A SEQUENTIAL COGNITIVE DIAGNOSIS MODEL FOR GRADED RESPONSE: MODEL DEVELOPMENT, Q-MATRIX VALIDATION, AND MODEL COMPARISON

BY WENCHAO MA

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, 2017

ABSTRACT OF THE DISSERTATION

A Sequential Cognitive Diagnosis Model for Graded Response: Model Development, Q-Matrix Validation, and Model Comparison

by Wenchao Ma

Dissertation Director: Jimmy de la Torre

Cognitive diagnosis models (CDMs) have received increasing attention in recent years. The goal of CDMs is to classify examinees into different latent classes with unique attribute patterns indicating mastery or nonmastery on a set of skills or attributes of interest. Although a large number of CDMs can be found in the literature, most of them are developed for dichotomous response data.

This dissertation proposes a general cognitive diagnosis model for a special type of polytomously scored items, where item categories are attained in a sequential manner, and explicitly associated with some attributes. The conditional probability of answering a category correctly given that the previous categories have been performed successfully is defined as *processing function*, and modeled using the generalized deterministic inputs, noisy “and” gate (G-DINA; de la Torre, 2011) model. The resulting model is referred to as the *sequential* G-DINA model. To relate response categories to attributes, a category-level Q-matrix is used. When the attribute and category association is specified a priori, the proposed model has the flexibility to allow different cognitive processes (e.g., conjunctive, disjunctive) to be modeled at different steps within a single item. This model can be extended for items, where categories cannot be explicitly linked to attributes, and for items with unordered categories. Item parameters of the proposed model are estimated using the marginal maximum likelihood estimation via expectation-maximization algorithm.

Like the traditional Q-matrix, the category-level Q-matrix is most likely to be developed by experts, and thus tends to be subjective. In this dissertation, a Q-matrix validation procedure is developed for the sequential G-DINA model to empirically identify and correct misspecifications in the category-level Q-matrix. This validation method is implemented in a stepwise manner based on the Wald test and an item discrimination index. Simulation studies are conducted to evaluate the performance of the proposed procedure in terms of the true positive and false positive rates.

A condensation rule is an important component for most CDMs, including the sequential G-DINA model, in that it specifies how the latent attributes are employed simultaneously to make a manifest item response. Although the G-DINA model has been used as the processing function, it is important to empirically determine whether the G-DINA model can be further constrained according to the cognitive processes involved in each step. In this dissertation, the performance of the Wald test and the likelihood ratio test are examined in determining the appropriate condensation rule for each step. More specifically, a simulation study is used to evaluate the Type I error and power of these hypothesis tests concerning whether the DINA model, DINO model, and *A*-CDM can be used in place of the G-DINA model as the processing function for the steps that involved more than one attribute.

Taken together, this dissertation develops a set of psychometric tools including statistical models and procedures for graded response data. These tools can facilitate the use of constructed-response items, which are typically scored polytomously, in cognitively diagnostic assessments. The performance of the proposed models and procedures are examined using both Monte Carlo simulation studies and real data.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Jimmy de la Torre, for his unwavering support, enlightening guidance, great patience, and continuous encouragement throughout my graduate studies at Rutgers. I was so fortunate to find an advisor who always has time for listening to my questions and offering suggestions. His editorial advice was also indispensable to the completion of this dissertation, as well as many of my other studies. I also appreciate his financial support for my graduate study and insightful advice on my career.

I would also like to extend my appreciation to my committee members: Dr. Chia-Yi Chiu, Dr. Drew Gitomer and Dr. Matthias von Davier, for their insightful comments and suggestions to my dissertation, and for their strong support in general.

I am grateful to my colleague Charlie Iaconongelo for proofreading many of my manuscripts. I would also like to thank my other ESM colleagues: Mehmet Kaplan, Lokman Akbay, Soo Lee, Immanuel Williams, Nathan Minchen, Eugene Geis, Ragip Terzi, Yan Sun and Yanhong Bian, as well as Miguel Sorrel, Levent Yakar, Kevin Carl Santos, Huiqin He, Hueying Tzou, for having many interesting and high-spirited discussions in Room 304. My thanks also go to Dr. Hao Song and Dr. Yuan Hong, who gave me chances to broaden my experience through internships.

Last, I would like to thank my family. The greatest support came from my loving wife, who makes my graduate study a wonderful journey. I am indebted to my father, my sister, my brother-in-law, and my parents-in-law, for their many years of support. Special thanks go to my passed mom, who did not have a chance to be well-schooled, but always encouraged me towards excellence.

Dedication

This dissertation is dedicated to my beloved wife

♥ Wenjing Guo ♥

for her whole-hearted support, and

for always being considerate and thoughtful, caring and encouraging.

I am truely thankful for having her in my life.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

1.1 Background

Educational assessments have played an increasing central role in evaluating students' learning. By establishing students' relative rankings along a proficiency continuum based on either item response theory (IRT) or classical test theory, a number of educational decisions, such as identifying students' level of mastery, offerring students scholarships, and examining students' readiness for further study, can be made accordingly (de la Torre & Minchen, 2014). In spite of the use of educational assessments for such purposes, they typically provide little information to facilitate classroom instruction. Many researchers such as Stiggins (2002) argue that educational assessment should be used to not only evaluate learning, but also support learning.

Cognitively diagnostic assessments (CDAs; de la Torre & Minchen, 2014) are designed to provide immediate diagnostic information to teachers and students so that the classroom instructions can be planned or modified accordingly. Unlike conventional large-scale educational assessments, the grain size of the skills or attributes to be measured in CDAs are typically finer, but the number can be much larger. As a result, it is usually reasonable to assume that there are only two statuses for each attribute: mastery or nonmastery. The goal of CDAs is to diagnose whether students have already mastered each attribute or not. Students with the same total score can have different attribute profiles, which offers additional finer-grained information about students' strengths and weaknesses.

To extract reliable and valid information from CDAs, cognitive diagnosis models (CDMs) have been developed. CDMs refer to a large set of psychometric approaches and have been named differently to emphasize their different aspects, such as restricted latent class models (e.g., Haertel, 1989), multiple classification latent class models (e.g., Maris, 1999), structured item response theory models (e.g., Rupp & Mislevy, 2007), structured located latent class models (e.g., Xu & von Davier, 2008), and diagnostic classification models (Rupp, Templin, & Henson, 2010). In this dissertation, CDM is used consistently to refer to the discrete-skills IRT-based latent class models as discussed in Roussos, Templin, and Henson (2007). Other diagnostic procedures will not be covered, such as the rule-space method (Tatsuoka, 1983), the attribute hierarchy method (Leighton, Gierl, & Hunka, 2004), Bayesian network models (Almond, DiBello, Moulder, & Zapata-Rivera, 2007), nonparametric approaches (Chiu & Douglas, 2013; Chiu, Douglas, & Li, 2009), and traditional IRT models used for diagnostic purposes (Stout, 2007).

A number of CDMs can be found in literature. Some of them are developed based on strong cognitive assumptions about the processes involved in problem solving, including, among others, the deterministic inputs, noisy "and" gate (DINA; Haertel, 1989) model, the deterministic inputs, noisy "or" gate (DINO; Templin & Henson, 2006) model, and the *additive* CDM (*A*-CDM; de la Torre, 2011). To better understand the relation among these models and make it possible to estimate various models within a single test, some general CDMs have been developed, such as the generalized DINA (G-DINA; de la Torre, 2011) model, the log-linear CDM (LCDM; Henson, Templin, & Willse, 2009) and the general diagnostic model (GDM; von Davier, 2008).

Two central components shared by these models are Q-matrix (Tatsuoka, 1983) and condensation rule (Maris, 1999). The former specifies the association between attributes and items. It is typically created by experts (see Tjoe & de la Torre, 2014, for a detailed development process of a Q-matrix for a proportional reasoning test), and assumed to be known in most CDM analyses. The latter specifies how attributes are "condensed" to produce an observed item response. For example, the DINA model is a conjunctive model assuming that to answer an item correctly, students are supposed to master all required attributes; whereas the DINO model is a disjunctive model assuming that students are expected to perform an item successfully as long as they have already mastered at least one required attribute.

Both Q-matrix and condensation rule are usually specified by experts before the CDM analysis and assumed to be correct. However, expert judgments can be subjective and their specifications may not always be accurate. A body of procedures have been developed to validate the Q-matrix (de la Torre, 2008; de la Torre & Chiu, 2016; Chiu, 2013; Liu, Xu, & Ying, 2013) and to select appropriate condensation rules (de la Torre, 2011; de la Torre & Lee, 2013; Ma, Iaconangelo, & de la Torre, 2016) empirically.

1.2 Motivation and Objectives

Despite a host of CDMs available, most of them are targeted for dichotomous responses, which are mainly from multiple-choice items. Constructed response items, however, may be more informative for diagnostic purposes in that students are allowed to show their problem-solving solutions explicitly. For example, Birenbaum and Tatsuoka (1987) found that constructed-response items were more appropriate for the diagnostic purpose by comparing a fraction addition test using open-ended and multiple-choice formats for diagnosing students' misconceptions. Similar conclusions have been drawn by Birenbaum, Tatsuoka, and Gutvirtz (1992), who also found that students used different cognitive processes when responding to items with different formats. For example, students may not really solve the problem in multiple-choice format as expected, but try to utilize the information in the options, which, sometimes, makes them more likely to achieve an incorrect solution.

Typically, although not always, constructed response items are scored polytomously, yielding graded responses with ordered categories. To calibrate this type of data, one commonly used strategy is to dichotomize responses so that they can be analyzed using existing dichotomous CDMs (e.g., Johnson et al., 2013; Su, 2013). However, the process of dichotomization often leads to loss of information. To handle polytomously scored items more appropriately, a few polytomous CDMs have been developed, such as the partial credit DINA model (de la Torre, 2010), the GDM for graded responses (von Davier, 2008), nominal response diagnostic model (Templin, Henson, Rupp, Jang, & Ahmed, 2008), and polytomous LCDM (Hansen, 2013), among others. A limitation shared by these polytomous models is that the required attributes for an item are assumed to be involved in all categories of the item, which, however, is not always the case.

The first major goal of this dissertation is to develop a new general cognitive diagnosis model, referred to as the sequential G-DINA model, for polytomously scored items that can overcome the aforementioned limitation. In particular, items that need to be solved through a sequence of steps are considered, such as $\sqrt{7.5/0.3-16}$ (Masters, 1982). Unlike other existing polytomous response CDMs, the sequential G-DINA model relaxes the assumption that all categories involve the same attributes, and takes the step and attribute association into consideration.

The sequential G-DINA model, like most other existing CDMs, relies on a Q-matrix. However, to consider the step and attribute association, the Q-matrix needs to be defined at the step level. No matter how the Q-matrix is defined, developing a Q-matrix by domain experts without any misspecifications remains challenging, and therefore, validating the Q-matrix empirically to identify the potential misspecified elements based on the collected data is an important research topic. Despite a few Q-matrix validation procedures available, none of them is developed for polytomous models.

The second goal of this dissertation is to develop a Q-matrix validation procedure that can be used along with the sequential G-DINA model for graded response items. Since the Q-matrix is defined at the step level for the proposed model, the Q-matrix validation method is able to identify which attributes are involved for each step. This Q-matrix validation procedure employs a formal hypothesis test, as well as an effect size measure. It is implemented step by step, and item by item.

Apart from the Q-matrix development, determining the appropriate condensation rule is another important but difficult task for domain experts. The fact that a large number of CDMs are available, on one hand, allows great flexibility in modeling complex cognitive processes involved in the problem solving, but, on the other hand, has also led uncertainties as to which model is the most suitable for an item. Choosing an appropriate model for an item should ensure that its condensation rule is in line with the way that students solve the problem. The use of an inappropriate condensation rule yields model misspecifications, which can result in questionable validity of further inference, as well as poor person attributes estimation (Rojas, de la Torre, & Olea, 2012). Constructed-response items that require students show their work explicitly enable us to better understand how students solve each question, but different steps of the problem-solving may involve different cognitive processes, and thus call for models with different condensation rules.

The third goal of this dissertation is to evaluate whether the Wald test and likelihood ratio test can be used in conjunction with the sequential G-DINA model to select appropriate condensation rule for each step. The Wald test has been previously used for this purpose for dichotomous response data (de la Torre & Lee, 2013; Ma et al., 2016), but it shows inflated Type I error under some conditions. This dissertation considers the Wald test using variance-covariance matrices calculated in various ways.

1.3 References

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement*, *44*, 341–359.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats - It does make a difference for diagnostic purposes. *Applied Psychological Measurement*, *11*, 385–395.

Birenbaum, M., Tatsuoka, K. K., & Gutvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement*, *16*, 353–363.

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.

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

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. (2010, July). *The partial-credit DINA model.* Paper presented at the international meeting of the Psychometric Society, Athens, GA.

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., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*, 355–373.

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

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321.

Hansen, M. (2013). *Hierarchical item response models for cognitive diagnosis* (Unpublished doctoral dissertation). University of California at Los Angeles.

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

Johnson, M., Lee, Y.-S., Sachdeva, R. J., Zhang, J., Waldman, M., & Park, J. Y. (2013, April). *Examination of gender differences using the multiple groups DINA model.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, California.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's Rule-Space approach. *Journal of Educational Measurement*, *41*, 205–237.

