol{\alpha}_{62}$	1	1	1	1	0	1			✓
$\boldsymbol{\alpha}_{31}$	0	1	1	1	1	0				$\boldsymbol{\alpha}_{63}$	1	1	1	1	1	0	✓	✓	✓
$\boldsymbol{\alpha}_{32}$	0	1	1	1	1	1				$\boldsymbol{\alpha}_{64}$	1	1	1	1	1	1	✓	✓	✓

Note. LC represents the possible latent classes; ✓ shows the permissible latent classes; A1 through A6 indicate the six attributes; Lin. is the linear hierarchy; Con. is the convergent hierarchy; Div. is the divergent hierarchy.

Appendix 3B: Permissible Latent Classes for Four Hypothetical Structures

	Attributes						Structures					Attributes						Structures			
LC	A1	A2	A3	A4	A5	A6	S^1	S^2	S^3	S^4	LC	A1	A2	A3	A4	A5	A6	S^1	S^2	S^3	S^4
$\boldsymbol{\alpha}_1$	0	0	0	0	0	0	✓	✓	✓	✓	$\boldsymbol{\alpha}_{33}$	1	0	0	0	0	0	✓	✓	✓	✓
$\boldsymbol{\alpha}_2$	0	0	0	0	0	1					$\boldsymbol{\alpha}_{34}$	1	0	0	0	0	1				
$\boldsymbol{\alpha}_3$	0	0	0	0	1	0					$\boldsymbol{\alpha}_{35}$	1	0	0	0	1	0				
$\boldsymbol{\alpha}_4$	0	0	0	0	1	1					$\boldsymbol{\alpha}_{36}$	1	0	0	0	1	1				
$\boldsymbol{\alpha}_5$	0	0	0	1	0	0					$\boldsymbol{\alpha}_{37}$	1	0	0	1	0	0				
$\boldsymbol{\alpha}_6$	0	0	0	1	0	1					$\boldsymbol{\alpha}_{38}$	1	0	0	1	0	1				
$\boldsymbol{\alpha}_7$	0	0	0	1	1	0					$\boldsymbol{\alpha}_{39}$	1	0	0	1	1	0				
$\boldsymbol{\alpha}_8$	0	0	0	1	1	1					$\boldsymbol{\alpha}_{40}$	1	0	0	1	1	1				
$\boldsymbol{\alpha}_9$	0	0	1	0	0	0					$\boldsymbol{\alpha}_{41}$	1	0	1	0	0	0				
$\boldsymbol{\alpha}_{10}$	0	0	1	0	0	1					$\boldsymbol{\alpha}_{42}$	1	0	1	0	0	1				
$\boldsymbol{\alpha}_{11}$	0	0	1	0	1	0					$\boldsymbol{\alpha}_{43}$	1	0	1	0	1	0				
$\boldsymbol{\alpha}_{12}$	0	0	1	0	1	1					$\boldsymbol{\alpha}_{44}$	1	0	1	0	1	1				
$\boldsymbol{\alpha}_{13}$	0	0	1	1	0	0					$\boldsymbol{\alpha}_{45}$	1	0	1	1	0	0				
$\boldsymbol{\alpha}_{14}$	0	0	1	1	0	1					$\boldsymbol{\alpha}_{46}$	1	0	1	1	0	1				
$\boldsymbol{\alpha}_{15}$	0	0	1	1	1	0					$\boldsymbol{\alpha}_{47}$	1	0	1	1	1	0				
$\boldsymbol{\alpha}_{16}$	0	0	1	1	1	1					$\boldsymbol{\alpha}_{48}$	1	0	1	1	1	1				
$\boldsymbol{\alpha}_{17}$	0	1	0	0	0	0					$\boldsymbol{\alpha}_{49}$	1	1	0	0	0	0	✓	✓	✓	✓
$\boldsymbol{\alpha}_{18}$	0	1	0	0	0	1					$\boldsymbol{\alpha}_{50}$	1	1	0	0	0	1				
$\boldsymbol{\alpha}_{19}$	0	1	0	0	1	0					$\boldsymbol{\alpha}_{51}$	1	1	0	0	1	0				✓
$\boldsymbol{\alpha}_{20}$	0	1	0	0	1	1					$\boldsymbol{\alpha}_{52}$	1	1	0	0	1	1				✓
$\boldsymbol{\alpha}_{21}$	0	1	0	1	0	0					$\boldsymbol{\alpha}_{53}$	1	1	0	1	0	0		✓	✓	✓
$\boldsymbol{\alpha}_{22}$	0	1	0	1	0	1					$\boldsymbol{\alpha}_{54}$	1	1	0	1	0	1				
$\boldsymbol{\alpha}_{23}$	0	1	0	1	1	0					$\boldsymbol{\alpha}_{55}$	1	1	0	1	1	0			✓	✓
$\boldsymbol{\alpha}_{24}$	0	1	0	1	1	1					$\boldsymbol{\alpha}_{56}$	1	1	0	1	1	1			✓	✓
$\boldsymbol{\alpha}_{25}$	0	1	1	0	0	0					$\boldsymbol{\alpha}_{57}$	1	1	1	0	0	0	✓	✓	✓	✓
$\boldsymbol{\alpha}_{26}$	0	1	1	0	0	1					$\boldsymbol{\alpha}_{58}$	1	1	1	0	0	1				
$\boldsymbol{\alpha}_{27}$	0	1	1	0	1	0					$\boldsymbol{\alpha}_{59}$	1	1	1	0	1	0		✓		✓
$\boldsymbol{\alpha}_{28}$	0	1	1	0	1	1					$\boldsymbol{\alpha}_{60}$	1	1	1	0	1	1		✓		✓
$\boldsymbol{\alpha}_{29}$	0	1	1	1	0	0					$\boldsymbol{\alpha}_{61}$	1	1	1	1	0	0	✓	✓	✓	✓
$\boldsymbol{\alpha}_{30}$	0	1	1	1	0	1					$\boldsymbol{\alpha}_{62}$	1	1	1	1	0	1				
$\boldsymbol{\alpha}_{31}$	0	1	1	1	1	0					$\boldsymbol{\alpha}_{63}$	1	1	1	1	1	0	✓	✓	✓	✓
$\boldsymbol{\alpha}_{32}$	0	1	1	1	1	1					$\boldsymbol{\alpha}_{64}$	1	1	1	1	1	1	✓	✓	✓	✓

Note. S^1, S^2, S^3, and S^4 are four distinct structures; LC represents the possible latent classes; ✓ shows the permissible latent classes; A1 through A6 indicate the six attributes.

Appendix 3C: Permissible Latent Classes for All Possible Hierarchical Structures

	Permissible latent classes							
Structure	000	100	010	001	110	101	011	111
A1 → A2 → A3	✓	✓			✓			✓
A1 → A3 → A2	✓	✓				✓		✓
A2 → A3 → A1	✓		✓				✓	✓
A2 → A1 → A3	✓		✓		✓			✓
A3 → A1 → A2	✓			✓		✓		✓
A3 → A2 → A1	✓			✓			✓	✓
A1 → {A3, A2}	✓	✓			✓	✓		✓
A2 → {A1, A3}	✓		✓		✓		✓	✓
A3 → {A1, A2}	✓			✓		✓	✓	✓
A1 → A2, A3	✓	✓		✓	✓	✓		✓
A1 → A3, A2	✓	✓	✓		✓	✓		✓
A2 → A1, A3	✓		✓	✓	✓		✓	✓
A2 → A3, A1	✓	✓	✓		✓		✓	✓
A3 → A1, A2	✓		✓	✓		✓	✓	✓
A3 → A2, A1	✓	✓		✓		✓	✓	✓
A1, A2, A3	✓	✓	✓	✓	✓	✓	✓	✓

Note. → specifies prerequisite relationships; and ✓ shows permissible latent class.

Appendix 3D: Ideal Response Patterns: DINA

	Latent classes						Ideal response patterns																	
Structure	α_1	α_2	α_3	α_4	α_5	α_6	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Linear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	0
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Convergent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	0	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	1	0	1	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	1	1
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
	1	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	1
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	0
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Divergent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	0	0	1	1	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
	1	0	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1
	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	1	1
	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	0	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	1	0	1	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	1	1	1
	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	1	1
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	0
	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note. DINA = deterministic input, noisy "and" gate model; and α_1 through α_6 are the attributes.

Appendix 3E: Ideal Response Patterns: DINO

	Latent classes						Ideal response patterns																	
Structure	α_1	α_2	α_3	α_4	α_5	α_6	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Linear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	1	0
	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	0
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Convergent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	1	0
	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0
	1	1	0	1	0	0	1	1	0	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	0
	1	1	1	0	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Divergent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	1	0
	1	0	0	1	0	0	1	0	0	1	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	0
	1	0	0	1	1	0	1	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	0
	1	0	0	1	0	1	1	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	1	0	0	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0
	1	1	0	1	0	0	1	1	0	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	0	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	0
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note. DINO = deterministic input, noisy "or" gate model; and α_1 through α_6 are the attributes.