Liu, J., Xu, G., & Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli*(5A), 1790–1817.

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*, 200–217.

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

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Rojas, G., de la Torre, J., & Olea, J. (2012). *Choosing between general and specific cognitive doagnosis models when the sample size is small.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, Vancouver, British Columbia.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills Diagnosis Using IRT-Based Latent Class Models. *Journal of Educational Measurement*, *44*, 293-311.

Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response models. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theories and applications* (pp. 205–240). Cambridge University Press.

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

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*, 758–765.

Stout, W. (2007). Skills diagnosis using irt-based continuous latent trait models. *Journal of Educational Measurement*, *44*, 313–324.

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

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

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

Templin, J. L., Henson, R. A., Rupp, A. A., Jang, E., & Ahmed, M. (2008, March). *Cognitive diagnosis models for nominal response data*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

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.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.

Xu, X., & von Davier, M. (2008). Fitting the structured general diagnostic model to naep data. *ETS Research Report Series*, 1–18.

Chapter 2

A Sequential Cognitive Diagnosis Model for Polytomous Responses

Abstract

This paper proposes a general polytomous cognitive diagnosis model for a special type of graded responses, where item categories are attained in a sequential manner, and associated with some attributes explicitly. To relate categories to attributes, a category-level Q-matrix is used. When the attribute and category association is specified a priori, the proposed model has the flexibility to allow different cognitive processes (e.g., conjunctive, disjunctive) to be modeled at different categories within a single item. This model can be extended for items, where categories cannot be explicitly linked to attributes, and for items with unordered categories. The feasibility of the proposed model is examined using simulated data. The proposed model is illustrated using the data from TIMSS 2007 assessment.

Note

This chapter is a reprint of the following publication with format adjustment:

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*, 253–275.

2.1 Introduction

Cognitive diagnosis models (CDMs) have received increasing attention recently. The goal of CDMs is to classify examinees into different latent classes with unique attribute patterns indicating mastery or nonmastery of a number of skills or attributes of interest. Students with the same total score according to item response theory (IRT) or classical test theory (CTT) can have different attribute patterns, which offers additional information about students' strengths and weaknesses thus informing instruction and remediation.

A host of CDMs can be found in the literature (for reviews, see DiBello, Roussos, & Stout, 2007; Rupp & Templin, 2008), and many of them are developed based on strong cognitive assumptions about the processes involved in problem-solving. For example, the deterministic inputs, noisy "And" gate (DINA; Haertel, 1989) model assumes that examinees are expected to answer an item correctly only when they possess all required attributes; whereas, the deterministic inputs, noisy "Or" gate (DINO; Templin & Henson, 2006) model assumes that, in principle, examinees are able to perform an item successfully as long as they master at least one required attribute. Some general CDM frameworks subsuming a number of commonly used CDMs have also been developed, such as the generalized DINA (G-DINA; de la Torre, 2011) model, the log-linear CDM (LCDM; Henson, Templin, & Willse, 2009) and the general diagnostic model (GDM; von Davier, 2008). Although developed from different perspectives, the G-DINA model and the LCDM are equivalent in their saturated forms, both of which are special cases of the GDM.

In spite of a number of CDMs available, most of them are targeted for dichotomous responses that stemmed primarily from multiple-choice items. The importance of constructed-response items, nevertheless, has been largely overlooked in cognitive diagnostic assessments recently. Theoretically, the constructed-response items are probably able to provide more evidence to support the inference about examinees' attribute patterns because they require examinees to explicitly show their problem-solving procedures. The merits of constructed-response items have also been empirically recognised. For example, Birenbaum and Tatsuoka (1987) administered a fraction addition test using open-ended and multiple-choice formats to diagnose students' misconceptions and found that open-ended items were more appropriate for the diagnostic purpose according to various criteria, such as the number of identified students' error types and diagnosis of students' sources of misconceptions. This conclusion has been further examined and verified by Birenbaum, Tatsuoka, and Gutvirtz (1992), who also found that students used different cognitive processes when responding to items with different formats. For example, students may not really solve the problem in multiple-choice format as expected, but try to utilize the information in alternatives, which, sometimes, makes them more likely to achieve an incorrect solution.

Typically, although not always, constructed-response items are scored polytomously, yielding graded response data with ordered categories. To calibrate this type of data, one commonly used strategy is to dichotomize them so that they can be analyzed using existing dichotomous CDMs (e.g., Johnson et al., 2013; Su, 2013). However, the process of dichotomization often results in loss of information. To deal with polytomously scored items more appropriately, a few polytomous CDMs have been developed, such as the partial credit DINA (PC-DINA; de la Torre, 2010) model, the GDM for graded responses (pGDM; von Davier, 2008), nominal response diagnostic model (NRDM; Templin, Henson, Rupp, Jang, & Ahmed, 2008), and polytomous LCDM (Hansen, 2013). However, none of these polytomous CDMs consider the possible relation between attributes and response categories. Unlike polytomous IRT models where the latent trait has an impact on students' responses to all categories, in CDMs, different categories could measure different attributes, as shown in an example in the next section. To take this information into account, a general polytomous CDM for graded responses has been developed in this paper. This model is referred to as the *sequential process model* to emphasize that a series of attributes is involved in the problem-solving process.

2.2 Attribute and Category Association

Suppose solving an item consists of a finite number of sequential steps, each of which involves some attributes. Also, suppose that students are scored according to how many successive steps they have successfully performed. Specifically, a student falls into the zero category if s/he fails the first step; the first category if s/he performs the first step correctly but fails the second step; and so forth. In doing so, responses to items with H steps have $H+1$ ordered categories, namely, category zero to category H.

Take $4\frac{1}{8}-\frac{3}{8}$ as an example. To solve this item, three steps may be involved. First, $4\frac{1}{8}$ is transformed to $3\frac{9}{8}$ to allow fraction subtraction; then, by subtracting the numerators of two fractions, $3\frac{6}{8}$ can be obtained; in the last step, $3\frac{6}{8}$ is simplified to $3\frac{3}{4}$. According to the attributes identified by Tasuoka (1990), students need to know (A1) borrow from whole number part, (A2) subtract numerators, and (A3) reduce answers to the simplest form to succeed in step 1, 2 and 3, respectively. This example is for illustrative purposes only, and items in practice can be more complex. For example, some steps could consist of multiple substeps that are not sequential and some substeps may need multiple attributes. Additionally, although response categories are assumed to be attained sequentially, different categories do not have to measure different attributes, nor must the attributes show any particular structure. For example, the attributes measured by lower categories do not have to be prerequisites to those required by higher categories.

To relate attributes to categories, the traditional Q-matrix (Tatsuoka, 1983) has been modified. The traditional Q-matrix is a $J \times K$ binary matrix specifying whether an attribute is measured by an item, where J is the test length and K is the number of attributes. Element q_{jk} at row j and column k is equal to 1 if attribute k is needed by item j, and 0 otherwise. For graded responses, a category level Q-matrix is developed in this paper, referred to as Q_C-matrix, where subscript C is used to denote *category*. Throughout this paper, item j is assumed to have $H_j + 1$ categories (i.e., $0, 1, \ldots, H_j$). The attribute and category association for item j is placed in H_j rows of the Q_C-matrix because category zero does not require any attribute. Each of H_j rows has K elements indicating which attributes are required by the category. In particular, element 1 indicates that the attribute is required by this category, and 0 indicates that the attribute is not. The Q_C-matrix is a $\sum_{j=1}^{J} H_j \times K$ binary matrix, and if all items are scored dichotomously, the Q_C-matrix is equivalent to the traditional Q-matrix.

Table 2.1 gives the Q_C-matrix for the item $4\frac{1}{8} - \frac{3}{8}$. The attribute and category association is specified in three rows to account for four categories. The required attributes for a category refer to the attributes required for the step that examinees need to solve to answer this category correctly after they have completed all previous steps successfully. For example, although the first two attributes are also indispensable to achieve category three, it is not necessary to specify $[1\ 1\ 1]$ because after examinees have already achieved category two, only the third attribute is needed to perform category three correctly. The Q_C-matrix defined in this way is referred to as the *restricted* Q_C-matrix.

To create the restricted Q_C-matrix, the attribute and category association must be known a priori. However, this information may not be available especially when CDMs are retrofitted to existing assessments. If so, it is reasonable to assume that all attributes required by an item are needed by each category of this item. The Q_C-matrix defined

Table 2.1: Restricted Q_C-matrix for $4\frac{1}{8} - \frac{3}{8} =?$

		Attributes		
Step	Category	A1	A2	A3
$3\frac{9}{8} - \frac{3}{8}$	1	1	0	0
$3\frac{6}{8}$	2	0	1	0
$3\frac{3}{4}$	3	0	0	1

Note: A1: borrow from whole number part; A2: subtract numerators; A3: reduce answers to the simplest form.

in this way is called the *unrestricted* Q_C-matrix. For the previous example, the unrestricted Q_C-matrix is given in Table 2.2.

Table 2.2: Unrestricted Q_C-matrix for $4\frac{1}{8} - \frac{3}{8} =?$

		Attributes		
Step	Category	A1	A2	A3
$3\frac{9}{8} - \frac{3}{8}$	1	1	1	1
$3\frac{6}{8}$	2	1	1	1
$3\frac{3}{4}$	3	1	1	1

Note: A1: borrow from whole number part; A2: subtract numerators; A3: reduce answers to the simplest form.

2.3 Sequential Process Model

When a test measures K attributes, examinees can be grouped into 2^K latent classes, each having unique attribute pattern, that is, $\boldsymbol{\alpha}_c = (\alpha_{c1}, \ldots, \alpha_{cK})$, where $c = 1, \ldots, 2^K$. $\alpha_{ck} = 1$ indicates attribute k is mastered by examinees in latent class c, and $\alpha_{ck} = 0$ indicates attribute k is not mastered by examinees in latent class c. Similar to Samejima (1995), we define the probability of examinees with attribute pattern $\boldsymbol{\alpha}_c$ answering category h of item j correctly provided that they have already completed the category $h-1$ successfully as the *processing function* of category h, denoted by $S_j(h|\boldsymbol{\alpha}_c)$, and we can reasonably assume

$$S_j(h|\boldsymbol{\alpha}_c) = \begin{cases} 1, & \text{if } h = 0 \\ 0, & \text{if } h = H_j + 1, \end{cases}$$

because examinees can always achieve category zero, but never achieve category $H_j + 1$. Students score h if and only if they answer category one to h correctly, and if h is not the highest category, category $h+1$ incorrectly; therefore, the category response function for item j can be expressed as

$$P(X_j = h|\boldsymbol{\alpha}_c) = [1 - S_j(h+1|\boldsymbol{\alpha}_c)] \prod_{x=0}^{h} S_j(x|\boldsymbol{\alpha}_c), \tag{2.1}$$

subject to the constraints

$$\sum_{h=0}^{H_j} P(X_j = h|\boldsymbol{\alpha}_c) = 1 \;\; \forall c,$$

where $h = 0, \ldots, H_j$, and $P(X_j = h|\boldsymbol{\alpha}_c)$ is the probability of examinees with attribute pattern $\boldsymbol{\alpha}_c$ scoring h on item j. It is reasonable to assume that the processing function $S_j(h|\boldsymbol{\alpha}_c)$ is a function of examinees' attribute patterns and the required attributes for category h of item j. The processing function is the kernel of the sequential process model, and can be formulated using most dichotomous CDMs. For example, if solving a step needs the possession of all required attributes, the DINA model can be used as the processing function. By parameterizing each category separately, the sequential process model allows different cognitive processes to be modeled at different categories within a single item.

2.3.1 Sequential G-DINA Model

In this paper, the G-DINA model (de la Torre, 2011) is used as the processing function because it offers a general framework subsuming several widely used CDMs. The resulting model is referred to as the *sequential G-DINA model*.

Like the G-DINA model, for item j, 2^K latent classes can be collapsed into $2^{K_j^*}$ latent groups with unique probabilities of success, where K_j^* is the number of required attributes for item j. For category h, $2^{K_j^*}$ latent groups can be further collapsed into $2^{K_{jh}^*}$ latent groups, where K_{jh}^* is the number of required attributes for category h of item j. Let $\boldsymbol{\alpha}_{ljh}^*$ be the reduced attribute vector for category h of item j consisting of the required attributes for this category only, where $l = 1, \cdots, 2^{K_{jh}^*}$. Without loss of generality, we can assume the first K_{jh}^* attributes are required for category h of item j, that is, $\boldsymbol{\alpha}_{ljh}^* = [\alpha_{l1}, \ldots, \alpha_{lk}, \ldots, \alpha_{lK_{jh}^*}]$. The processing function $S_j(h|\boldsymbol{\alpha}_c)$ can be written as $S_j(h|\boldsymbol{\alpha}_{ljh}^*)$, and formulated using the identity link G-DINA model:

$$\begin{aligned} S_j(h|\boldsymbol{\alpha}_{ljh}^*) =& \phi_{jh0} + \sum_{k=1}^{K_{jh}^*} \phi_{jhk}\alpha_{lk} + \sum_{k'=k+1}^{K_{jh}^*} \sum_{k=1}^{K_{jh}^*-1} \phi_{jhkk'}\alpha_{lk}\alpha_{lk'} + \cdots \\ & + \phi_{jh12\cdots K_{jh}^*} \prod_{k=1}^{K_{jh}^*} \alpha_{lk}, \end{aligned} \tag{2.2}$$

where ϕ_{jh0} is the intercept, ϕ_{jhk} is the main effect due to α_{lk}, $\phi_{jhkk'}$ is the two-way interaction effect due to α_{lk} and $\alpha_{lk'}$, and $\phi_{jh12\cdots K_{jh}^*}$ is K_{jh}^*-way interaction effect due to α_{lk} to $\alpha_{lK_{jh}^*}$. ϕ_{jh0} represents the processing function of category h for examinees who master none of required attributes, ϕ_{jhk} is the change of processing function of category h due to the mastery of attribute k, and interaction coefficients represent the change in the processing function of category h due to the mastery of all relevant attributes that is over and above all impact of lower order effects. Like the G-DINA model, the processing function can also be defined using log or logit link function. For category h of item j, there are $2^{K_{jh}^*}$ item parameters, as in, $\boldsymbol{\phi}_{jh} = \{\phi_{jh0}, \phi_{jh1}, \cdots, \phi_{jh12\cdots K_{jh}^*}\}$. By defining processing functions $\boldsymbol{S}_j(h|\boldsymbol{\alpha}_{jh}^*) = \{S_j(h|\boldsymbol{\alpha}_{ljh}^*)\}$, $\boldsymbol{\phi}_{jh}$ can be derived from $\boldsymbol{S}_j(h|\boldsymbol{\alpha}_{jh}^*)$ directly because equation (2) can be expressed as $\boldsymbol{S}_j(h|\boldsymbol{\alpha}_{jh}^*) = \boldsymbol{M}_{jh}\boldsymbol{\phi}_{jh}$, where $\boldsymbol{M}_{jh}$ is an invertible design matrix of dimension $2^{K_{jh}^*} \times 2^{K_{jh}^*}$ (See de la Torre, 2011, for details about the design matrix). This implies that processing functions can also be viewed as item parameters, though this is not true if constraints are added to the processing functions.

As shown by de la Torre (2011), by setting appropriate constraints in the G-DINA model, the DINA model, DINO model, *A*-CDM, linear logistic model (LLM; Maris, 1999) and reduced reparametrized unified model (R-RUM; Hartz, 2002) can be obtained. Those models can also be specified as the processing functions using similar constraints in the sequential G-DINA model. Please refer to de la Torre (2011) for details about the appropriate constraints.

The sequential G-DINA model can use either restricted or unrestricted $\mathrm{Q_C}$-matrix. For notational convenience, the sequential G-DINA model using restricted and unrestricted $\mathrm{Q_C}$-matrix are called restricted and unrestricted sequential G-DINA model, and abbreviated as RS-GDINA model and US-GDINA model, respectively. The use of restricted $\mathrm{Q_C}$-matrix allows us to model different underlying processes at different response categories. The use of unrestricted $\mathrm{Q_C}$-matrix, on the other hand, provides a possible solution to account for the uncertainty in the attribute and category association. When the attribute and category association is available, the RS-GDINA model may be preferred theoretically because it usually estimates fewer item parameters than the US-GDINA model. Regarding the aforementioned example, the RS-GDINA model has six item parameters but the US-GDINA model has 24, which implies that additional 18 parameters for this single item need to be estimated when using the unrestricted $\mathrm{Q_C}$-matrix. Nevertheless, the practical consequence of estimating extra parameters needs further empirical examination.

2.3.2 Parameter Estimation

Item parameters of the sequential G-DINA model can be estimated using the marginal maximum likelihood estimation approach via Expectation Maximization (MMLE/EM) algorithm (Bock & Aitkin, 1981). Let $\boldsymbol{\alpha}^*_{lj}$ be the reduced attribute pattern for the l^{th} collapsed latent group for item j, where $l = 1, \ldots, 2^{K^*_j}$. Also, let X_{ij} be the response of examinee i to item j, where $i = 1, \ldots, N$. Under the assumption of local independence, the conditional probability of the response vector $\boldsymbol{X}_i$ can be written as

$$P(\boldsymbol{X}_i|\boldsymbol{\alpha}^*_{lj}) = \prod_{j=1}^{J}\prod_{h=0}^{H_j} P(X_j = h|\boldsymbol{\alpha}^*_{lj})^{I(X_{ij}=h)},$$

where $I(X_{ij} = h)$ is an indicator variable evaluating whether X_{ij} is equal to h. The MMLE/EM algorithm implements E-step and M-step iteratively item by item until convergence. In particular, for item j, based on the provisional item parameter estimates and the distribution of reduced latent classes $p(\boldsymbol{\alpha}^*_{lj})$, E-step calculates the expected number of examinees with attribute pattern $\boldsymbol{\alpha}^*_{lj}$ scoring in category h, that is,

$$\bar{r}_{ljh} = \sum_{i=1}^{N} I(X_{ij} = h) P(\boldsymbol{\alpha}^*_{lj}|\boldsymbol{X}_i),$$

where $P(\boldsymbol{\alpha}^*_{lj}|\boldsymbol{X}_i)$ is the posterior probability of examinee i having reduced attribute pattern $\boldsymbol{\alpha}^*_{lj}$, and can be calculated by

$$P(\boldsymbol{\alpha}^*_{lj}|\boldsymbol{X}_i) = \frac{P(\boldsymbol{X}_i|\boldsymbol{\alpha}^*_{lj})p(\boldsymbol{\alpha}^*_{lj})}{\sum_{l=1}^{2^{K^*_j}} P(\boldsymbol{X}_i|\boldsymbol{\alpha}^*_{lj})p(\boldsymbol{\alpha}^*_{lj})}.$$

In the M-step, the following object function ℓ needs to be maximized with respect to item parameters $\boldsymbol{\phi}_j$, which is a vector of length $\sum_{h=1}^{H_j} 2^{K^*_{jh}}$ when the step function is the G-DINA model, using some general optimization techniques,

$$\ell = \sum_{l=1}^{2^{K^*_j}} \sum_{h=0}^{H_j} \bar{r}_{ljh} \log\left[\hat{P}(X_j = h|\boldsymbol{\alpha}^*_{lj})\right].$$

Two steps are repeated until convergence. In this study, the Nelder and Mead's (1965) simplex method is used for M-step optimization. It is one of the most popular derivative-free optimization technique applicable for multidimensional nonlinear problems. After generating a geometric simplex, its convergence is guided by moving the simplex appropriately (Nelder & Mead, 1965). It should be noted that although the Nelder and Mead method is robust and the default method for the optim function in R programming language (R Core Team, 2015), other optimization techniques such as quasi-Newton methods can also be employed as alternatives. For estimating the joint attribute distribution, an empirical Bayes method (Carlin & Louis, 2000) is adopted. Specifically, the prior distribution of latent classes is uniform at the beginning, and then updated after each EM iteration based on the posterior distribution, as in, $p(\boldsymbol{\alpha}_c) = \sum_{i=1}^{N} P(\boldsymbol{\alpha}_c|\boldsymbol{X}_i)/N$.

Please note that the above MMLE/EM algorithm is suitable to both RS-GDINA model and US-GDINA model. However, for the US-GDINA model using the G-DINA model as the processing function, item parameter estimates in M-step can be obtained via closed-form solutions; therefore, the general optimization routine is not necessary. After a few algebraic manipulations and simplifications, we have

$$\hat{P}(X_j = h|\boldsymbol{\alpha}_{lj}^*) = \frac{\sum_{i=1}^{N} I(X_{ij} = h)P(\boldsymbol{\alpha}_{lj}^*|\boldsymbol{X}_i)}{\sum_{i=1}^{N} P(\boldsymbol{\alpha}_{lj}^*|\boldsymbol{X}_i)}.$$

By substituting $\hat{P}(X_j = h|\boldsymbol{\alpha}_{lj}^*)$ into equation 4.2, the marginal likelihood estimates of $S_j(h|\boldsymbol{\alpha}_{lj}^*)$ can be obtained. Then, item parameter ϕ can be estimated via the least-square method as introduced by de la Torre (2011). After estimating item parameters, expected a posteriori (EAP) can be used to estimate individuals' attribute patterns.

3.3. Relations with existing polytomous CDMs

Although the US-GDINA model is originally developed for ordered responses with unknown category and attribute association, it has been found to be suitable for nominal response data as well. In particular, the US-GDINA model can be shown to be equivalent to the NRDM (Templin et al., 2008) and the PC-DINA model (de la Torre, 2010) when the processing function is the G-DINA and DINA model, respectively. The equivalence becomes evident when we view all of them as the CDM counterparts of Bock's (1972) nominal response model involving direct estimation of category response functions. For instance, the category response function of the NRDM can be reparameterized using the identity link as

$$P(X_j = h|\boldsymbol{\alpha}_{lj}^*) = \delta_{jh0} + \sum_{k=1}^{K_j^*} \delta_{jhk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*}\sum_{k=1}^{K_j^*-1} \delta_{jhkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{jh12\cdots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{lk}, \tag{2.3}$$

with the constraint of $\sum_{h=0}^{H_j} P(X_j = h|\boldsymbol{\alpha}_{lj}^*) = 1$. It can be shown that estimating $\boldsymbol{\delta} = \{\delta_{jh0}, \delta_{jh1}, \cdots, \delta_{jh12\cdots K_{jh}^*}\}$ is equivalent to estimating the category response function $P(X_{ij} = h|\boldsymbol{\alpha}_{lj}^*)$ because they can be derived directly from each other. Bearing this in mind, the category response function of the sequential G-DINA model in equation 4.2 can then act as a bridge between these two models. Please refer to the online appendix for more details.

In spite of the equivalence, the importance of the US-GDINA model should not be overlooked. Both RS-GDINA and US-GDINA models are special cases of the sequential G-DINA model, with the only difference in how the Q_C-matrix is constructed to reflect our knowledge, or lack thereof, of the category and attribute association. The development of the US-GDINA model allows us to see that the NRDM and PC-DINA model are special cases of the sequential G-DINA model (i.e., when responses are treated as nominal data). As a result, the proposed sequential G-DINA model can serve as a very general model framework so that researchers are able to calibrate simultaneously a number of different CDMs for dichotomous, ordered, or unordered polytomous responses with or without specific assumptions about the cognitive processes (e.g., conjunctive, disjunctive or additive) for a single assessment. In addition, when fitted to ordered responses where categories are attained sequentially, the processing functions from the US-GDINA model can provide extra information that the NRDM and PC-DINA model typically do not provide. Lastly, the MMLE/EM algorithm developed for the sequential G-DINA model is another contribution of this work in that it provides a much faster alternative to the MCMC algorithm originally used for NRDM (Templin et al., 2008).

2.4 Simulation Study

Two simulation studies were conducted to evaluate the performance of the sequential G-DINA model under various conditions. The processing functions used in the simulation studies were the G-DINA model, unless otherwise stated. Study 1 examined (1) whether parameters of the sequential G-DINA model can be recovered accurately based on the proposed estimation algorithm; (2) whether the sequential G-DINA model can provide more accurate person classifications than the G-DINA model using dichotomized responses; and (3) whether the attribute and category association can be used to improve parameter recovery for the sequential G-DINA model.

The appropriateness of the RS-GDINA model depends upon whether the observed processing functions are in accordance with that predicted by the attribute and category association. If the predicted processing functions deviate from the observed ones dramatically, the US-GDINA model may be more appropriate because it relaxes the assumption about the attribute and category association. Study 2 investigated the impact of the discrepancy between the observed and predicted processing functions on parameter estimation. Likelihood ratio test (LRT), Akaike information criterion (AIC; Akaike, 1974), and Bayesian information criterion (BIC; Schwarz, 1978) have been widely used for model comparison within the CDM context (e.g., DeCarlo, 2011; Kunina-Habenicht, Rupp, & Wilhelm, 2012; de la Torre & Lee, 2013). Study 2 also examined whether these indices can be used to select the appropriate sequential G-DINA model under various degrees of discrepancy.

Sixteen polytomous items and five dichotomous items were used for the data simulation. Five attributes were measured by these items. The restricted Q_C-matrix is given in Table 4.1, where all attributes are ensured to be measured the same number of times.