Chapter 4

Impact of Inexact Hierarchy and Q-matrix on Q-matrix Validation and Structure Selection

4.1 Introduction

Implementation of the most, if not all, of the CDMs requires a constructed Q-matrix that describes the association between the items and attributes needed to complete the items (de la Torre, 2008). Q-matrix embodies the cognitive specifications in test construction (Leighton, Gierl, & Hunka, 2004), and it needs to be correctly specified to provide maximum information in cognitively diagnostic assessment (de la Torre, 2008). However, Q-matrix construction depends on content-experts' judgments and this subjective process may result in misspecifications. Recent research (e.g., Chiu, 2013; de la Torre, 2008; de la Torre & Chiu, 2016) showed the negative effect of misspecified Q-matrix in item calibration, which may consequently degrade attribute classification accuracy. Furthermore, when the correctness of the Q-matrix is not verified, misspecifications in the Q-matrix may result in model misfit (de la Torre, 2008; de la Torre & Chiu, 2016).

To address misspecifications in Q-matrix that may emerge due to misjudgments of experts, several parametric and nonparametric Q-matrix validation methods have been proposed (Chiu, 2013). Some of these methods are designed for specific CDMs (e.g., DINA; Junker & Sijtsma, 2001 and DINO; Templin & Henson, 2006), whereas others can also be used with a general model (e.g., G-DINA; de la Torre,

2011). Viability of the data-driven Q-matrix validation methods have been tested under various factors including the correlational and higher order relationships among the attributes. Yet, neither higher-order nor correlational dependency among the attributes can represent the conditions where the attributes are truly hierarchical. One purpose of this study is to investigate the impact of correctly and incorrectly specified hierarchical structures on Q-matrix validation.

A Q-matrix used for likelihood based structured-model selection (i.e., hierarchy selection) may not always be correct. The second purpose of this manuscript is to make the hierarchical structure selection more realistic with the employment of misspecified Q-matrices. Thus, the problem here is twofold: (1) a misspecified Q-matrix can adversely impact the hierarchical structure selection, and (2) selection of an erroneous structure can also affect Q-matrix validation procedure. This study aims to document the reciprocal impact of misspecified Q-matrix and inexact hierarchical structure on hierarchy selection and Q-matrix validation, respectively.

The remainder of the manuscript consists of the following sections. In the next section, recent development on empirical Q-matrix validation will be presented. In the third section, three empirical Q-matrix validation methods used in this study will be summarized. Fourth section will provide brief summary of the CDMs used throughout this study. Simulation study and its results will be given in the fifth section, respectively. The conclusion of the study will be the last section of the manuscript.

4.2 Background

To provide empirical information to validate the provisional Q-matrix constructed by content experts, de la Torre (2008) proposed an empirically based Q-matrix validation method, which is referred to as the *delta* (δ) *method*. Although he

introduced and implemented the method in conjunction with the DINA model, the method can also be used with the DINO model after modifying the rule for classifying examinees into group $\eta_j = 1$ and $\eta_j = 0$. The method was established on the idea of maximizing the difference in the probabilities of correct response between the groups $\eta_j = 1$ and $\eta_j = 0$ (de la Torre, 2008). He demonstrated that correct q-vector maximizes the differences in the probabilities of correct response for the two distinct groups, whereas the misspecification of any q-entry shrinks the gap due to either a higher guessing or slip parameter.

It should be noted here that Feng (2013) showed the applicability of the delta method, employing sequential search algorithm proposed by de la Torre (2008), in conjunction with reparameterized unified model (Reduced-RUM: Hartz & Roussos, 2005) and DINO models. In her dissertation study, she incorporated the sequential search based on the posterior distribution of attribute patterns and Bayesian model selection criteria. Therefore, she formulated the variation of the delta method as a two-stage validation method.

To make empirical Q-matrix validation viable with more general models, de la Torre and Chiu (2016) extended the δ-method such that the new method operates within the G-DINA framework. This general method is also based upon an item discrimination index referred to as G-DINA discrimination index. The G-DINA discrimination index was proposed in the same paper and is denoted as ς^2. It should be noted here that this index is also item specific. The main principle on which the general validation method was developed is that the correct q-vector results in homogeneous latent groups with respect to the probability of success (de la Torre & Chiu, 2016). De la Torre and Chiu (2016) showed that the correct q-vector is expected to result in groups producing a ς^2, which approximates the maximum ς^2.

On top of the two EM-based Q-matrix validation methods explained above, DeCarlo (2012) proposed a Bayesian model-based Q-matrix validation method that

is only practicable with reparameterized DINA (R-DINA: DeCarlo, 2010) model. Requirement for identification of presumptive misspecified elements of a Q-matrix is the apparent shortcoming of this method. The method treats the q-entries that are conceivably wrong as random variables to be estimated.

To be freed from the challenges encountered in MLE-based Q-matrix validation methods, several nonparametric counterparts have been developed recently. The need for sophisticated and expensive software, high sample size requirement, and being sensitive to starting values are among the challenges encountered in parametric methods. Chiu (2013) proposed Q-matrix refinement method to identify and correct misspecified elements in the matrix. The method uses the residual sum of squares (RSS) between the observed and expected responses as a loss function to identify misspecified q-entries. This method employs a nonparametric classification method, developed by Chiu and Douglas (2013), to assign examinee class memberships, which are then used to obtain examinees' expected responses. It should be noted here that to obtain examinees' expected responses, the nonparametric classification method postulates a specific CDM that supposedly underlies the examinee responses.

Liu, Xu, and Ying (2012) proposed a Q-matrix estimation approach by means of minimizing a loss function. The approach needs only the information of dependent structure of examinee responses as input. Although viability of this approach does not depend on item parameters or attribute distributions, additional information such as the partial information about the Q-matrix can be incorporated in the estimation process to increase the efficiency in terms of computational time and correctness of the resulting Q-matrix. Furthermore, Barnes (2010) proposed a data mining technique, referred to as Q-matrix method, which uses students response data to create a Q-matrix. This method extracts a Q-matrix using a hill-climbing algorithm with the expectation that the extracted Q-matrix can be useful for obtaining diagnostic information. Although Barnes (2010) underscored the usefulness of this initial work on

Q-matrix construction, she pointed out that the extracted and the expert Q-matrices did not often coincide.

4.3 Empirical Q-matrix Validation Methods

4.3.1 The Sequential EM-Based Delta Method

Recall when a test is associated with K attributes, there are 2^K, possible attribute profiles, and $2^K - 1$ possible q-vector that specifies required attributes for item j. Based on the examinees' deterministic responses, any q-vector for item j partitions the examinees into two distinct groups in both the DINA and DINO models (i.e., group $\eta_j = 1$ and $\eta_j = 0$ for the DINA model, and group $\omega_j = 1$ and $\omega_j = 0$ for the DINO model). Then, the q-vector for item j corresponding to $\boldsymbol{\alpha}_l$ is regarded as correct if it maximizes the difference in the probabilities correct response between the two examinee groups (de la Torre, 2008). This statement can be expressed as,

$$q_j = \arg\max_{\alpha_l}[P(X_j = 1|\eta_{ll'} = 1) - P(X_j = 1\eta_{ll'} = 0)] = \arg\max_{\alpha_l}[\delta_{jl}], \tag{4.1}$$

for l, $l' = 1, 2, , 2^K - 1$, where $\eta_{ll'} = \prod_{k=1}^{K} \alpha_{l'k}^{\alpha_{lk}}$.

De la Torre (2008) further explained that the q-vector maximizing the equation above minimizes the sum of the guessing and slip parameters for the same item. Accordingly, he noted that although $\delta_j = (1 - s_j) - g_j$ changes as q-vector changes, δ_j could be considered as an item-specific discrimination index when the Q-matrix is correctly specified (de la Torre, 2008). De la Torre (2008) demonstrated that when an unnecessary attribute was specified, the guessing parameter increases; and omission of a required attribute increases the slip parameter. Therefore, he claimed that any specification error in the q-vector shrinks the δ_j. Then, to find the correct q-vector, starting from single attribute q-vectors, this algorithm keeps track of changes in δ_j

and allows a new attribute in the q-vector as long as the gain in δ_j is meaningful.

4.3.2 General Method of Empirical Q-matrix Validation

This general method of empirical Q-matrix validation is based on the G-DINA index, ς^2, which is a measure of the weighted variance of the probability of success for a particular attribute distribution (de la Torre & Chiu, 2016). For the formal definition of the ς^2, let the total number of attributes measured be K and the first K^* attributes be requisite for the item where $K' \leq K^* \leq K''$. Further let $w(\alpha_1, \ldots, \alpha_{K''})$ and $p(\alpha_1, \ldots, \alpha_{K''})$ be the weight and success probability of $(\alpha_1, \ldots, \alpha_{K''})$, respectively. Also define

$$w(\boldsymbol{\alpha}_{K':K''}) = \sum_{\alpha_1=0}^{1} \ldots \sum_{\alpha_{K'-1}=0}^{1} w(\boldsymbol{\alpha}_{1:K''}), \tag{4.2}$$

and

$$p(\boldsymbol{\alpha}_{K':K''}) = \frac{\sum_{\alpha_1=0}^{1} \cdots \sum_{\alpha_{K'-1}=0}^{1} w(\boldsymbol{\alpha}_{1:K''}) p(\boldsymbol{\alpha}_{1:K''})}{w(\boldsymbol{\alpha}_{K':K''})}. \tag{4.3}$$

Then, given the definitions for the key components above, de la Torre and Chiu (2016) formulated the index as

$$\varsigma^2 = \varsigma^2_{K':K''} = \sum_{\alpha_{K'}=0}^{1} \ldots \sum_{\alpha_{K''}=0}^{1} w(\boldsymbol{\alpha}_{K':K''}) [p(\boldsymbol{\alpha}_{K':K''}) - \bar{p}(\boldsymbol{\alpha}_{K':K''})]^2, \tag{4.4}$$

where $\bar{p}$ is the mean of the probability of success across all the $2^{K''-K'+1}$ possible patterns of $p(\boldsymbol{\alpha}_{K':K''})$ (de la Torre & Chiu, 2016).