Table 2.3: Restricted Q_C-matrix for data simulation

Item	Category	A1	A2	A3	A4	A5	Item	Category	A1	A2	A3	A4	A5
1	1	1	0	0	0	0	11	1	1	1	0	0	0
1	2	0	1	0	0	0	11	2	0	0	0	0	1
2	1	0	0	1	0	0	12	1	1	1	1	0	0
2	2	0	0	0	1	0	12	2	0	0	0	1	1
3	1	0	0	0	0	1	13	1	1	1	0	0	0
3	2	1	0	0	0	0	13	2	0	0	1	1	1
4	1	0	0	0	0	1	14	1	1	0	1	0	0
4	2	0	0	0	1	0	14	2	0	0	0	1	0
5	1	0	0	1	0	0	14	3	0	0	0	0	1
5	2	0	1	0	0	0	15	1	0	0	0	0	1
6	1	1	0	0	0	0	15	2	0	0	1	1	0
6	2	0	1	1	0	0	15	3	0	1	0	0	0
7	1	0	0	1	0	0	16	1	1	0	0	0	0
7	2	0	0	0	1	1	16	2	0	1	0	0	0
8	1	0	0	0	0	1	16	3	0	0	1	1	0
8	2	1	1	0	0	0	17	1	1	0	0	0	0
9	1	0	0	0	1	1	18	1	0	1	0	0	0
9	2	0	0	1	0	0	19	1	0	0	1	0	0
10	1	0	1	0	1	0	20	1	0	0	0	1	0
10	2	1	0	0	0	0	21	1	0	0	0	0	1

2.4.1 Study 1

2.4.1.1 Design

Sample size and item quality were controlled in this study. Examinees of size $N = 500$, 1000, 2000 or 4000 were drawn from a uniform attribute distribution. Item responses were simulated based on the RS-GDINA model using the restricted Q_C-matrix in Table 4.1. Item j was of high quality when $S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{1}) = 0.9$ and $S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{0}) = 0.1$, moderate quality when $S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{1}) = 0.8$ and $S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{0}) = 0.2$, and low quality when $S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{1}) = 0.7$ and $S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{0}) = 0.3$, for all categories. When $K^*_{jh} > 1$, the processing functions for latent classes with $\boldsymbol{\alpha}^*_{ljh}$ not equal to $\mathbf{0}$ or $\mathbf{1}$, that is, $S_j(h|\boldsymbol{\alpha}^*_{ljh} \not\subset \{\mathbf{0},\mathbf{1}\})$, were drawn from a uniform distribution $U[S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{0}), S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{1})]$. The processing functions were simulated with the monotonicity constraint that examinees mastering additional attributes would not have a lower processing function. Note that, to easily control item quality, the processing functions are manipulated directly instead of ϕ because either of them can be viewed as item parameters when the G-DINA model is used as the processing function. Based on simulated processing functions, category response functions of item j can be calculated, as in, $\boldsymbol{P}_{lj} = \left[P(X_j = 0|\boldsymbol{\alpha}^*_{lj}), \ldots, P(X_j = H_j|\boldsymbol{\alpha}^*_{lj})\right]$. Responses of examinees with attribute pattern $\boldsymbol{\alpha}^*_{lj}$ were generated from a Bernoulli and generalized Bernoulli distribution with parameters of $\boldsymbol{P}_{lj}$, if item j is scored dichotomously and polytomously, respectively. To reduce Monte Carlo sampling errors, 100 data sets were generated in each condition.

Both the US-GDINA model and the RS-GDINA model were fitted to simulated data. To fit the US-GDINA model, the unrestricted Q_C-matrix needs to be constructed from Table 4.1. Specifically, for each item, attributes required by a category are also assumed to be required by all other categories. Take item 15 as an example, in the unrestricted Q_C-matrix, all three categories require the last four attributes. To fit the G-DINA model, polytomous responses were dichotomized in two ways. For one, partial credit and full marks were converted to 1; for the other, only full marks were converted to 1, and partial credit was transformed to 0. In either case, the q-vector of an item in the traditional Q-matrix is specified to measure all required attributes for each category of this item. For example, the q-vector for item 15 is [0 1 1 1 1]. The code for implementing the MMLE/EM algorithm presented in the previous section was written in R language (R Core Team, 2015), and can be requested from the first author.

Item parameter recovery was examined using the root mean square error (RMSE) of the estimated category response function for each latent class from the true, that is,

$$RMSE = \sqrt{\frac{\sum_{r=1}^{R}\sum_{c=1}^{2^K}\sum_{j=1}^{J}\left[\hat{P}^{(r)}(X_j = h|\boldsymbol{\alpha}_c) - P^{(r)}(X_j = h|\boldsymbol{\alpha}_c)\right]^2}{J \times 2^K \times R}},$$

where J, K and R are the number of items, attributes and replications, respectively, and $\hat{P}^{(r)}(X_j = h|\boldsymbol{\alpha}_c)$ and $P^{(r)}(X_j = h|\boldsymbol{\alpha}_c)$ are the estimated and true probability of scoring in category h of item j for examinees with attribute pattern $\boldsymbol{\alpha}_c$ for the rth replication, respectively. Note that RMSE was only calculated for the sequential G-DINA model.

Person parameter recovery was evaluated using the proportion of correctly classified attribute vectors (PCV) defined as,

$$PCV = \frac{\sum_{r=1}^{R}\sum_{i=1}^{N} I^{(r)}[\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i]}{N \times R},$$

where $I^{(r)}[\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i]$ is an indicator variable evaluating whether the estimated attribute vector matches the true for the rth replication.

2.4.1.2 Results

Figure 2.1 gives RMSEs of the RS-GDINA model and the US-GDINA model under various conditions. It is worth emphasizing that the data were generated using the RS-GDINA model; therefore, to evaluate item parameter recovery, we only focused on the RMSEs of the RS-GDINA model. The RMSEs of the RS-GDINA model were between 0.012 and 0.090, with the largest value occurring when $N = 500$ and item quality was low. Although, as expected, sample size and item quality have an impact on parameter estimation, the fact that the maximum RMSE was less than 0.1 shows that item parameters can be recovered accurately based on the proposed estimation algorithm.

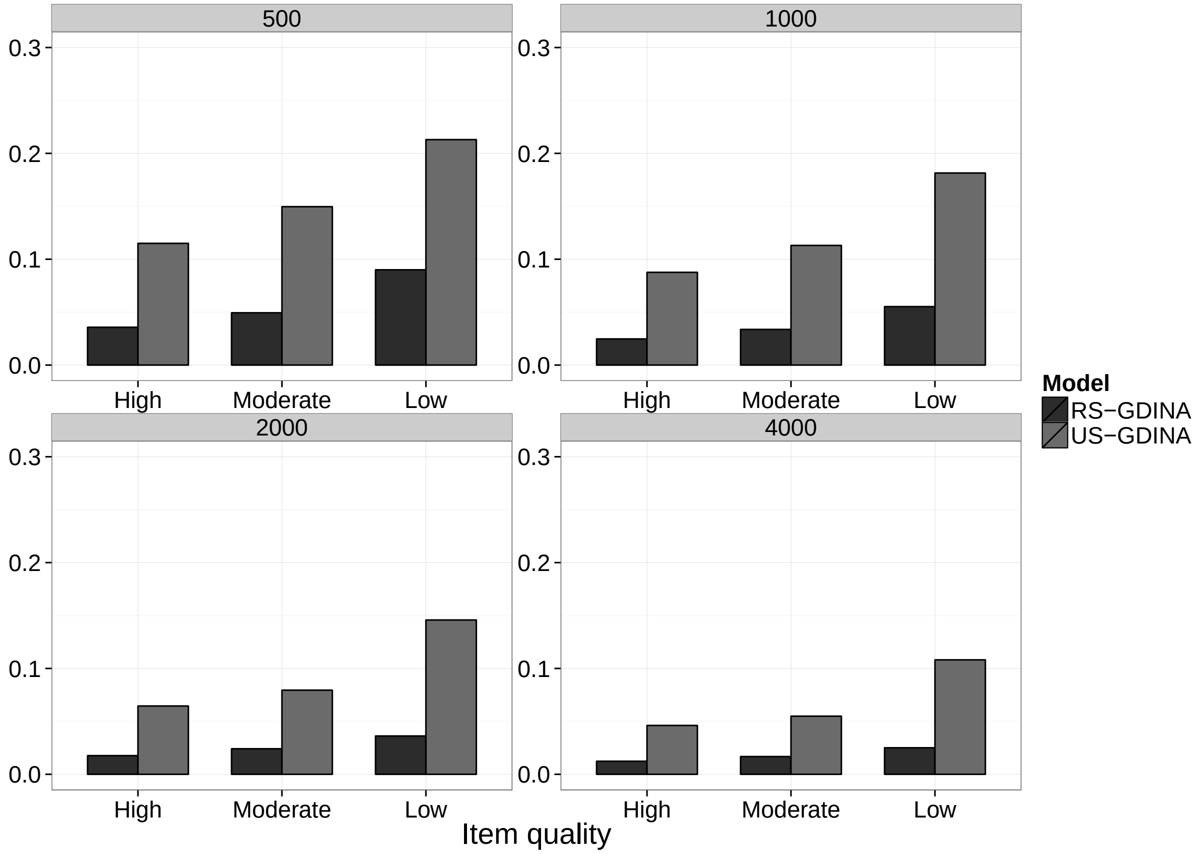


Figure 2.1: RMSE of the sequential G-DINA models

In addition, the RS-GDINA model always had smaller RMSEs than the US-GDINA model, as shown in Figure 2.1. The difference in RMSE between these two models can be larger than 0.1, when item quality was low and sample size was relatively small. For example, the RMSE of the RS-GDINA model was lower than that of the US-GDINA model by 0.126, when $N = 1000$ and item quality was low. This suggests that the attribute and category association can provide important information for accurate item parameter estimation for the sequential G-DINA model, especially under the condition of low item quality and small sample sizes.

Table 2.4 gives PCV for the sequential G-DINA model and the G-DINA model. The results showed that the way of dichotomization influenced the classification rates

Table 2.4: PCV for the sequential G-DINA models and the G-DINA model

N	Item Quality	RS-GDINA	US-GDINA	GDINA1	GDINA2
	High	0.917	0.901	0.759	0.742
500	Moderate	0.679	0.601	0.442	0.371
	Low	0.310	0.223	0.182	0.142
	High	0.921	0.911	0.790	0.752
1000	Moderate	0.692	0.641	0.476	0.415
	Low	0.354	0.254	0.199	0.147
	High	0.923	0.918	0.799	0.761
2000	Moderate	0.697	0.674	0.501	0.441
	Low	0.372	0.295	0.217	0.152
	High	0.924	0.921	0.809	0.765
4000	Moderate	0.702	0.690	0.534	0.456
	Low	0.384	0.339	0.240	0.168

Note: RS-GDINA: the restricted sequential G-DINA model; US-GDINA: the unrestricted sequential G-DINA model; GDINA1: the G-DINA model (examinees get one point as long as they get partial credit); GDINA2: the G-DINA model (examinees get one point only if they get full marks).

of the G-DINA model. Converting both partial credit and full marks to one point (denoted by GDINA1) produced better person classification rates than converting only full marks to one point (denoted by GDINA2) up to 7.8%, with the maximum difference occurring when $N = 4000$ and item quality was moderate. However, this conclusion may not hold when other dichotomous CDMs rather than the G-DINA model are employed, and therefore, needs further investigation.

The sequential G-DINA model produced better person classifications than the G-DINA model fitted to dichotomized responses across all conditions. This conclusion holds regardless of the way of dichotomization. For example, when item quality was moderate and $N = 500$, the RS-GDINA model outperformed the GDINA1 by 23.7% in terms of the PCV. Additionally, the RS-GDINA model produced better person classifications than the US-GDINA model across all conditions. The difference in PCV between these two models can be noticeable when item quality was low and sample size was small. For example, with items of low quality, the RS-GDINA model outperforms the US-GDINA model by 8.7% and 10.0% when sample sizes were 500 and

1000, respectively. When item quality was high or sample size was large, nevertheless, the difference tended to be negligible. These results imply that when enough information has been provided through large sample size and high-quality items, the category and attribute association can offer limited extra information to improve the classification; whereas when sample size is small and item quality is low, the category and attribute association can be very important for accurate attribute estimation.

2.4.2 Study 2

2.4.2.1 Design

This study explored whether LRT, AIC and BIC can be used to choose between the RS-GDINA model and the US-GDINA model. The factors examined in this study included sample size, item quality, fitted models, and magnitude of disturbances. The settings of sample size and item quality were the same as the previous study. To quantify the uncertainty in attribute and category association, small or large disturbances were added to the simulated processing functions. Specifically, the processing functions of each item were first simulated based on the RS-GDINA model using the restricted $\mathrm{Q_C}$-matrix in Table 4.1. Take the first item as an example. When item quality is moderate, the simulated processing functions of the second category are $S(h=2|\boldsymbol{\alpha}_{ljh}^{*}=0)=0.2$ and $S(h=2|\boldsymbol{\alpha}_{ljh}^{*}=1)=0.8$, or equivalently, $S(h=2|\boldsymbol{\alpha}_{lj}^{*}=00)=S(h=2|\boldsymbol{\alpha}_{lj}^{*}=10)=0.2$ and $S(h=2|\boldsymbol{\alpha}_{lj}^{*}=01)=S(h=2|\boldsymbol{\alpha}_{lj}^{*}=11)=0.8$. Then, random disturbances $\boldsymbol{\epsilon}$ were added to the simulated processing functions of $\boldsymbol{\alpha}_{lj}^{*}$. $\varepsilon \sim U[-0.1,0.1]$ indicates a small disturbance and $\varepsilon \sim U[-0.2,0.2]$ a large disturbance, where U represents the uniform distribution. Large disturbance implies a large discrepancy between the data and the RS-GDINA model. If the processing function is greater than 1 or less than 0 after adding the disturbance, it is set to be 0.99 and 0.01, respectively. Study 1 can be viewed as a condition where $\varepsilon = 0$. In each condition, 100 data sets were generated and fitted by both the RS-GDINA model and the US-GDINA model. Please note that adding random disturbances in simulating data based on the RS-GDINA model is equivalent to simulating data from the US-GDINA model. However, the size of disturbances can help quantify how "wrong" the pre-specified attribute and category association is.

AIC and BIC were used for model comparison. LRT was also employed as the RS-GDINA model is nested within the US-GDINA model. Specifically, $\Delta\chi^2$ can be calculated as the difference in -2 log likelihood of two models. There are $\sum_{j=1}^{J}\sum_{h=1}^{H_j} 2^{K_{jh}^*}$ item parameters and $2^K - 1$ latent class parameters, that is, 145 and 449 parameters for the RS-GDINA and US-GDINA models, respectively, based on the $\mathrm{Q_C}$-matrix in Table 3. Accordingly, $\Delta\chi^2$ follows a χ^2 distribution with 304 degrees of freedom. LRTs were conducted at the significant level of 0.05. To understand the properties of LRT, AIC and BIC, the proportion of choosing the US-GDINA model by each statistic was examined. The PCV based on the models selected by LRT, AIC and BIC was calculated as well.

2.4.2.2 Results

Table 2.5 gives the proportion of choosing the US-GDINA model for LRT, AIC and BIC. The results for $\varepsilon = 0$ represented type I errors, which were obtained by reanalyzing the data in Study 1. When $\varepsilon = 0$, AIC and BIC correctly chose the RS-GDINA model for all replications across all conditions; whereas LRT yielded inflated type I errors (ranging from 0.11 to 1) when items quality was moderate or low.

When noises were added (i.e., $\varepsilon = 0.1$ and $\varepsilon = 0.2$), LRT was always able to identify this deviation from the RS-GDINA model and chose the US-GDINA model under all conditions. BIC, however, consistently chose the RS-GDINA model when $\varepsilon = 0.1$, with only one exception occurring when item quality was high and sample size was

Table 2.5: Proportion of choosing the US-GDINA model

		$\varepsilon = 0$			$\varepsilon = 0.1$			$\varepsilon = 0.2$		
N	Item Quality	LRT	AIC	BIC	LRT	AIC	BIC	LRT	AIC	BIC
500	High	0.14	0.00	0.00	1.00	0.01	0.00	1.00	1.00	0.00
	Moderate	0.90	0.00	0.00	1.00	0.00	0.00	1.00	0.99	0.00
	Low	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.31	0.00
1000	High	0.08	0.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
	Moderate	0.57	0.00	0.00	1.00	0.02	0.00	1.00	1.00	0.00
	Low	1.00	0.00	0.00	1.00	0.01	0.00	1.00	0.98	0.00
2000	High	0.02	0.00	0.00	1.00	1.00	0.00	1.00	1.00	0.98
	Moderate	0.24	0.00	0.00	1.00	0.90	0.00	1.00	1.00	0.14
	Low	1.00	0.00	0.00	1.00	0.03	0.00	1.00	1.00	0.00
4000	High	0.06	0.00	0.00	1.00	1.00	0.12	1.00	1.00	1.00
	Moderate	0.11	0.00	0.00	1.00	1.00	0.00	1.00	1.00	1.00
	Low	0.96	0.00	0.00	1.00	0.67	0.00	1.00	1.00	0.00

Note: LRT: likelihood ratio test; AIC: Akaike information criterion (Akaike, 1974); BIC: Bayesian information criterion (Schwarz, 1978).

4000. When $\varepsilon = 0.2$, BIC still tended to select the RS-GDINA model, especially when sample size was relative small or items quality was low.

When $\varepsilon = 0.2$, the proportion of choosing the US-GDINA model for AIC was greater than 98% in all conditions, except when sample size was 500 and item quality was low, where the proportion was 31%. With small disturbances, AIC preferred the US-GDINA model when items were of high quality and samples were of relatively large sizes. For example, when $N = 500$ and 4000, the proportions of choosing US-GDINA model for AIC were less than 1% and greater than 67%, respectively. When $N = 1000$ or 2000, the proportion of choosing US-GDINA model for AIC increased as item quality improved.

Table 2.6 gives the PCV for the RS-GDINA model, the US-GDINA model and selected models using LRT, AIC, and BIC when disturbances were added. The PCV for RS-GDINA model and US-GDINA model when $\varepsilon = 0$ were given in Table 2.4. Considering that both AIC and BIC always chose the RS-GDINA model when $\varepsilon = 0$, the PCV results for $\varepsilon = 0$ are omitted from Table 2.6. When $\varepsilon = 0.1$, the RS-GDINA model produced comparable or even better classification rates than the US-GDINA model, especially when sample sizes were relatively small and item quality was low. With large disturbances, the US-GDINA model outperformed the RS-GDINA model in all conditions, with one exception occurring when $N = 500$ and item quality was low. The difference in PCV between these two models was up to 8.5%, with the maximum value occurring when $N = 4000$ and items were of low quality.

To compare LRT, AIC and BIC, the higher and lower values of PCV of the RS-GDINA model and the US-GDINA model are used as the upper and lower benchmarks, respectively. Across all conditions, selected models based on these three statistics yielded comparable or higher values of PCV than the lower benchmark, which implies that all of them are useful for model selection. The PCV of models selected using LRT can be lower than the upper benchmark up to 10%, with the value of 10% occurring when $\varepsilon = 0$, $N = 1000$, and item quality was low (which is not given in Table 2.6). Note that, in this condition, LRT always chose the US-GDINA model incorrectly for all replications, as shown in Table 2.5. For BIC, the maximum difference in PCV between the selected models and the upper benchmark is 8.5%, which occurred when $\varepsilon = 0.2$, $N = 4000$ and items were of low quality. Lastly, AIC selected the optimal models when $\varepsilon = 0.2$, which yielded the same PCV as the upper benchmark. When $\varepsilon = 0.1$, AIC also produced almost optimal model selections - the maximum difference in PCV between the selected models and the upper benchmark is 0.7%, occurring when $N = 4000$ and item quality is low. These results suggest that compared with LRT and BIC, models selected by AIC yielded desirable person classification rates in terms of PCV under all conditions.

Table 2.6: PCV of the sequential G-DINA models and selected models using LRT, AIC and BIC

		$\varepsilon = 0.1$					$\varepsilon = 0.2$				
N	Item Quality	LRT	AIC	BIC	RS-GDINA	US-GDINA	LRT	AIC	BIC	RS-GDINA	US-GDINA
500	High	0.927	0.926	0.926	0.926	0.927	0.935	0.935	0.910	0.910	0.935
	Moderate	0.644	0.690	0.690	0.690	0.644	0.757	0.757	0.719	0.719	0.757
	Low	0.252	0.324	0.324	0.324	0.252	0.330	0.347	0.347	0.347	0.330
1000	High	0.934	0.934	0.928	0.928	0.934	0.947	0.947	0.921	0.921	0.947
	Moderate	0.676	0.698	0.698	0.698	0.676	0.790	0.790	0.726	0.726	0.790
	Low	0.286	0.359	0.360	0.360	0.286	0.403	0.403	0.387	0.387	0.403
2000	High	0.940	0.940	0.930	0.930	0.940	0.951	0.951	0.950	0.923	0.951
	Moderate	0.709	0.709	0.707	0.707	0.709	0.806	0.806	0.744	0.733	0.806
	Low	0.335	0.383	0.384	0.384	0.335	0.468	0.468	0.412	0.412	0.468
4000	High	0.943	0.943	0.934	0.933	0.943	0.952	0.952	0.952	0.925	0.952
	Moderate	0.720	0.720	0.708	0.708	0.720	0.814	0.814	0.814	0.734	0.814
	Low	0.384	0.389	0.396	0.396	0.384	0.501	0.501	0.416	0.416	0.501

Note: LRT: likelihood ratio test; AIC: Akaike information criterion (Akaike, 1974); BIC: Bayesian information criterion (Schwarz, 1978); RS-GDINA: the restricted sequential G-DINA model; US-GDINA: the unrestricted sequential G-DINA model.

2.5 Real Data Illustration

2.5.1 Data

The data for this illustration were a subset of the data originally used by Lee, Park, and Taylan (2011), and were taken from booklets 4 and 5 of TIMSS 2007 fourth grade mathematics assessment. Responses of 823 students to 12 of 25 items involving eight of the original 15 attributes identified by Lee et al. (2011) were used in the current study. The definitions of the attributes are given in Table 2.7. Two of the five constructed-response items (i.e., Items 3 and 9) were scored polytomously with three ordered response categories (i.e., 0, 1 and 2). Items M031242A and M031242B, referred to as Items 7a and 7b, respectively, are related to a common stimulus as shown in Figure 2.2. The former requires students to complete tables using the information in two posters, whereas the latter requires one of the correct answers: "3 (as long as does not contradict Part A [i.e., Item 7a] including table empty or incomplete)", or "number(s) correct according to a complete but erroneous table in Part A OR indicates no match according to a complete but erroneous table in Part A" (Foy & Olson, 2009, p. 95). The scoring rule for the latter item implies a heavy dependence between the two items, which has also been found by Hansen (2013) when examining testlet effect. Although it is possible for students to solve Item 7b independently, it is more straightforward, and thus more likely for them to obtain the answer directly by reading from the tables completed in Item 7a. A further examination of students' responses showed that only three out of 823 students answered Item 7b correctly, but not Item 7a. This suggests that we can consider the two items as a single polytomous item to handle the testlet effect, and at the same time, to allow for answering Item 7a successfully as a prerequisite to answering Item 7b correctly for most, if not all, students. By removing the responses of three students who answer Item 7b correctly, but not Item 7a, and combining Items 7a and 7b as a single polytomous item, the responses of 820 students to 11 items were analyzed.

Table 2.7: Attribute definitions for TIMSS 2007 data

A1. Representing, comparing, and ordering whole numbers as well as demonstrating knowledge of place value.
A2. Recognize multiples, computing with whole numbers using the four operations, and estimating computations.
A3. Solve problems, including those set in real life contexts (for example, measurement and money problems).
A4. Find the missing number or operation and model simple situations involving unknowns in number sentence or expression.
A5. Describe relationships in patterns and their extensions; generate pairs of whole numbers by a given rule and identify a rule for every relationship given pairs of whole numbers.
A6. Read data from tables, pictographs, bar graphs, and pie charts.
A7. Comparing and understanding how to use information from data.
A8. Understanding different representations and organizing data using tables, pictographs, and bar graphs.

Note: This table is modified from Lee et al. (2011).

The sequential G-DINA model was fitted to all items. Note that for dichotomously scored items, the sequential G-DINA model is equivalent to the G-DINA model. We used a and b to denote category 1 and 2, respectively, for polytomously scored items 3 and 9. For example, 9a and 9b represent the first and second categories of Item 9, respectively. To fit the model, the q-vector for each category of Items 3, 7 and 9 need to be derived from the original item level q-vector developed by Lee et al. (2011). Item 3 requires students to complete a bar graph by drawing two bars based on the information in a table. Students can get a score of one if only one bar is drawn correctly, and two if both bars are drawn correctly (Foy & Olson, 2009, p. 85). Therefore, it is clear that category 3a is a prerequisite to 3b. Because both categories require the same operation, they measure the same set of attributes, as in, A1, A6 and A8. When viewed as an independent item, Item 7b requires three attributes (i.e., A2, A3 and A7; Lee et al., 2011); however, when Items 7a and 7b are assumed to be attained sequentially, only A7 (i.e., comparing and understanding how to use information from data) is necessary

Table 2.8: Q_C-matrix for TIMSS 2007 data

			Attributes							
Item	TIMSS Item No.	Category	A1	A2	A3	A4	A5	A6	A7	A8
1	M041052	1	1	1	0	0	0	0	0	0
2	M041281	1	0	1	1	0	1	0	0	0
3a	M041275	1	1	0	0	0	0	1	0	1
3b	M041275	2	1	0	0	0	0	1	0	1
4	M031303	1	0	1	1	0	0	0	0	0
5	M031309	1	0	1	1	0	0	0	0	0
6	M031245	1	0	1	0	1	0	0	0	0
7a	M031242A	1	0	1	1	0	1	0	0	0
7b	M031242B	2	0	0	0	0	0	0	1	0
8	M031242C	1	0	1	1	0	1	0	1	0
9a	M031247	1	0	1	1	1	0	0	0	0
9b	M031247	2	0	1	1	1	0	0	0	0
10	M031173	1	0	1	1	0	0	0	0	0
11	M031172	1	1	1	0	0	0	1	0	1

Note: Polytomous items are shown in bold; This Q_C-matrix is modified from Lee et al. (2011).

for Item 7b. Finally, Item 9 requires students to solve a problem in real life context. Students get a score of one if they provide a correct problem-solving procedure, but with computational errors; or if they provide correct final solution without their work shown. If both the solution and procedure are correctly presented, students get a score of two (Foy & Olson, 2009, p. 98). Although this scoring rubric implies that category 9a is a prerequisite to 9b, it is not clear that which attributes are involved in each category. In particular, it is difficult to determine that which attributes are used if only the final solution is provided without their work shown, and that which attributes are involved when computational errors occur. Therefore, we used the unrestricted Q_C-matrix for Item 9, and assumed that both categories required A2, A3 and A4. The Q_C-matrix is given in Table 2.8.

Based on the Q_C-matrix, there were $\sum_{j=1}^{11}\sum_{h=1}^{H_j} 2^{K_{jh}^*} = 102$ item parameters and $2^8 - 1 = 255$ latent class parameters. $S(h|\boldsymbol{\alpha}_{ljh}^*)$ were constrained to be equal to or greater than $S(h|\boldsymbol{\alpha}_{l'jh}^*)$ whenever $\boldsymbol{\alpha}_{ljh}^* \succ \boldsymbol{\alpha}_{l'jh}^*$, similar to de la Torre (2011). Also, the lower and upper bounds for processing functions were set to 0.001 and 0.999, respectively. The MMLE/EM algorithm described in the previous section with a convergence criterion of 0.001 was used for this analysis. It should be noted that, due to the relatively small number of examinees and items, large number of attributes, and possible misspecifications in the Q_C-matrix, the results should be interpreted with caution.