De la Torre and Chiu (2016) introduced a general method for empirical Q-matrix validation by employing ς^2. This validation method was developed based on the idea that the correct q-vector results in homogeneous latent groups with respect to the probability of success (de la Torre & Chiu, 2016). However, they argued that any q-vector resulting in latent groups with homogeneous within-group probabilities of success can be referred to as appropriate, from which the most parsimonious one

is defined as the correct q-vector. The correct q-vector is expected to be the most parsimonious q-vector such that resulting ς^2 approximates the maximum ς^2.

4.3.3 The Q-matrix Refinement Method

Chiu (2013) developed the *Q-matrix refinement method* with the intent of identifying and correcting misspecified elements in a Q-matrix. The operational logic of the method is minimizing the RSS between the examinee responses and expected responses of an item. Thus, the CDM that gives rise to observed examinee responses needs to be known to be able to generate the expected responses. Furthermore, this method employs a nonparametric classification method, developed by Chiu and Douglas (2013), to assign examinees into latent classes.

To define the method, let Y_{ij} be observed response of examinee i to item j and η_{ij} be the expected response of the same examinee to the same item. Then, squared residual becomes $(Y_{ij} - \eta_{ij})^2$. The RSS for item j is defined as

$$RSS_j = \sum_{i=1}^{N}(Y_{ij} - \eta_{ij})^2 = \sum_{m=1}^{2^K}\sum_{i \in C_m}(Y_{ij} - \eta_{jm})^2 \tag{4.5}$$

where C_m is the latent class m, and N is the total number of examinees in the sample (Chiu, 2013). Because the expected responses are class-specific, the index for ideal response changes in the right hand-side of the equation (see Chiu, 2013 and Chiu & Douglas, 2013 for details of class-specific expected responses).

In her paper, Chiu justified that when examinee classification is correct, the RSS of the correct q-vector is supposed to be the minimum among all the RSSs produced by $2^K - 1$ possible q-vectors. Therefore, a correct Q-matrix is expected to yield minimum RSS for the entire test. Moreover, the algorithm of the method is initialized with the item that has the highest RSS produced by the provisional Q-matrix. After considering possible change in the q-vector of the initial item, the

algorithm searches for a new item that has the next highest RSS. Essentially, this is an iterative method and the algorithm terminates when all items are checked, and RSS of all items are the minimum in their current q-vector forms.

4.4 The Cognitive Diagnosis Models

Up to date, many cognitive diagnosis models (CDMs) have been proposed. These models differ in terms of the assumptions on the relationships between attributes and test performance. Several well recognized specific CDMs such as the *deterministic input, noisy "and" gate* (DINA; de la Torre, 2009b, Junker and Sijtsma, 2001), *deterministic input, noisy "or" gate* (DINO; Templin and Henson, 2006), and *additive-CDM* (A-CDM; de la Torre, 2011) can be derived from the *generalized deterministic inputs, noisy "and" gate* (G-DINA: de la Torre, 2011) model.

The IRF of the *generalized-DINA model* (G-DINA; de la Torre, 2011) under the *identity link* is

$$P(\alpha_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \qquad (4.6)$$

where K_j^* represents the number of required attributes for the j^{th} item (notice that K_j^* is item specific and does not represents the total number of attributes measured by a test); l represents a particular attribute pattern out of $2^{K_j^*}$ possible attribute patterns; δ_{j0} is the intercept for the item j; δ_{jk} is the main effect due to α_k; $\delta_{jkk'}$ represents interaction effect due to α_k and $\alpha_{k'}$; and $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$ (de la Torre, 2011). The G-DINA model splits examinees into $2^{K_j^*}$ latent groups for item j based on the probability of answering item j correctly.

DINA Model: The DINA model in known to be one of the most parsimonious model as it contains only two item parameters (i.e., guessing and slip). This

model is referred to as a conjunctive model (de la Torre, 2011; de la Torre & Douglas, 2004) because it assumes that missing one of the several required attributes for an item is the same as having none of the required attributes (de la Torre, 2009b; Rupp & Templin, 2008). This assumption is statistically represented by the *conjunctive condensation function* (Maris, 1995, 1999). Given an examinee's attribute profile, $\boldsymbol{\alpha}_i$, and the j^{th} row of the Q-matrix (i.e., attribute specification of j^{th} item) the conjunctive condensation rule generates a deterministic response ($\eta_{ij} = 1$ or 0) through the function

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}. \tag{4.7}$$

Furthermore, item response function (IRF) of the DINA model has a probabilistic component, which allows probability of *slipping* an item when an examinee possesses all required attributes. This probabilistic component also allows an examinee lacking at least one required attribute to *guess* the correct response. The probabilities of slipping and guessing for item j are denoted as $s_j = P(X_{ij} = 0|\eta_{ij} = 1)$ and $g_j = P(X_{ij} = 1|\eta_{ij} = 0)$, respectively, where X_{ij} is the observed response of examinee i to item j. Given s_j and g_j, the IRF of the DINA model is written as

$$P(X_j = 1|\boldsymbol{\alpha}_i) = P(X_j = 1|\eta_{ij}) = g_j^{(1-\eta_{ij})}(1 - s_j)^{\eta_{ij}} \tag{4.8}$$

where $\boldsymbol{\alpha}_i$ is examinee's attribute pattern among 2^K possible attributes patterns; η_{ij} is the expected response of examinee i to item j; and g_j and s_j are guessing and slip parameters, respectively (de la Torre, 2009a). It should be noted here that g_j and $(1 - s_j)$ correspond to δ_{j0} and $\delta_{j12K_j^*}$, respectively, in the G-DINA model representation. Therefore, the G-DINA is reduced to the DINA model by setting all the parameters but δ_{j0} and $\delta_{j12\ldots K_j^*}$ to zero.

DINO Model: The DINO model is the disjunctive counterpart of the DINA model. It assumes that having one of the several required attributes is sufficient to answer an item successfully (Rupp & Templin, 2008; Templin & Rupp, 2006). It is the same thing saying that having one of the required attributes produces the same success probability as having all of the required attributes. Due to this nature of the model, given an examinee's attribute profile, $\boldsymbol{\alpha}_i$, and the j^{th} row of the Q-matrix, the deterministic response (i.e., $\omega_{ij} = 1$ or 0) for the model is obtained by the function

$$\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}}. \tag{4.9}$$

Therefore, the DINO model also splits examinees into two distinct groups. The first group consists of examinees possessing at least one of the required attributes for item j, and the second group is constituted by the examinees who mastered none of the required attributes.

The DINO model also has two item parameters; $s_j^* = P(X_{ij} = 0|\omega_{ij} = 1)$ and $g_j^* = P(X_{ij} = 1|\omega_{ij} = 0)$. Then, $1 - s_j^*$ is the probability that examinee i correctly answers item j given that the examinee has mastered at least one of the required attributes. Likewise, g_j^* stands for the probability that examinee i correctly answers item j when the examinee has not mastered any required attribute. The item response function of the DINO model is written as

$$P(X_j = 1|\boldsymbol{\alpha}_i) = P(X_j = 1|\omega_{ij}) = {g_j}^{(1-\omega_{ij})}(1 - s_j)^{\omega_{ij}} \tag{4.10}$$

where $\boldsymbol{\alpha}_i$ is examinee's attribute pattern; ω_{ij} is the expected response of examinee i to item j; and g_j^* and s_j^* are guessing and slip parameters for item j, respectively (Templin & Rupp, 2006).

The DINO model can also be derived from the G-DINA model by setting $\delta_{jk} = -\delta_{jk'k''} = \ldots = (-1)^{K_j^*+1}\delta_{j12\ldots K_j^*}$ (de la Torre, 2011). In words, the DINO

Table 4.1: The Q-Matrix

	Attributes							Attributes					
Item	A1	A2	A3	A4	A5	A6	*Item*	A1	A2	A3	A4	A5	A6
1	1	0	0	0	0	0	11	0	0	0	0	1	1
2	0	1	0	0	0	0	12	1	0	0	0	0	1
3	0	0	1	0	0	0	13	1	1	1	0	0	0
4	0	0	0	1	0	0	14	0	1	1	1	0	0
5	0	0	0	0	1	0	15	0	0	1	1	1	0
6	0	0	0	0	0	1	16	0	0	0	1	1	1
7	1	1	0	0	0	0	17	1	0	0	0	1	1
8	0	1	1	0	0	0	18	1	1	0	0	0	1
9	0	0	1	1	0	0	19	1	0	0	0	0	0
10	0	0	0	1	1	0	20	0	0	0	0	0	1

Note. A1 through A6 are the measured attributes.

model is obtained from the G-DINA by constraining the main and the interaction effects to be equal with alternating sign that allows only two probabilities; $\delta_{j0} = g_j^*$ and $\delta_{j0} + \delta_{jk} = 1 - s_j^*$.