2.5.2 Results

Given in Table 2.9 are the estimated processing functions of the 11 items for each of the specific reduced attribute patterns. There are $2^{K^*_{jh}}$ processing functions for category h of item j associated with the reduced attribute patterns given on the top of the table. It should be noted that the same reduced attribute pattern for different items may not represent the same set of attributes. For example, Items 5 and 6 each have four reduced attribute patterns, but they refer to different attributes (i.e., A2 and A3 for Item 5, and A2 and A4 for Item 6).

For Item 7, students who have mastered all the required attributes for category 1 (i.e., A2, A3 and A5) have a 95.5% chance of answering this category correctly; however, those who lack the required attributes have only 8.8% chance of being correct. After completing category 1 correctly, students who have mastered A7 and those who have yet to master the attribute have 99.9% and 0.1% chance of being correct on category 2, respectively. Furthermore, students' responses to category 1 do not appear to satisfy the conjunctive assumption that lacking one of the required attributes produces identical processing functions. For example, the processing function for students who have mastered A5 but not A2 and A3 is approximately twice as high as that for students who have mastered A3 only, both of which are much higher than that for students who have not mastered any required attribute. It can also be noted that the conjunctive assumption may not hold well for most categories requiring two or more attributes. A such fitting the DINA model to the data, as in Lee et al. (2011) and Hansen (2013), might be an oversimplification.

A close scrutiny of the processing functions reveals that for most categories, students who mastered all the required attributes have very high probabilities of success (e.g., the processing functions are greater than 0.95 for 12 out of 14 categories); whereas those who mastered none of the required attributes have low probabilities of success (e.g., the processing functions are less than 0.15 for 9 out of 14 categories). Similar to de la Torre (2008), $S_j(h|\mathbf{1}) - S_j(h|\mathbf{0})$ can be defined as a category discrimination index for category h of item j. For 14 categories of 11 items in this data, the discrimination indices range from 0.488 to 0.998, with the mean of 0.79. This means that overall, most categories can be considered very discriminating.

In addition, compared with fitting NRDM to Item 9, fitting the US-GDINA model provided more information about response categories. For instance, mastering A4 contributed considerably to the processing functions for 9b, but not for 9a, which implies a possible misspecification in the q-vector of 9a because A4 does not seem to be necessary for this category. Similar findings can be observed for Item 1 and 6. For example, A2 has a trivial contribution to the success probability for Item 6, and therefore, may not be necessary. These, however, need to be investigated further.

Lastly, although the processing functions can be interpreted in a straightforward manner as above, the category response functions can also be derived easily and interpreted accordingly. For example, students who mastered all the required attributes for Item 3 have 99.8% chance of getting a score of two; whereas those who only mastered A1 have 72.2% chance of getting a score of one, but only 1% chance of getting a score of two.

Table 2.9: Estimates of processing functions for TIMSS 2007 data analysis.

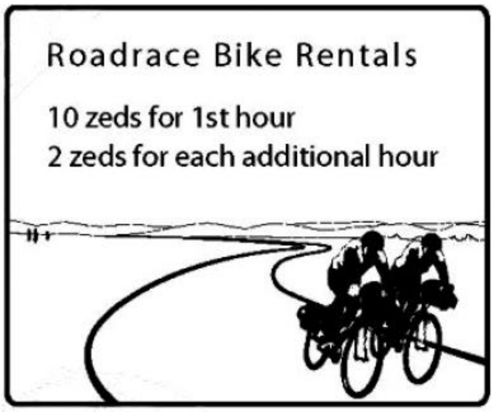
		Attribute Pattern															
		0	1														
		00	10	01	11												
		000	100	010	001	110	101	011	111								
Item	Category	0000	1000	0100	0010	0001	1100	1010	1001	0110	0101	0011	1110	1101	1011	0111	1111
1	1	0.511	0.941	0.511	0.999												
2	1	0.260	0.908	0.444	0.487	0.908	0.908	0.617	0.999								
3a	1	0.002	0.732	0.999	0.999	0.999	0.999	0.999	0.999								
3b	2	0.013	0.013	0.540	0.535	0.999	0.999	0.999	0.999								
4	1	0.452	0.868	0.781	0.973												
5	1	0.122	0.865	0.305	0.961												
6	1	0.001	0.001	0.914	0.999												
7a	1	0.088	0.621	0.427	0.854	0.770	0.899	0.891	0.955								
7b	2	0.001	0.999														
8	1	0.257	0.257	0.377	0.646	0.347	0.377	0.999	0.347	0.988	0.457	0.646	0.999	0.718	0.999	0.999	0.999
9a	1	0.072	0.145	0.452	0.072	0.646	0.145	0.470	0.646								
9b	2	0.001	0.645	0.532	0.532	0.715	0.762	0.532	0.762								
10	1	0.095	0.700	0.761	0.999												
11	1	0.355	0.835	0.835	0.355	0.355	0.835	0.835	0.835	0.835	0.835	0.415	0.835	0.835	0.835	0.999	0.999

Note: Polytomous items are shown in bold.

Posters for two sports clubs that rent bikes are shown below.

Mountain Bike Rentals

8 zeds for 1st hour
3 zeds for each additional hour



Roadrace Bike Rentals

10 zeds for 1st hour
2 zeds for each additional hour

A. Use the information in the posters to complete the tables.

Mountain Bike rental	
Hours	Cost (zeds)
1	8
2	11
3	
4	
5	
6	

Roadrace Bike Rentals	
Hours	Cost (zeds)
1	10
2	12
3	
4	
5	
6	

B. For what number of hours are the rental costs the same at the two clubs?

Answer: ____________________________________

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

Figure 2.2: Item M031242A and M031242B from TIMSS 2007 assessment

2.6 Summary and Discussion

In this paper, we developed a new polytomous CDM for graded responses (i.e., the sequential G-DINA model). Unlike other existing polytomous CDMs, by taking the attribute and category association into account, the sequential G-DINA model is able to model different cognitive processes at different response categories when these categories are completed in a sequential manner. Although initially developed for graded responses, the sequential G-DINA model is also suitable for unordered categorical responses when the unrestricted Q_C-matrix is used. The simulation study shows that the proposed estimation algorithm can produce accurate item and person parameter recovery.

The RS-GDINA model and the US-GDINA model were distinguished in this paper to account for possible uncertainty in attribute and category association. LRT, AIC and BIC were used to compare the RS-GDINA and US-GDINA models empirically. Based on the simulation study, selected models based on AIC can produce almost optimal person classifications in all simulated conditions. LRT and BIC yielded worse results in some conditions.

The development of the sequential G-DINA model has important practical implications in that it opens the possibility of relating response categories to attributes of interest. In particular, when writing polytomous items for cognitive diagnostic assessment, item writers may consider whether it is possible to link categories with attributes. In doing so, more diagnostic information may be extracted, which in turn can lead to more accurate person classifications.

Despite promising results, only the psychometric framework has been developed in this study. This offers researchers and practitioners a flexible tool to analyze polytomous items, but it is only a beginning of exploiting the value of polytomous items in cognitive diagnostic assessment. Additional research along this line is needed. For example, this study only used one Q_C-matrix, and in future studies, researchers can consider various Q_C-matrices to examine the impact of the Q_C-matrix on parameter recoveries. Additionally, like most other CDMs, the sequential G-DINA model is a single-strategy model assuming that all examinees use the same strategy, which, nevertheless, may not necessary be the case in practice. For example, to solve $4\frac{1}{8} - \frac{3}{8}$, another strategy is to convert the mixed number to improper fraction prior to further operations. Multiple-strategies issues have been considered in the context of dichotomous response data. For example, Mislevy (1996) considered a mixture model for estimating the strategy being used. De la Torre and Douglas (2008) developed a multiple-strategy DINA model which allows students to use different strategies for different items. A major difference for graded response data stemming from constructed-response items is that the strategy used by each student for each item is probably observable if students show their work explicitly, and therefore estimating the strategy being used is not needed. It would be straightforward to incorporate multiple Q_C-matrices into the sequential G-DINA model, in conjunction with indicator variables showing the strategies being used by each students for each item.

Although the sequential G-DINA model does not make any assumption about the attribute structures, if attributes are structured, it seems intuitively more reasonable to assess lower level attributes at lower level categories. How the attribute structures can be incorporated in the sequential G-DINA model would be an interesting topic to examine in the future. Also, in the simulation studies, five single-attribute items were included to ensure the Q_C-matrix is complete (Chiu, Douglas, & Li, 2009). However, it is worthwhile to further investigate whether the completeness in the sequential G-DINA model can be achieved using single-attribute specifications at the category rather than item level. Furthermore, although the authors have noted the relationship between the sequential G-DINA model and the NRDM (Templin et al., 2008) and PC-DINA model (de la Torre, 2010), it is still not clear how the proposed model relates to other polytomous CDMs, such as pGDM (von Davier, 2008) and polytomous LCDM (Hansen, 2013). This needs further investigation. Lastly, because of the sequential mechanism underlying the proposed model, it is appropriate for items with analytic scoring rubrics. At this point, it is not clear if the sequential G-DINA model is applicable to items that are scored using holistic rubrics.

2.7 References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats - It does make a difference for diagnostic purposes. *Applied Psychological Measurement*, *11*, 385–395.

Birenbaum, M., Tatsuoka, K. K., & Gutvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement*, *16*, 353–363.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. New York, NY: Chapman & Hall.

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

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8–26.

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. (2010, July). *The partial-credit DINA model.* Paper presented at the international meeting of the Psychometric Society, Athens, GA.

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

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*, 355–373.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). A review of cognitively diagnostic assessment and a summary of psychometric models. In R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics* (Vol. 26, pp. 979–1030). Amsterdam, Netherlands: Elsevier.

Foy, P., & Olson, J. F. (2009). *TIMSS 2007 user guide for the international database*. Chestnut Hill, MA: International Study Center, Boston College.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321.

Hansen, M. (2013). *Hierarchical item response models for cognitive diagnosis* (Unpublished doctoral dissertation). University of California at Los Angeles.

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

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

Johnson, M., Lee, Y.-S., Sachdeva, R. J., Zhang, J., Waldman, M., & Park, J. Y. (2013, April). *Examination of gender differences using the multiple groups DINA model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, California.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59–81.

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*, 144–177.

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

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Education al Measurement*, *33*, 379-416.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The computer journal*, *7*, 308-313.

R Core Team. (2015). *R: A language and environment for statistical computing.* [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from `https://www.R-project.org/`

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.

Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika*, *60*, 549-572.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

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

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

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

Templin, J. L., Henson, R. A., Rupp, A. A., Jang, E., & Ahmed, M. (2008, March). *Cognitive diagnosis models for nominal response data*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.

2.8 Appendix

Appendix 2A: Equivalence between the Sequential G-DINA Model and Other Polytomous CDMs

The connection between the unrestricted sequential G-DINA (US-GDINA) model and other existing polytomous CDMs including the partial credit DINA (PC-DINA; de la Torre, 2010) model and nominal response diagnostic model (NRDM; Templin, Henson, Rupp, Jang, & Ahmed, 2008) are elaborated below.

The US-GDINA model and PC-DINA model

The probability of scoring category h of item j for examinees with reduced attribute pattern $\boldsymbol{\alpha}_{lj}^*$ for the PC-DINA model is given by

$$P(X_j = h|\boldsymbol{\alpha}_{lj}^*) = \begin{cases} g_{jh}, & \text{if } \boldsymbol{\alpha}_{lj}^* \neq \mathbf{1} \\ 1 - s_{jh}, & \text{if } \boldsymbol{\alpha}_{lj}^* = \mathbf{1} \end{cases},$$

where $h = 1, \dots, H_j$ if item response $X_j \in \{0, 1, \dots, H_j\}$. Note that although we still use $h = 1, \dots, H_j$ for convenience, these categories are not necessarily ordered. The PC-DINA model has $2 \times H_j$ parameters for item j, and g_{jh} is the probability of scoring category h of item j for examinees who do not master all the attributes required by this item, and s_{jh} is the probability of failure on category h of item j for examinees who master all the required attributes.

For the US-GDINA model, when the processing function is assumed to be the DINA model, $S_j(h|\boldsymbol{\alpha}_{ljh}^*)$ can be fomulated as

$$S_j(h|\boldsymbol{\alpha}_{lj}^*) = \begin{cases} g_{jh}^*, & \text{if } \boldsymbol{\alpha}_{lj}^* \neq \mathbf{1} \\ 1 - s_{jh}^*, & \text{if } \boldsymbol{\alpha}_{lj}^* = \mathbf{1} \end{cases},$$

because $S_j(h|\boldsymbol{\alpha}^*_{lj}) = S_j(h|\boldsymbol{\alpha}^*_{ljh})$ for $h = 1,\ldots,H_j$, when all H_j categories have the same q-vectors. Note that g^*_{jh} is the probability of answering category h correctly given that category $h-1$ has been answered correctly for examinees who do not master all required attributes, and s^*_{jh} is the probability of answering category h incorrectly given that category $h-1$ has been answered correctly for examinees who master all required attributes. Note that if responses are nominal, g^*_{jh} and s^*_{jh} are not interpretable, but they can still be used for model parameterizations.

There are $2 \times H_j$ item parameters for item j as well. Based on the category response function (see Equation 1 in the paper), g_{jh} and s_{jh} can be calculated directly from g^*_{jh} and s^*_{jh}. For example, assuming item j requires two attributes, and has three categories (i.e, 0, 1 and 2), it is easy to find the following relationship:

$$\begin{aligned}
g_{j1} &= g^*_{j1}(1-g^*_{j2}) \\
g_{j2} &= g^*_{j1}g^*_{j2} \\
s_{j1} &= 1-(1-s^*_{j1})s^*_{j2} \\
s_{j2} &= 1-(1-s^*_{j1})(1-s^*_{j2}).
\end{aligned}$$

The US-GDINA model and NRDM

Denote $\boldsymbol{\alpha}_c = (\alpha_{c1},\ldots,\alpha_{cK})$ as the attribute pattern, where $c = 1,\ldots,2^K$. The probability of scoring category h of item j for examinees with attribute pattern $\boldsymbol{\alpha}_c$ for the NRDM is given by

$$P(X_j = h|\boldsymbol{\alpha}_c) = \frac{\exp(z_{jch})}{\sum_{x=0}^{H_j}\exp(z_{jcx})},$$

where $z_{jch} = \lambda_{0jh} + \sum_{k=1}^{K}\lambda_{1jkh}\alpha_{ck}q_{jk} + \sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}\lambda_{2jkk'h}\alpha_{ck}q_{jk}\alpha_{ck'}q_{jk'} + \ldots$. Note that z_{jch} includes the intercept, K main effects, and all possible interactions, and q_{jk} is the element of row j and column k in the traditional Q-matrix. To ensure the identifiability of the NRDM, Templin et al. (2008) added the following constraints for main effects and all interaction terms:

$$\begin{aligned}
&\sum_{h=0}^{H_j} \lambda_{0jh} = 0 \quad \forall j;\\
&\sum_{h=0}^{H_j} \lambda_{1jkh} = 0 \quad \forall j,k;\\
&\sum_{h=0}^{H_j} \lambda_{2jkk'h} = 0 \quad \forall j,k,k';\\
&\vdots
\end{aligned}$$

Like the sequential G-DINA model, we can assume that the first K_j^* attributes are required for item j, and again, let $\boldsymbol{\alpha}_{lj}^* = (\alpha_{l1}, \dots, \alpha_{lK_j^*})$ be the reduced attribute pattern, where $l = 1, \dots, 2^{K_j^*}$. Then, $P(X_j = h|\boldsymbol{\alpha}_c)$ can be reparameterized using the identity link:

$$P(X_j = h|\boldsymbol{\alpha}_{lj}^*) = \delta_{jh0} + \sum_{k=1}^{K_j^*} \delta_{jhk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jhkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{jh12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}, \tag{2.4}$$

with the constraint of $\sum_{h=0}^{H_j} P(X_j = h|\boldsymbol{\alpha}_{lj}^*) = 1$. Note that only the required attributes are considered in this reparameterization because the attributes that are not required do not contribute to the probability of success. Denote $P(X_j = h|\boldsymbol{\alpha}_{lj}^*)$ as $P_{jh}(\boldsymbol{\alpha}_{lj}^*)$ for simplicity. Equation 2.4 can be expressed equivalently in matrix form as follows,

$$\boldsymbol{P}_{jh} = \boldsymbol{M}_j \boldsymbol{\delta}_{jh},$$

where both $\boldsymbol{P}_{jh} = \{P_{jh}(\boldsymbol{\alpha}_{lj}^*)\}$ and $\boldsymbol{\delta}_{jh} = \{\delta_{jh0}, \delta_{jh1}, \cdots, \delta_{jh12\cdots K_{jh}^*}\}$ are vectors of length $2^{K_j^*}$. $\boldsymbol{M}_j$ is a design matrix of dimension $2^{K_j^*} \times 2^{K_j^*}$ (de la Torre, 2011). For example, when $K_j^* = 3$,

$$\begin{bmatrix} P_{jh}(000) \\ P_{jh}(100) \\ P_{jh}(010) \\ P_{jh}(001) \\ P_{jh}(110) \\ P_{jh}(101) \\ P_{jh}(011) \\ P_{jh}(111) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} \delta_{jh0} \\ \delta_{jh1} \\ \delta_{jh2} \\ \delta_{jh3} \\ \delta_{jh12} \\ \delta_{jh13} \\ \delta_{jh23} \\ \delta_{jh123} \end{bmatrix}.$$

Because $\boldsymbol{M}_j$ is invertible (de la Torre, 2011), $\boldsymbol{\delta}_{jh}$ can be derived directly if $\boldsymbol{P}_{jh}$ are known, and vice versa. In other words, we can view either $\boldsymbol{\delta}_{jh}$ or $\boldsymbol{P}_{jh}$ as item parameters.

Similarly, because $\boldsymbol{S}_j(h|\boldsymbol{\alpha}_{lj}^*) = \boldsymbol{M}_j\boldsymbol{\phi}_{jh}$ for the US-GDINA model (See Equation 2 in the paper), either processing functions $\boldsymbol{S}_j(h|\boldsymbol{\alpha}_{lj}^*)$ or $\boldsymbol{\phi}_{jh}$ can be regarded as item parameters. Note that this is not true if the processing function is not the saturated G-DINA model such as the DINA or DINO model in that the $\boldsymbol{M}_j$ is not a squared matrix any more. As a result, to show the equivalence between NRDM and the US-GDINA model, we only need to show the relation between $\boldsymbol{S}_j(h|\boldsymbol{\alpha}_{lj}^*)$ and $\boldsymbol{P}_{jh}$, which is given as the category response function (i.e., Equation 1 in the paper). In other words, $\boldsymbol{P}_{jh}$ can be calculated from $\boldsymbol{S}_j(h|\boldsymbol{\alpha}_{lj}^*)$ directly based on the category response function in the paper, which implies that when item parameter $\boldsymbol{\phi}_{jh}$ of the US-GDINA model are obtained, the item parameters $\boldsymbol{\delta}_{jh}$ of NRDM can be derived accordingly.

Appendix 2B: RMSE and PCV for the Sequential G-DINA Model Under the Higher-order Attribute Distribution

This appendix, which has not previously been published, gives the simulation results about the performance of the sequential G-DINA model when attributes were generated from a higher-order IRT model. Specifically, attribute patterns were generated from a higher-order distribution, where the probability of mastering attribute *k* for individual *i* were defined as:

$$P(\alpha_k = 1|\theta_i, \delta_k) = \frac{\exp(\theta_i - \delta_k)}{1 + \exp(\theta_i - \delta_k)},$$

where θ_i represents the ability of examinee *i*, and was drawn from the standard normal distribution; and δ_k is the difficulty of attribute *k*, which was randomly drawn from one of the five equal intervals from -1.5 to 1.5. The same settings as the simulation study in Chapter 2 were considered except the attribute distribution. Under each condition, 50 data sets were simulated.

Figure 2.3 gives the RMSE, which represents item parameter recovery, and Figure 2.4 gives the PCV, which indicates the person attribute recovery. Compared with the results under uniform attribute distribution, similar results were observed when attributes were generated from the higher-order attribute distribution.

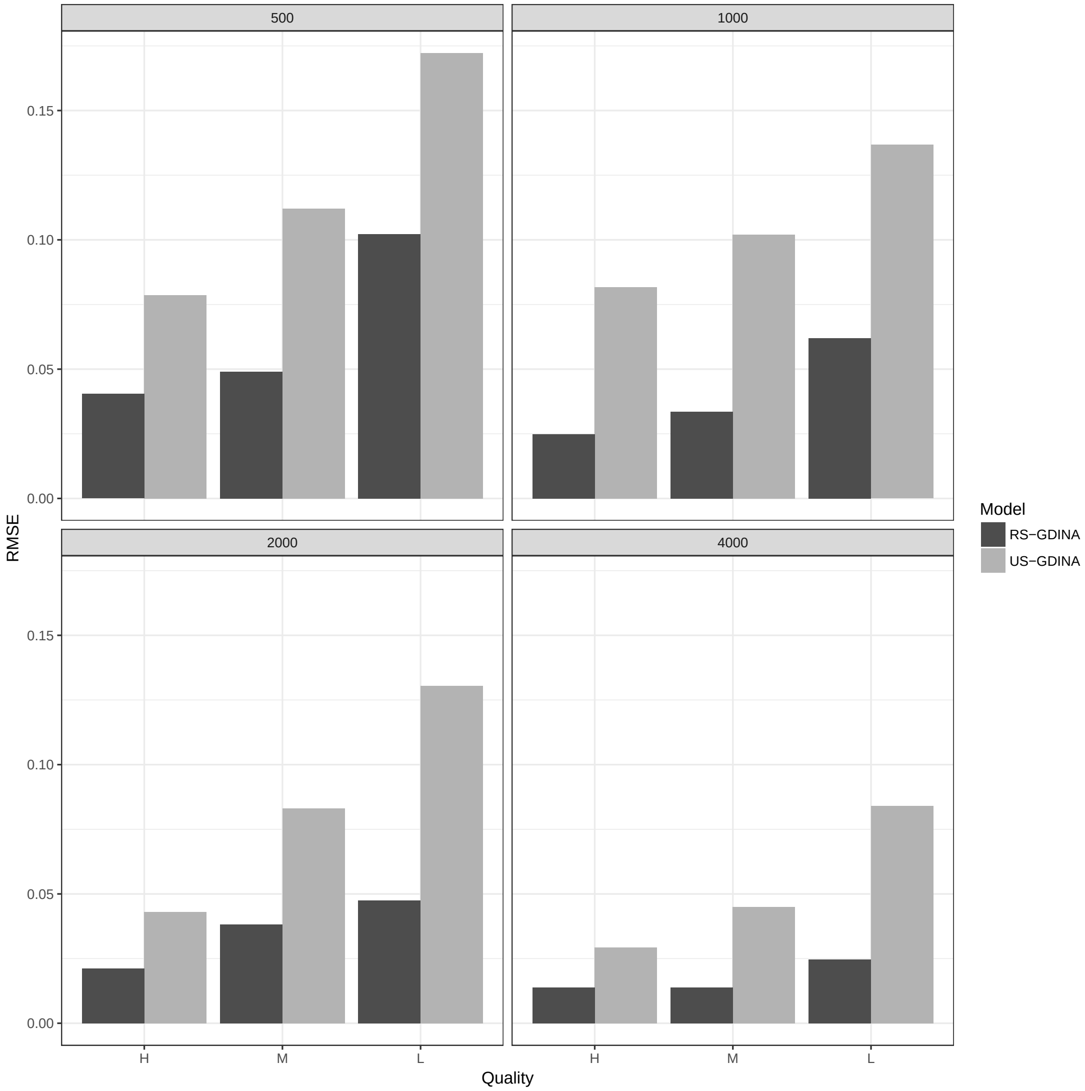


Figure 2.3: RMSE of the sequential G-DINA model under the higher-order attribute distribution

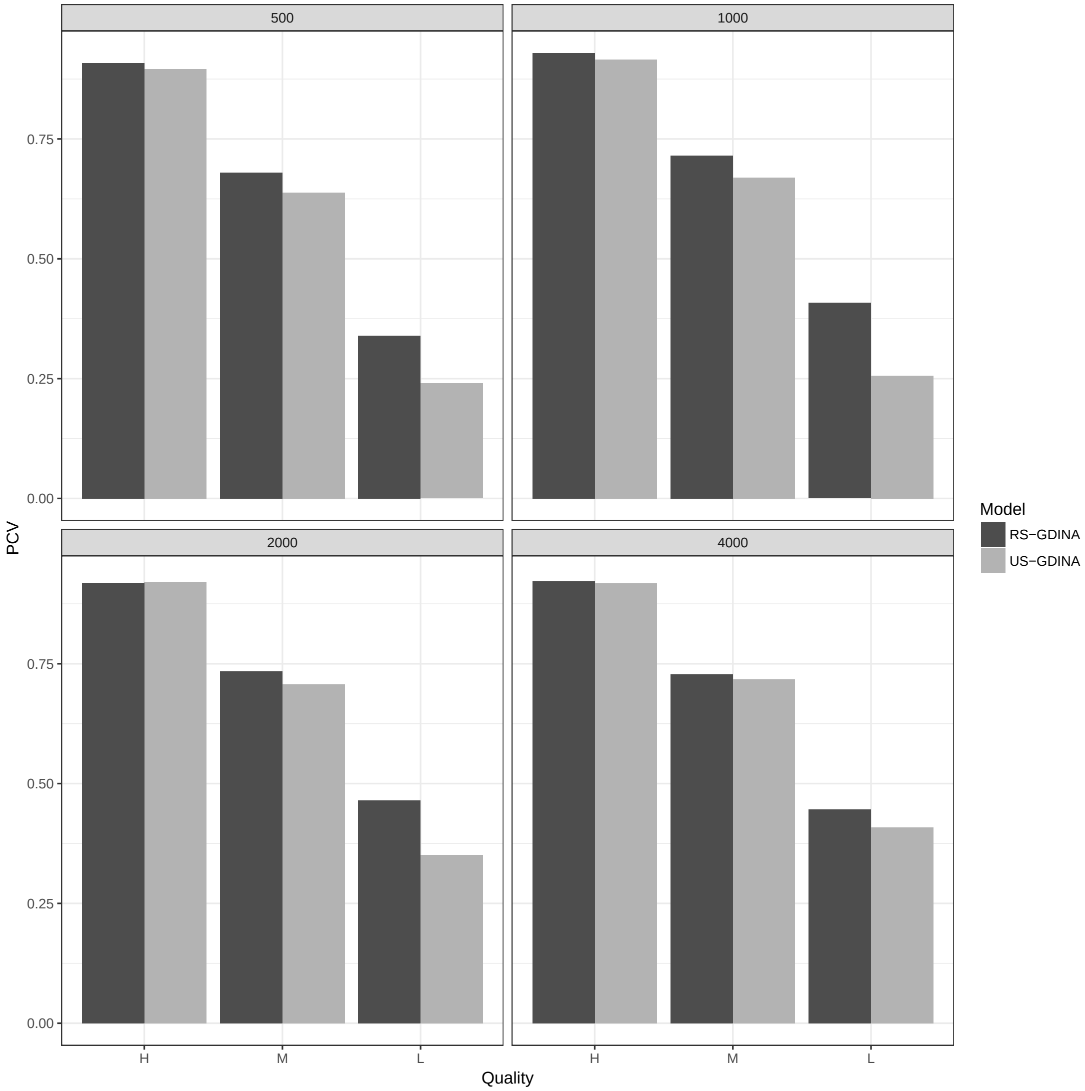


Figure 2.4: PCV of the sequential G-DINA model under the higher-order attribute distribution