4.5 Simulation Study

4.5.1 Design and Analysis

To investigate the impact of misspecified Q-matrix on attribute hierarchy selection, three misspecified Q-matrices and the true Q-matrix were used. The generating Q-matrix is given in Table 4.1. Ideal response patterns based on the permissible latent classes generated by this Q-matrix can be found in Appendices 4B and 4C. These show that all permissible latent classes are identifiable. Table 4.2 shows the type of misspecifications and the items they were applied to. Furthermore, three hypothetical hierarchical attribute structures, given in Figure 4.1, were considered as candidate structures. Among the three, S^1 (i.e., structure 1) is the most stringent structure, and S^3 is the most liberal one. Accordingly, S^1 allows only seven latent classes whereas 10

Table 4.2: Misspecified Items and the Types of Misspecifications

		True q-vectors						Q^M DINA						Q^M DINO					
Type	Item	α_1	α_2	α_3	α_4	α_5	α_6	α_1	α_2	α_3	α_4	α_5	α_6	α_1	α_2	α_3	α_4	α_5	α_6
OS	8	0	1	1	0	0	0	0	1	1	**1**	0	0	**1**	1	1	0	0	0
	10	0	0	0	1	1	0	0	0	0	1	1	**1**	0	0	**1**	1	1	0
US	8	0	1	1	0	0	0	0	1	**0**	0	0	0	0	**0**	1	0	0	0
	10	0	0	0	1	1	0	0	0	0	1	**0**	0	0	0	0	**0**	1	0
OUS	8	0	1	1	0	0	0	**1**	1	**0**	0	0	0	**1**	1	**0**	0	0	0
	10	0	0	0	1	1	0	0	0	0	**0**	1	**1**	0	0	0	**0**	1	**1**

Note. OS = Overspecified q-vector; US=Underspecified q-vector; OUS=Over- and underspecified q-vector; Q^M DINO = misspecified q-vectors for DINO; Q^M DINA = misspecified q-vectors for DINA.

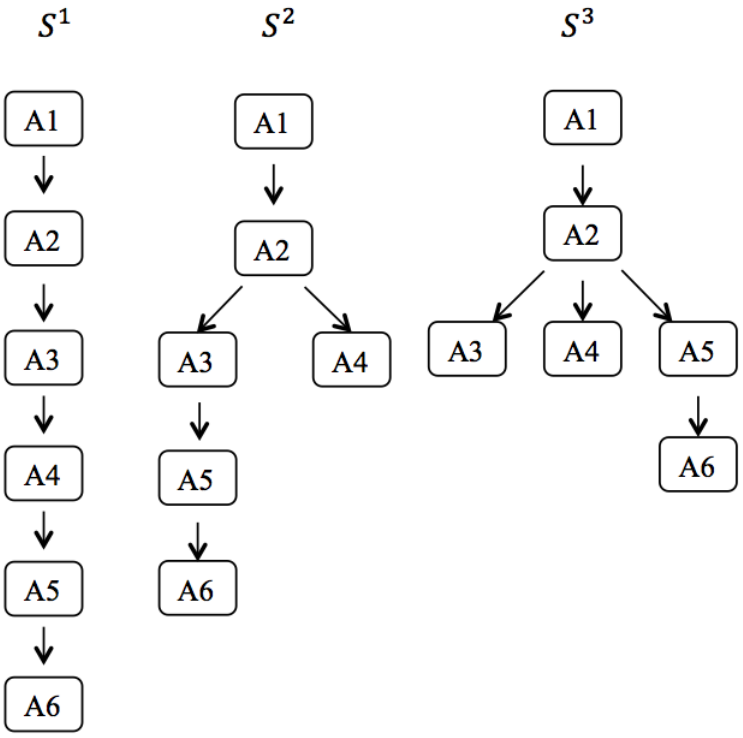
and 14 latent classes are permissible under S^2 and S^3, respectively. The permissible latent classes corresponding to these hierarchical structures are given in Appendix 4A. The sets of latent classes associated with the three hierarchical structures hold the following relationships:

$$\boldsymbol{L}^1 \subset \boldsymbol{L}^2 \subset \boldsymbol{L}^3,$$

where $\boldsymbol{L}$ indicates the set of permissible latent classes. By employing each of the misspecified and true Q-matrices, one structure was selected among the four candidates (i.e., S^1, S^2, S^3, and S^U) where S^U stands for independent attribute structure. The LRT, AIC, and BIC model selection criteria were taken into consideration in the structure (i.e., structured-model) selection.

Moreover, to examine the impact of a hierarchical attribute structure on Q-matrix validation, each of the misspecified Q-matrix and the true Q-matrix were validated with datasets generated based on unstructured and hierarchical (i.e., S^2 in Figure 4.1) attribute structures. Unstructured attribute condition was used to improve the comparability of validation results under hierarchical and independent attribute structures. Another reason to include unstructured attribute conditions in the fact that not all three methods were evaluated simultaneously in one study.

Figure 4.1: Three Hypothetical Hierarchies



Attribute generation followed a uniform distribution of the permissible latent classes. All three recognized empirical Q-matrix validation methods summarized earlier were employed in a validation procedure to assess their viability under hierarchical attribute structures. To control the impact of item quality and sample sizes, three levels of item qualities (i.e., higher, mixed, and lower) and two levels of sample size (i.e., $N = 1000$ and $N = 500$) were considered. All factors that were taken into account are summarized in Table 4.3. In all conditions, data were generated using the same Q-matrix (see Table 4.1).

The test length and number of attributes were fixed to twenty and six, respectively. Throughout the study, the two parsimonious CDMs, namely, the DINA and DINO models, were considered. For the purpose of data generation, the lowest and highest success probabilities (i.e., $P(\mathbf{0})$ and $P(\mathbf{1})$) were generated from $U(0.05, 0.20)$ and $U(0.80, 0.95)$ for the higher-quality (HQ) item conditions. In other words, slip

Table 4.3: Factors to Be Considered in Study III

CDM	Sample size	Item quality	Q-matrix	Candidate structure	Selection criterion	Validation method
DINA	500	HQ	Q^T	S^1	LRT	δ
DINO	1000	MQ	Q^U	S^2	AIC	ς^2
		LQ	Q^O	S^3	BIC	RSS
			$Q^{O\&U}$	S^U		

Note. CDM = cognitive diagnosis model; DINA = deterministic input, noisy "and" gate model; DINO = deterministic input, noisy "or" gate model; HQ = higher quality; MQ= mixed quality; LQ = lower quality; S^1, S^2, and S^3 = hypothetical hierarchies given in Figure 4.1; S^U = unstructured hierarchy; LRT = likelihood ratio test; AIC = Akaike informatin criterion; BIC = Bayesian information criterion; Q^T = true Q-matrix; Q^O = overspecified Q-matrix; Q^U = underspecified Q-matrix; $Q^{O\&U}$ = over- and underspecified Q-matrix; δ=delta method; ς^2=general Q-matrix validation method; and RSS=Q-matrix refinement method.

and guessing parameters were drawn from $U(.05, 0.20)$ for HQ items. For the lower-quality (LQ) items, the highest and lowest success probabilities were drawn from $U(0.15, 0.30)$ and $U(0.70, 0.85)$, respectively, so that the slip and guessing parameters were drawn from $U(0.15, 0.30)$. The mixed-quality items had item parameters drawn from $U(0.05, 0.30)$. For each condition, 100 data sets were generated and analyzed. Data generation and all analyses except statistical refinement of Q-matrix were performed using the OxMetrics programming language (Doornik, 2011). Q-matrix validation based on statistical refinement was carried out using the R-package NPCD version 1.0-7 (Zheng & Chiu, 2014).

4.5.2 Results

Results of Hierarchy Selection

Correct structure selection rates based on LRT, AIC, and BIC selection criteria are reported in Table 4.4 in terms of null hypothesis rejection rates for various conditions where either correctly specified Q-matrix or one of the misspecified Q-matrices is employed. The results when the DINA model was fitted are presented in the upper panel, whereas the corresponding results for the DINO are given in the

Table 4.4: Correct Structured-Model (i.e., hierarchy) Selection Rates by the Q-Matrix Types

			$N = 500$									$N = 1000$								
			LRT			AIC			BIC			LRT			AIC			BIC		
CDM	IQ	Q-matrix	S1	S3	S4	S1	S3	S4	S1	S3	S4	S1	S3	S4	S1	S3	S4	S1	S3	S4
DINA	HQ	True	1.00	.01	.00	1.00	.01	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00
		Over	1.00	.01	.00	1.00	.01	.00	1.00	.00	.00	1.00	.02	.00	1.00	.02	.00	1.00	.00	.00
		Under	1.00	.00	.00	1.00	.01	.00	1.00	.00	.00	1.00	.00	.01	1.00	.02	.00	1.00	.00	.00
		O & U	1.00	.00	.00	1.00	.02	.00	1.00	.00	.00	1.00	.02	.08	1.00	.03	.01	1.00	.00	.00
	MQ	True	1.00	.01	.00	1.00	.01	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00
		Over	1.00	.01	.00	1.00	.01	.00	.97	.00	.00	.99	.00	.00	.99	.01	.00	.99	.00	.00
		Under	1.00	.00	.00	1.00	.03	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00
		O & U	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.01	.02	1.00	.01	.00	1.00	.00	.00
	LQ	True	1.00	.00	.00	1.00	.02	.00	1.00	.00	.00	1.00	.00	.00	1.00	.01	.00	1.00	.00	.00
		Over	.98	.00	.00	.99	.00	.00	.97	.00	.00	1.00	.02	.00	1.00	.04	.00	.99	.00	.00
		Under	1.00	.00	.00	1.00	.01	.00	1.00	.00	.00	1.00	.02	.00	1.00	.02	.00	1.00	.00	.00
		O & U	1.00	.00	.00	1.00	.00	.00	1.00	.00	.00	1.00	.01	.00	1.00	.01	.00	1.00	.00	.00
DINO	HQ	True	1.00	.00	.00	1.00	.02	.00	1.00	.00	.00	1.00	.02	.00	1.00	.03	.00	1.00	.00	.00
		Over	1.00	.02	.00	1.00	.04	.00	1.00	.00	.00	1.00	.03	.00	1.00	.03	.00	1.00	.00	.00
		Under	.87	.00	.00	.93	.01	.00	.69	.00	.00	.91	.03	.00	.92	.04	.00	.83	.00	.00
		O & U	.97	.09	.00	.98	.12	.00	.87	.00	.00	.96	.10	.02	.96	.13	.00	.93	.08	.00
	MQ	True	1.00	.01	.00	1.00	.02	.00	.96	.00	.00	1.00	.01	.00	1.00	.04	.00	1.00	.00	.00
		Over	1.00	.01	.00	1.00	.02	.00	.95	.00	.00	1.00	.04	.00	1.00	.07	.00	1.00	.00	.00
		Under	.80	.01	.00	.88	.04	.00	.27	.00	.00	.91	.03	.00	.96	.04	.00	.56	.00	.00
		O & U	.94	.15	.00	.95	.17	.00	.62	.00	.00	.95	.20	.00	.96	.25	.00	.75	.04	.00
	LQ	True	.95	.02	.00	.97	.04	.00	.58	.00	.00	1.00	.02	.00	1.00	.03	.00	.92	.00	.00
		Over	.94	.03	.00	.97	.04	.00	.54	.00	.00	1.00	.03	.00	1.00	.07	.00	.90	.00	.00
		Under	.47	.00	.00	.64	.00	.00	.07	.00	.00	.69	.03	.00	.78	.03	.00	.12	.00	.00
		O & U	.77	.06	.00	.85	.07	.00	.22	.00	.00	.91	.07	.00	.95	.11	.00	.51	.00	.00

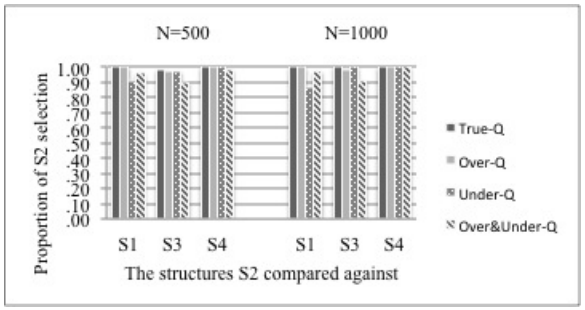
Note. N = sample size; LRT = likelihood ratio test; AIC = Akaike information criterion; BIC = Bayesian information criterion; IQ = item quality; S1:S4 = hierarhchical attribute structures; HQ = higher quality; MQ = mixed quality; LQ = lower quality; True = True Q-matrix; Over = Overspecified Q-matrix; Under=Underspecified Q-matrix; and O & U=Over- and underspecified Q-matrix.

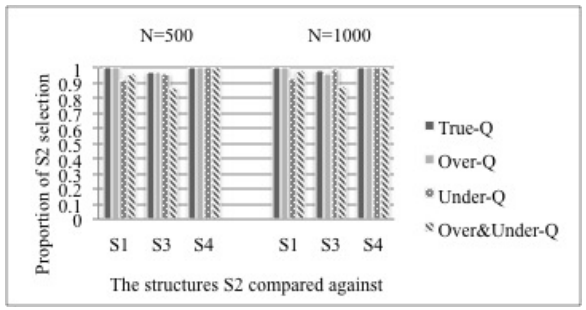
lower panel of the table. For each of the misspecified Q-matrices, items 8 and 10 were manipulated so that 10% of the q-vectors in the Q-matrix were misspecified. Scrolling across the rows of the table presents the results based on smaller and larger sample sizes. Within them, LRT, AIC, and BIC results are given from left to right, respectively. Scrolling down the columns shows results observed under three levels of item qualities. Because generating hierarchy was S^2, a correct structure selection is observed when the model selection methods reject S^1 in favor of S^2 and retain S^2 when it is compared with S^3 and S^4. For example, under the DINA model, higher item quality, and 1000 examinees case, the null hypothesis of more parsimonious model fits the data as well was always rejected when S^1 and S^2 compared when true Q-matrix was used. Similarly, the null hypotheses were always retained when the structured model based on S^2 was compared against structured models of S^3 and S^4.

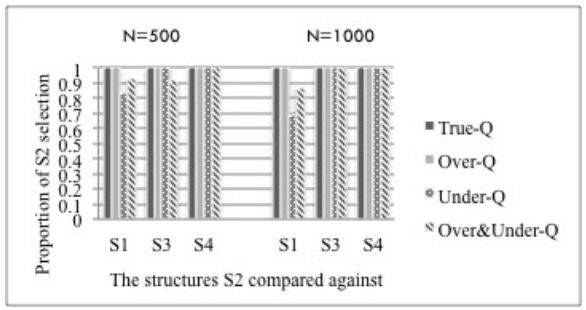
One of the apparent differences was observed between the DINA and DINO model results. In the DINA case, all three model selection methods selected the true structure at least 96% of the time regardless of the sample size and the Q-matrix type. However, impact of sample size and Q-matrix type, as well as model selection method, was substantial in the DINO model. This difference might be due to the generating Q-matrix, which may be more informative for the DINA model. Because the impact of studied factors are clearer in the DINO conditions, only the DINO results will be examined in detail.

Proportion of time the true hierarchy (i.e., S^2) was selected when it was compared against S^1, S^3, and S^4 are presented as bar-graphs in Figure 4.2. The upper panel of the figures shows the results based on higher item quality conditions, whereas the lower part consists of the results of lower item quality conditions. Results based on LRT, AIC, and BIC are given from left to right in both panels. First of all, impact of sample size was much stronger under lower item quality conditions. With larger sample size, all three model selection methods tended to select more strict hierarchy

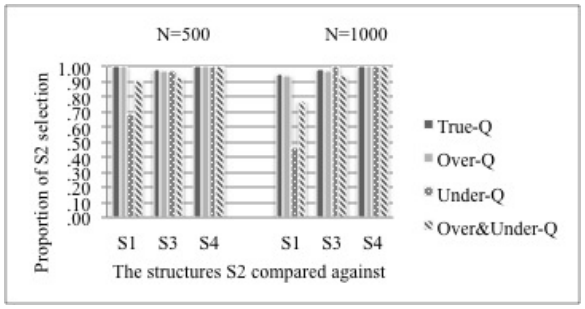
Figure 4.2: Proportion of True Hierarchy Selection

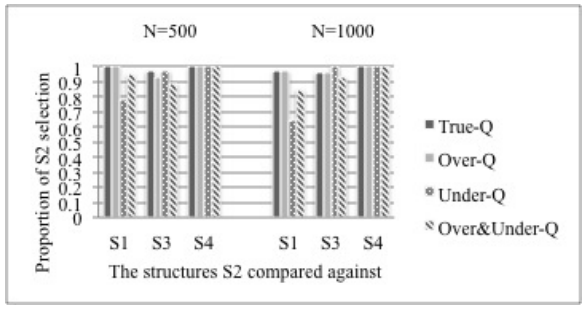


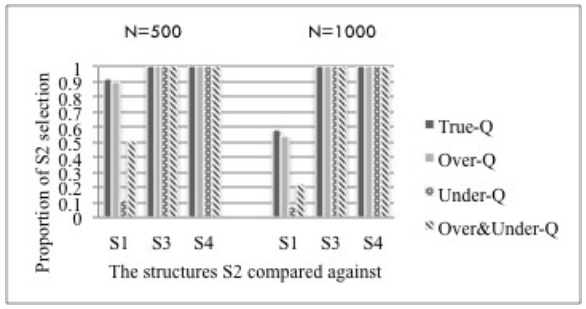




(a) Higher Item Quality







(b) Lower Item Quality

(i.e, S^1). One expected result was that all three model selection methods selected the true hierarchy more often when the difference between the hierarchies, in terms of latent class discrepancy, increased. In other words, LRT, AIC and BIC selected the models based on S^2 more often when it was compared against S^4 than it was compared against S^3.

It is interesting to note that overspecified Q-matrix resulted in just a slight decrease in correct hierarchy selection compared to the true Q-matrix. Yet, when an underspecified Q-matrix was used, the correct hierarchical structure selection substantially decreased across all conditions. The decrease was in the range of .10 to .25 in higher item quality conditions, whereas it varied from .20 to .80 under lower item quality conditions. For example, under lower item quality and 1000 examinees, LRT, AIC, and BIC have selected the true hierarchy among the all four candidates about 98 %, 97%, and 92% with a true Q-matrix, respectively. These percentages dropped down to 66%, 75%, and 12%, respectively, with employment of underspecified Q-matrix. Moreover, impact of over- and underspecified Q-matrix varied within .05 to .15 under higher item quality conditions and within .07 to .45 for lower item quality cases.

Although AIC was superior to LRT, correct attribute structure selection rates of the both model selection methods were comparable. In general, correct hierarchy selection rates of BIC were relatively lower. Nevertheless, BIC tended to select more parsimonious model more often than LRT and AIC. For instance, when over- and underspecified Q-matrix was used, under larger sample size and mixed item quality condition, comparison of S^1 and S^2 resulted in selection of S^2 in 95%, 96%, and 75% based on LRT, AIC, and BIC, respectively. However, these percentages were 80%, 75%, and 96%, respectively, when S^2 was compared against S^3.

Results of Q-matrix Validation

The Q-matrix validation results based on the DINA model are presented in a 2×2 contingency tables in Table 4.5. The contingency tables report the true-positive, true-negative, false-positive, and false-negative rates. False-negative and false-positive rates are analogous to Type-I and Type-II error rates. The true-negative rate shows the proportion of corrected misspecified elements or vectors in the Q-matrix. Further, true-positives indicate the proportion of the correctly specified vectors, which were retained. True-negative and true-positive rates are also referred to as *sensitivity* and *specificity*, respectively (de la Torre & Chiu, 2016).

Although not presented here, the results obtained under the DINO model were similar to the ones obtained under the DINA model. Also, smaller sample size (i.e., $N = 500$) caused substantial reduction in true-negative rates (i.e., sensitivity) of the validation methods. This reduction was the smallest for the *delta method* and the largest for the *general Q-matrix validation method*. Because impact of attribute structure can clearly be seen from the DINA results with 1000 examinees given in Table 4.5, only these results will be discussed in this section. Sensitivity and specificity of the validation methods for hierarchical attribute structure conditions are given on the left hand side of the table, whereas, corresponding rates when attributes are independent are presented at the right hand side.

It should be noted here that these results are at the attribute-vector level. When a suggested q-vector is in between the explicit and implicit counterparts of the generating q-vector, item was counted toward true-positive. For example, for *explicit* generating q-vector of 001000 under the DINA model, suggested q-vectors of 011000, 101000, and 111000 are counted toward true-positive (i.e., validation method retained the q-vector 001000) as the first two attributes are prerequisites for the third attribute. In other words, in order to retain a q-vector in the provisional Q-matrix, the most complex independent attributes in the DINA model must be specified by

Table 4.5: Attribute-Vector Validation Rates by Attribute Structures: The DINA Model

			Hierarchical Attributes						Unstructured Attributes					
			δ		RSS		ς		δ		RSS		ς	
Q-matrix	IQ	Item type	Corr.	Ret.	Corrr.	Ret.	Corr.	Ret.	Corr.	Ret.	Corr.	Ret.	Corr.	Ret.
Over	HQ	Misspecified	.77	.23	.77	.23	.83	.17	.86	.14	.79	.21	.85	.15
		Correct	.01	.99	.04	.96	.02	.98	.02	.98	.03	.97	.02	.98
	MQ	Misspecified	.74	.26	.70	.30	.76	.24	.83	.17	.71	.29	.81	.19
		Correct	.01	.99	.04	.96	.02	.98	.02	.98	.04	.96	.03	.97
	LQ	Misspecified	.71	.29	.63	.37	.69	.31	.74	.26	.63	.37	.74	.26
		Correct	.02	.98	.06	.94	.02	.98	.04	.96	.06	.94	.05	.95
Under	HQ	Misspecified	.72	.28	.72	.28	.52	.48	.82	.18	.76	.24	.79	.21
		Correct	.03	.97	.04	.96	.01	.99	.01	.99	.02	.98	.01	.99
	MQ	Misspecified	.69	.31	.63	.37	.42	.58	.78	.22	.69	.31	.74	.26
		Correct	.03	.97	.07	.93	.01	.99	.01	.99	.03	.97	.02	.98
	LQ	Misspecified	.58	.42	.56	.44	.31	.69	.71	.29	.61	.39	.68	.32
		Correct	.05	.95	.07	.93	.02	.98	.03	.97	.03	.97	.03	.97
O & U	HQ	Misspecified	.74	.26	.65	.35	.53	.47	.77	.23	.72	.28	.80	.20
		Correct	.05	.95	.03	.97	.11	.89	.02	.98	.04	.96	.02	.98
	MQ	Misspecified	.69	.31	.59	.41	.45	.55	.73	.27	.65	.35	.77	.23
		Correct	.05	.95	.05	.95	.14	.86	.03	.97	.06	.94	.02	.98
	LQ	Misspecified	.61	.39	.52	.48	.28	.72	.63	.37	.60	.40	.72	.28
		Correct	.07	.93	.06	.94	.19	.81	.05	.95	.09	.91	.04	.96

Note. δ=delta method; ς^2=general Q-matrix validation method; RSS=Q-matrix refinement method; IQ = item quality; Corr. = corrected; Ret. = retained; Over = Overspecified Q-matrix; Under=Underspecified Q-matrix; O & U=Over- and underspecified Q-matrix; HQ = higher quality; MQ = mixed quality; and LQ = lower quality.

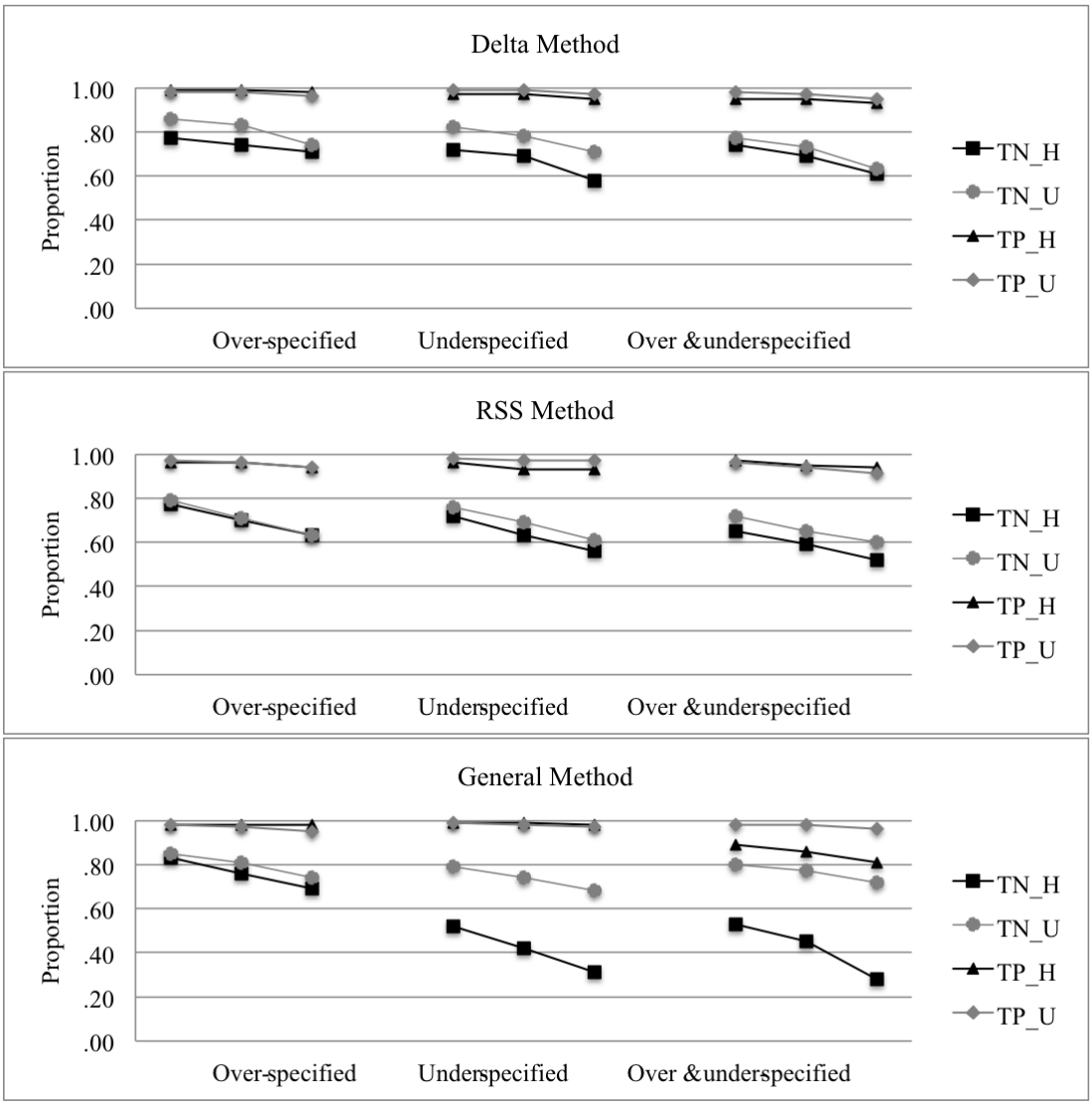
the suggested q-vector. Similarly, a q-vector is retained as long as the most basic independent attribute(s) is/are specified by the suggested q-vector in the DINO model.

Comparison of results of the *delta method* under the hierarchical and unstructured attribute conditions showed that there was 5% to 10% reduction in the methods sensitivity. Observed reductions in over- and underspecified Q-matrix conditions were relatively smaller than the ones observed when all misspecifications were the same type (i.e., either underspecified or overspecified). In other respects, the observed specificity of the delta method was at least .95 under both hierarchical and unstructured attribute conditions. Table 4.5 also displays up to 8% reduction in the sensitivity of the *Q-matrix refinement method* when attributes follow the hierarchy. For this validation method, specificity was well over 90% and the gap between the hierarchical and unstructured attribute conditions were about 2%.

In the *general method of Q-matrix validation*, when only overspecifications occurred, sensitivity of the method slightly reduced. This reduction was augmented by lower item quality (e.g., reduction up to 5%). When underspecified Q-matrix and over- and underspecified Q-matrix were validated, the gap between the sensitivity results of the unstructured and hierarchical attribute structures were much larger. When underspecified Q-matrix was validated, observed sensitivity of the unstructured attribute conditions were .79, .74, and .68 for the higher, mixed, and lower item qualities, respectively. These sensitivity rates were .55, .42, and .31, respectively, under the hierarchical attributes. As can be seen the reduction in the sensitivity was augmented by the decrease in the item quality. Similar results were present for validation of over- and underspecified Q-matrix.

When specificity was considered, reduction was observed only when over- and underspecified Q-matrix was validated. Specifically, .09, .12, and .15 in the specificity rates of general Q-matrix validation method were observed under higher, mixed, and

Figure 4.3: Proportion of Sensitivity and Specificity



lower item qualities, respectively. In all other conditions, specificity rates of the hierarchical and unstructured attributes did not differ much, where observed specificity rates were about and over .95.

The sensitivity and specificity of the three validation methods were compared in Figure 4.3. The figure consists of three sets of graphs created for three different validation methods. Each of the graphs has six lines indicating the sensitivity and specificity rates of validation method under the higher, mixed, and lower item quality conditions. Based on these graphs we can conclude that impact of item quality was

only substantial for the general Q-matrix validation method, where the reduction in the sensitivity increased with the lower item quality. In all other conditions the lines based on hierarchical and unstructured attributes appears to be parallel. The graphs also show that specificity rates of the validation methods under the two attribute structures were generally high, and not too different.

As depicted in the figure, the sensitivity rates of the validation methods were not similar under two different attribute structures. For the delta method and statistical refinement, hierarchical attribute structure resulted up to 10% reduction in the sensitivity rates. The gap between the hierarchical and unstructured attributes was even larger for the general Q-matrix validation method. However, the sensitivity of the method when over-specified Q-matrix was used was high in both attribute structures – up to 37% and 44% reductions were observed under hierarchical attribute conditions when under-specified and over- and under-specified Q-matrices were employed.

4.6 Conclusion and Discussion

This study was conducted to report on; (1) the impact of Q-matrix misspecification on attribute hierarchy selection and (2) the performance of recently proposed and well accepted Q-matrix validation methods under the hierarchical attribute structures. Two simulation studies were conducted to accomplish these purposes, and the simulation results presented in the previous section. Results of the first simulation study showed that underspecified Q-matrix substantially decreased the correct attribute hierarchy selection rates under the DINO model. Although all three misspecified Q-matrix types downgraded the correct hierarchy selection, negative impact of overspecified Q-matrix was minimum. It was also noted that BIC's performance was relatively poor in comparison to LRT and AIC.

In the light of the results of the second simulation study, it can be concluded that performance of the Q-matrix validation methods were relatively poor when attribute are hierarchically structured. Although the specificity of the methods did not change much, their sensitivities, especially of the general Q-matrix validation method, substantially reduced. Change in the methods' sensitivity also varied across the types of Q-matrix misspecifications. Reduction in the sensitivity caused by the hierarchical attribute structure was minimum when the Q-matrix was over-specified.

Given the fact that Q-matrix correction performance of all three Q-matrix validation methods are fully or partially affected under hierarchical attribute structure, modifications to these methods to improve their performance under hierarchies might be a potential future direction for Q-matrix validation research. It would be interesting to see whether their performance increases when the search algorithms described in de la Torre (2008) and de la Torre & Chiu (2016) are modified so that only the latent classes allowed by the hierarchical structures are used in the model estimation. Moreover, the implementation of the Q-matrix refinement method can also be carried out with a modification in the algorithm described in Chiu (2013). Modification can be applied to *Step* 1 in the algorithm so that the nonparametric classification method (Chiu & Douglas, 2013) estimates examinees' class memberships using the ideal response patterns corresponding only to the permissible latent classes.

4.7 References

Barnes, T. (2010). Novel derivation and application of skill matrices: The q-matrix method. In C. Ramero, S. Vemtora, M. Pechemizkiy, & R. S. J. de Baker (Eds.), *Handbook on educational data mining* (pp. 159-172). Boca Raton, FL: Chapman & Hall.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37,* 598-618.

Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification, 30,* 225-250.

DeCarlo, L. T. (2010). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35,* 8-26.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36,* 447-468.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA Model: Development and applications. *Journal of educational measurement, 45,* 343-362.

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33,* 163-183.

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

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76,* 179-199.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81,* 253-273.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69,* 333-353.

Doornik, J. A. (2011). *Object-oriented matrix programming using Ox (Version 6.20).* London: Timberlake Consultants Press.

Feng, Y. (2013). *Estimation and Q-matrix validation for diagnostic classification models* (Unpublished Doctoral Dissertation). University of South Carolina: Columbia, South Carolina.

Hartz, S. M., & Roussos, L. A. (2005). *The fusion model for skills diagnosis: Blending theory with practice.* ETS Research Report. Princeton, NJ: Educational Testing Service.

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.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuokas rule-space approach. *Journal of Educational Measurement, 41,* 205-237.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied psychological measurement, 36,* 548-564.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60,* 523-547.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64,* 187-212.

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

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods, 11(3),* 287-305.

Zheng, Y. & Chiu, C.-Y. (2014). NPCD: Nonparametric methods for cognitive diagnosis. R package version 1.0-7. http://CRAN.R-project.org/package=NPCD

4.8 Appendices

Appendix 4A: Permissible Latent Classes for Three Hypothetical Structures

	Attributes						Structures				Attributes						Structures		
LC	A1	A2	A3	A4	A5	A6	S^1	S^2	S^3	LC	A1	A2	A3	A4	A5	A6	S^1	S^2	S^3
$\boldsymbol{\alpha}_1$	0	0	0	0	0	0	✓	✓	✓	$\boldsymbol{\alpha}_{33}$	1	0	0	0	0	0	✓	✓	✓
$\boldsymbol{\alpha}_2$	0	0	0	0	0	1				$\boldsymbol{\alpha}_{34}$	1	0	0	0	0	1			
$\boldsymbol{\alpha}_3$	0	0	0	0	1	0				$\boldsymbol{\alpha}_{35}$	1	0	0	0	1	0			
$\boldsymbol{\alpha}_4$	0	0	0	0	1	1				$\boldsymbol{\alpha}_{36}$	1	0	0	0	1	1			
$\boldsymbol{\alpha}_5$	0	0	0	1	0	0				$\boldsymbol{\alpha}_{37}$	1	0	0	1	0	0			
$\boldsymbol{\alpha}_6$	0	0	0	1	0	1				$\boldsymbol{\alpha}_{38}$	1	0	0	1	0	1			
$\boldsymbol{\alpha}_7$	0	0	0	1	1	0				$\boldsymbol{\alpha}_{39}$	1	0	0	1	1	0			
$\boldsymbol{\alpha}_8$	0	0	0	1	1	1				$\boldsymbol{\alpha}_{40}$	1	0	0	1	1	1			
$\boldsymbol{\alpha}_9$	0	0	1	0	0	0				$\boldsymbol{\alpha}_{41}$	1	0	1	0	0	0			
$\boldsymbol{\alpha}_{10}$	0	0	1	0	0	1				$\boldsymbol{\alpha}_{42}$	1	0	1	0	0	1			
$\boldsymbol{\alpha}_{11}$	0	0	1	0	1	0				$\boldsymbol{\alpha}_{43}$	1	0	1	0	1	0			
$\boldsymbol{\alpha}_{12}$	0	0	1	0	1	1				$\boldsymbol{\alpha}_{44}$	1	0	1	0	1	1			
$\boldsymbol{\alpha}_{13}$	0	0	1	1	0	0				$\boldsymbol{\alpha}_{45}$	1	0	1	1	0	0			
$\boldsymbol{\alpha}_{14}$	0	0	1	1	0	1				$\boldsymbol{\alpha}_{46}$	1	0	1	1	0	1			
$\boldsymbol{\alpha}_{15}$	0	0	1	1	1	0				$\boldsymbol{\alpha}_{47}$	1	0	1	1	1	0			
$\boldsymbol{\alpha}_{16}$	0	0	1	1	1	1				$\boldsymbol{\alpha}_{48}$	1	0	1	1	1	1			
$\boldsymbol{\alpha}_{17}$	0	1	0	0	0	0				$\boldsymbol{\alpha}_{49}$	1	1	0	0	0	0	✓	✓	✓
$\boldsymbol{\alpha}_{18}$	0	1	0	0	0	1				$\boldsymbol{\alpha}_{50}$	1	1	0	0	0	1			
$\boldsymbol{\alpha}_{19}$	0	1	0	0	1	0				$\boldsymbol{\alpha}_{51}$	1	1	0	0	1	0			✓
$\boldsymbol{\alpha}_{20}$	0	1	0	0	1	1				$\boldsymbol{\alpha}_{52}$	1	1	0	0	1	1			✓
$\boldsymbol{\alpha}_{21}$	0	1	0	1	0	0				$\boldsymbol{\alpha}_{53}$	1	1	0	1	0	0		✓	✓
$\boldsymbol{\alpha}_{22}$	0	1	0	1	0	1				$\boldsymbol{\alpha}_{54}$	1	1	0	1	0	1			
$\boldsymbol{\alpha}_{23}$	0	1	0	1	1	0				$\boldsymbol{\alpha}_{55}$	1	1	0	1	1	0			✓
$\boldsymbol{\alpha}_{24}$	0	1	0	1	1	1				$\boldsymbol{\alpha}_{56}$	1	1	0	1	1	1			✓
$\boldsymbol{\alpha}_{25}$	0	1	1	0	0	0				$\boldsymbol{\alpha}_{57}$	1	1	1	0	0	0	✓	✓	✓
$\boldsymbol{\alpha}_{26}$	0	1	1	0	0	1				$\boldsymbol{\alpha}_{58}$	1	1	1	0	0	1			
$\boldsymbol{\alpha}_{27}$	0	1	1	0	1	0				$\boldsymbol{\alpha}_{59}$	1	1	1	0	1	0		✓	✓
$\boldsymbol{\alpha}_{28}$	0	1	1	0	1	1				$\boldsymbol{\alpha}_{60}$	1	1	1	0	1	1		✓	✓
$\boldsymbol{\alpha}_{29}$	0	1	1	1	0	0				$\boldsymbol{\alpha}_{61}$	1	1	1	1	0	0	✓	✓	✓
$\boldsymbol{\alpha}_{30}$	0	1	1	1	0	1				$\boldsymbol{\alpha}_{62}$	1	1	1	1	0	1			
$\boldsymbol{\alpha}_{31}$	0	1	1	1	1	0				$\boldsymbol{\alpha}_{63}$	1	1	1	1	1	0	✓	✓	✓
$\boldsymbol{\alpha}_{32}$	0	1	1	1	1	1				$\boldsymbol{\alpha}_{64}$	1	1	1	1	1	1	✓	✓	✓

Note. S^1, S^2, and S^3 are three distinct structures; LC represents the possible latent classes; ✓ shows the permissible latent classes; A1 through A6 indicate the six attributes.

Appendix 4B: Ideal Response Patterns: DINA

	Latent classes						Ideal response patterns																	
Structure	α_1	α_2	α_3	α_4	α_5	α_6	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Linear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	0
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Convergent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	0	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	1	0	1	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	1	1
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
	1	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	1
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	0
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Divergent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	0	0	1	1	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
	1	0	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1
	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	1	1
	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	0	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	1	0	1	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0
	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	1	1	1
	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1	1	1
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	0	0	0	0	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	0
	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	0	0	1	1	1
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note. DINA = deterministic input, noisy "and" gate model; and α_1 through α_6 are the attributes.

Appendix 4C: Ideal Response Patterns: DINO

	Latent classes						Ideal response patterns																			
Structure	α_1	α_2	α_3	α_4	α_5	α_6	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Linear	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	1	0
	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	0
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Convergent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	1	0
	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0
	1	1	0	1	0	0	1	1	0	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	0
	1	1	1	0	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Divergent	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	1	0
	1	0	0	1	0	0	1	0	0	1	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	0
	1	0	0	1	1	0	1	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	0
	1	0	0	1	0	1	1	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	1	0	0	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0
	1	1	0	1	0	0	1	1	0	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	0	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1	0
	1	1	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	1	0
	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note. DINO = deterministic input, noisy "or" gate model; and α_1 through α_6 are the attributes.

Chapter 5

Summary

Cognitive model of task performance (Leighton & Gierl, 2007) based assssments aim to identify the attribute mastery status of examinees. These assessments are, in general, called cognitively diagnostic assessment (CDA; de la Torre & Minchen, 2014). CDAs are developed purposefully to serve as formative assessments that may lead to modifications in teaching and learning activities (DiBello & Stout, 2007). Recent political changes intensified the need for formative assessments and led to an increase in the popularity of CDAs. As an offshoot, statistical models to extract diagnostic information from CDAs were proposed. These models are referred to as cognitive diagnosis models (CDMs) or diagnostic classification models (DCMs) (de la Torre & Minchen, 2014).

Attributes assembled into CDA may have hierarchical structure such that mastery of some attributes require mastery of more basic attributes (de la Torre, Hong, & Deng, 2010; Leighton, Gierl, & Hunka, 2004; Templin & Bradshaw, 2014). When this is the case, hierarchical structure needs to be taken into account; otherwise, calibration results of the conventional CDMs may be biased or less accurate, which in turn may result in less accurate attribute mastery profiles. As such they may not be appropriate and useful (Templin & Bradshaw, 2014).

With a general aim to address importance of consideration of hierarchical attribute structure in cognitive diagnosis modeling framework, this dissertation studied estimation approaches that can be employed under hierarchical attribute structures. These estimation approaches are obtained by using structured (i.e., constrained) and

unstructured (i.e., unconstrained) versions of a Q-matrix and prior distribution. To determine the impact of structured Q-matrix and prior distribution on item parameter estimation and examinee classification, a simulation study was carried out in the first study. The results showed that more accurate and precise item parameter estimates were obtained when either the Q-matrix or prior distribution was structured.

Furthermore, performance of estimation approaches on examinee classification were also investigated in the first study. Although structured Q-matrix resulted in higher attribute and vector-level attribute estimation in the DINO case, it yielded lower attribute and vector level attribute estimation under the DINA model. Results also indicated that both the structured and unstructured versions of the Q-matrix yielded identical item parameter estimation and examinee classification when prior distribution was structured. The highest attribute and vector-level correct classification rates were obtained with structured prior distributions where only latent classes allowed by the hierarchy were involved in estimation.

In the first study, prior distribution and Q-matrix were structured based on known hierarchies. However, hierarchical structure of attributes may not be known in real world applications although it needs to be correctly specified; otherwise, an incorrect hierarchy may substantially degrade estimation accuracy. Deriving a hierarchy based on expert opinions and verbal data analyses are the current practices (Cui & Leighton, 2009; Gierl, Zheng, & Cui, 2008). However, both approaches are subjective so that they may result in multiple attribute hierarchies. The second study of this dissertation addressed this subjectivity by proposing a model-fit based empirical exhaustive search algorithm to identify hierarchical relationships among the attributes. The viability of the model selection methods for hierarchy selection was also examined in the second study.

Results of the second study showed that the LRT based exhaustive search successfully generates an R-matrix that specifies all prerequisite relationships when

more complex attributes have only one direct prerequisite attribute (e.g., in linear or divergent hierarchies). However, when a more complex attribute is mastered as long as one of its multiple prerequisites is mastered (e.g., convergent hierarchy), the search algorithm yields in an R-matrix that allows some additional latent classes in the estimation procedure. In other words, search algorithm produces a more liberal hierarchy than the true one. Even in such cases, exhaustive search eliminates many of the non-existing latent classes. It should be noted that likelihood-based exhaustive search algorithm is a tool that can be used along with conventional hierarchy identification methods. This study further showed that model selection based on LRT, AIC, and BIC are potentially viable in the selection of the correct attribute hierarchy when several alternative are available. Performance of LRT and AIC is sufficiently high; in contrast, BIC works better in search algorithm, but it does poorly in hierarchy selection.

CDM implementation requires construction of a Q-matrix that embodies the cognitive specifications in test construction (Leighton et al., 2004). Q-matrix needs to be correctly specified to obtain maximum information from a CDM estimation (de la Torre, 2008). Construction of a Q-matrix heavily depends on content expert opinions and this subjective process may result in misspecifications in the Q-matrix. Empirical Q-matrix validation methods developed for Q-matrix correction have been tested in variety of conditions using either simulated or real data sets. The third study of this dissertation was carried out to report viability of validation methods under hierarchical attribute structures. Based on the simulation results, it was observed that performance of all examined Q-matrix validation methods was more or less lower under hierarchical attribute structures. The sensitivity of the validation methods, especially of the general Q-matrix validation method, significantly decreased. The sensitivities were least affected when the Q-matrix was overspecified. In comparison, there was not much difference in validation methods' specificity across the conditions

involving hierarchical and nonhierarchical attribute structures.

The second purpose of the third study was to report on the impact of Q-matrix misspecification on attribute hierarchy selection. A simulation study was conducted to accomplish this purpose. Results showed that although all types of Q-matrix misspecifications decreased the correct hierarchy selection rates, the observed drop was minimum when Q-matrix overspecified; whereas the maximum decrease was observed with underspecified Q-matrix.

Considering the three studies as a whole, this dissertation showed the importance of taking attribute hierarchy into account in CDM implementations. The first study demonstrated the extent to which a structured Q-matrix or prior distribution was useful in hierarchical attribute cases; the second study showed that model-fit based exhaustive search can be a useful tool in attribute hierarchy identification procedures; and the third study demonstrated that, when it is not taken into account, hierarchical attribute structure not only harms item parameter estimation and examinee classification, but also endangers Q-matrix validation practices.

Modifications to the Q-matrix validation methods to improve their performances under hierarchies might be a direction for future Q-matrix validation studies. For instance, search algorithms described in de la Torre (2008) and de la Torre & Chiu (2016) can be modified so that only permissible latent classes defined by the hierarchy are used in the model estimation procedure. Similarly, Q-matrix refinement method can also be modified such that the nonparametric classification method (Chiu & Douglas, 2013) estimates class-memberships based on ideal response patterns of the permissible latent classes. Another future research direction is to extend these studies to more general CDMs.

5.1 References

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

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47,* 227-249.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicologia Educative, 20,* 89-97.

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44,* 285-291.

Gierl, M. J., Zheng, Y., & Cui, Y. (2008). Using the attribute hierarchy method to identify and interpret cognitive skills that produce group differences. *Journal of Educational Measurement, 45*, 65-89.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26 (2),* 3-16.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuokas rule-space approach. *Journal of Educational Measurement, 41,* 205-237.

No Child Left Behind Act of 2001, Pub. L. No. 1-7-110 (2001).

Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*, 317-339.