Chapter 3

An Empirical Q-Matrix Validation Method for the Sequential G-DINA Model

Abstract

The Q-matrix, which is an item and attribute association matrix, is a core component of most cognitive diagnosis models. The Q-matrix is typically developed by domain experts, and thus tends to be subjective. Validating the Q-matrix empirically is important in that misspecifications in the Q-matrix could produce erroneous attribute estimation. Unlike existing Q-matrix validation procedures that are developed for dichotomous responses, this paper proposes a method to empirically detect and correct the misspecifications in the Q-matrix for graded response data based on the sequential G-DINA model. The proposed Q-matrix validation procedure is implemented in a stepwise manner based on the Wald test and an effect size measure, and its feasibility is examined using simulation studies. A dataset from TIMSS 2007 Mathematics assessment is used to illustrate the proposed method.

3.1 Introduction

Cognitive diagnosis models (CDMs) refer to a set of psychometric models that aim to group individuals into distinct latent classes based on their skill profiles. A more generic term for skills is attributes, which are typically, although not always, assumed as binary latent variables, and an attribute profile indicates which attribute individuals have possessed and which they have not. In educational contexts, CDM analyses could provide diagnostic information about a student's strengths and weaknesses on a set of fine-grained skills to facilitate classroom instruction and learning. CDMs have also shown their potentials in the application in other fields such as psychological disorder diagnosis and personnel selection (Sorrel et al., 2016; de la Torre, van der Ark, & Rossi, 2015; Templin & Henson, 2006).

A host of CDMs have been proposed (for reviews, see DiBello, Roussos, & Stout, 2007; Rupp, Templin, & Henson, 2010), but most are designed for dichotomous responses. Examples include the deterministic inputs, noisy "and" gate (DINA; Haertel, 1989) model, which assumes that examinees are expected to answer an item correctly only when they possess all required attributes, and the generalized DINA (G-DINA; de la Torre, 2011) model, which is a general model subsuming several CDMs. To deal with polytomously scored items appropriately, a few polytomous response CDMs, such as the sequential G-DINA model (W. Ma & de la Torre, 2016) and the general diagnostic model (von Davier, 2008), have been developed as well.

Regardless of their parameterizations, most CDMs rely on a Q-matrix, or an item and attribute association matrix (Tatsuoka, 1983), which specifies whether an attribute is measured by an item. The importance of the Q-matrix in CDM analyses cannot be overemphasized, and it has been widely recognized that a misspecified Q-matrix can degrade item parameter estimation, produce poor model-data fit, and result in erroneous attribute estimation (e.g., Rupp & Templin, 2008; Chiu, 2013). The Q-matrix is typically created by experts (See Tjoe & de la Torre, 2014, for a detailed development process of a Q-matrix for a proportional reasoning test), and assumed to be correct in most CDM analyses. However, expert judgment tends to be subjective and therefore, some entries in the Q-matrix may be not accurate. As noted by DeCarlo (2012), for example, Tatsuoka's (1983) fraction subtraction data has been used for more than twenty years, but its Q-matrix is still in debate: researchers have suggested various modifications to some entries (e.g., de la Torre, 2008; DeCarlo, 2011; de la Torre & Chiu, 2016).

A variety of procedures have been developed to empirically identify and correct misspecified entries in a Q-matrix. For example, Cen, Koedinger, and Junker (2005) considered fitting CDMs with several competing Q-matrices, and comparing the obtained relative model-data fit indices, such as Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978). The Q-matrix that produces the best model-data fit indices is preferred. A major limitation of this approach is that the number of competing Q-matrices may be too large to be manageable.

DeCarlo (2012) developed a Q-matrix validation procedure from a Bayesian perspective for a reparameterized DINA model. The entries in the Q-matrix are treated as random variables, and estimated along with all other parameters using the Markov chain Monte Carlo (MCMC) method. A limitation of this method is that the entries with possible misspecifications need to be known a priori. In another study, Chiu (2013) proposed a nonparametric Q-matrix validation approach by minimizing the residual sum of squares (RSS) between the observed and the corresponding ideal responses. This method was justified by showing that the RSS between the observed and the corresponding ideal responses based on the correct q-vector for an item was less than that based on a misspecified q-vector.

Another Q-matrix validation method was developed by de la Torre (2008), which is called the EM-based delta-method for the DINA model. In his paper, an item discrimination index, δ_j, is defined as the difference in probability of success for item j between examinees who mastered all the required attributes and those who have not. This method is based on δ_j and implemented in a sequential manner. Although promising results can be found based on the simulation study, this method is only applicable for the DINA model. The G-DINA discrimination index ς^2, which has been recently proposed by de la Torre and Chiu (2016), can be viewed as a generalization of δ_j. The ς^2 method is very flexible in that it does not assume the particular CDM forms, as long as they are subsumed by the G-DINA model. Another procedure without the assumption about the forms of the CDMs is developed by Chen (2017) using a set of model fit measures based on residuals. This procedure is carried out iteratively and could be time-consuming.

Despite a number of Q-matrix validation procedures available, none of them is developed for CDMs for graded responses. This could impede the use of constructed-response items in diagnostic assessments because constructed-response items are typically scored polytomously. This study attempts to fill this gap by developing a Q-matrix validation procedure for the sequential G-DINA model (W. Ma & de la Torre, 2016) for graded response data. The sequential G-DINA model is a general CDM suitable for items that need to be solved through a sequence of steps. The proposed Q-matrix validation procedure attempt to determine which attributes are involved for each step of the problem-solving. For dichotomous items, the sequential G-DINA model is equivalent to the G-DINA model and therefore, the proposed methods can also be used for dichotomous responses as long as the underlying CDM is subsumed by the G-DINA model.

3.2 Overview of the Sequential G-DINA Model

Suppose K attributes are involved in a test with J items, and let $X_{ij} \in \{0,\ldots,H_j\}$ be the response of individual i to item j, where H_j is the highest response category for item j. The sequential G-DINA model (W. Ma & de la Torre, 2016) assumes that students' problem solving can be decomposed into a sequence of steps, each involving one or more attributes. A binary q-vector of length K, $\boldsymbol{q}_{jh} = \{q_{jhk}\}$, is associated with category h of item j, and $q_{jhk} = 1$ indicates that the attribute k is required by step h of item j, and $q_{jhk} = 0$ otherwise. A collection of all q-vectors produces a category level Q-matrix, or Q_C-matrix, with the dimensions of $\sum_{j=1}^{J} H_j \times K$. Note that although this paper aims to develop procedures to validate the Q_C-matrix, we still use the general term, Q-matrix validation, to be consistent with the literature.

The K binary attributes lead to 2^K latent classes with unique attribute patterns, namely, $\boldsymbol{\alpha}_c = (\alpha_{c1},\ldots,\alpha_{cK})$, where $c = 1,\ldots,2^K$. Element $\alpha_{ck} = 1$ indicates that attribute k is mastered by individuals in latent class c, and $\alpha_{ck} = 0$ indicates attribute k is not mastered by individuals in the same latent class. The probability for individual i with attribute pattern $\boldsymbol{\alpha}_c$ performing step h correctly provided that s/he has already completed step $h-1$ successfully is referred to as the *processing function* (W. Ma & de la Torre, 2016) and can be expressed as,

$$S_j(h|\boldsymbol{\alpha}_c) = P(X_{ij} \geq h | X_{ij} \geq h-1, \boldsymbol{\alpha}_c) = \frac{P(X_{ij} \geq h|\boldsymbol{\alpha}_c)}{P(X_{ij} \geq h-1|\boldsymbol{\alpha}_c)}.$$

An individual's response falls into category h if she performs the first h steps correctly but fails step $h+1$, and thus the conditional probability of obtaining score h on item j can be written as

$$P(X_{ij} = h|\boldsymbol{\alpha}_c) = [1 - S_j(h+1|\boldsymbol{\alpha}_c)] \prod_{x=0}^{h} S_j(x|\boldsymbol{\alpha}_c),$$

where $S_j(h=0|\boldsymbol{\alpha}_c) \equiv 1$ and $S_j(h=H_j+1|\boldsymbol{\alpha}_c) \equiv 0$.

This model framework is referred to as the sequential process model (W. Ma & de la Torre, 2016); but different names have been used to refer to this type of model in different contexts such as the continuation ratio model (Agresti, 2013; Mellenbergh, 1995), the sequential model (Tutz, 1997) and the step model (Verhelst, Glas, & de Vries, 1997).

The processing function can be defined using any dichotomous CDMs. The sequential G-DINA model is obtained when the G-DINA model (de la Torre, 2011) is used. Specifically, for step h of item j, 2^K latent classes can be collapsed into $2^{K^*_{jh}}$ latent groups, where K^*_{jh} is the number of required attributes for this step. Let $\boldsymbol{\alpha}^*_{ljh}$ be the reduced attribute pattern for step h of item j consisting of the required attributes for this step only, where $l = 1, \cdots, 2^{K^*_{jh}}$. Without loss of generality, we can assume the first K^*_{jh} attributes are required for category h of item j, that is, $\boldsymbol{\alpha}^*_{ljh} = (\alpha_{l1}, \ldots, \alpha_{lk}, \ldots, \alpha_{lK^*_{jh}})$. The processing function for the sequential G-DINA model is given by

$$g\left[S_j(h|\boldsymbol{\alpha}^*_{ljh})\right] = \phi_{jh0} + \sum_{k=1}^{K^*_{jh}} \phi_{jhk}\alpha_{lk} + \sum_{k'=k+1}^{K^*_{jh}} \sum_{k=1}^{K^*_{jh}-1} \phi_{jhkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \phi_{jh12\cdots K^*_{jh}} \prod_{k=1}^{K^*_{jh}} \alpha_{lk},$$

where $g[\cdot]$ is the identity, log or logit link function. ϕ_{jh0} is the intercept, ϕ_{jhk} is the main effect due to attribute k, $\phi_{jhkk'}$ is the two-way interaction effect due to attributes k and k', and $\phi_{jh12\cdots K^*_{jh}}$ is K^*_{jh}-way interaction effect due to all required attributes. Note that the G-DINA model is equivalent to the loglinear CDM (Henson, Templin, & Willse, 2009), both of which subsume a number of reduced models by setting appropriate constraints. This allows different cognitive processes to be modeled at different steps within a single item. For example, if solving a step needs the possession of all required attributes, the DINA model can be used as the processing function, whereas if it needs the mastery of at least one required attribute, the DINO model (Templin & Henson, 2006) can be used as the processing function.

3.3 Category-level GDINA Discrimination Index

The G-DINA discrimination index (GDI) is originally proposed by de la Torre and Chiu (2016) for empirically validating the Q-matrix in conjunction with the G-DINA model for dichotomous responses. It can be extended for the sequential G-DINA model and defined in a straightforward manner for each nonzero category of a polytomously scored item. Specifically, for category h of item j, the GDI can be formulated by

$$\varsigma_{jh}^2 = \sum_{l=1}^{2^{K_{jh}^*}} p(\boldsymbol{\alpha}_{ljh}^*) \left[S_j(h|\boldsymbol{\alpha}_{ljh}^*, \boldsymbol{q}_{jh}) - \bar{S}_{jh} \right]^2,$$

where $p(\boldsymbol{\alpha}_{ljh}^*)$ is the posterior probability of the latent group with the reduced attribute pattern $\boldsymbol{\alpha}_{ljh}^*$, and

$$\bar{S}_{jh} = \sum_{l=1}^{2^{K_{jh}^*}} p(\boldsymbol{\alpha}_{ljh}^*) S_j(h|\boldsymbol{\alpha}_{ljh}^*, \boldsymbol{q}_{jh}).$$

The category level GDI is the variance of success probabilities for category h of item j for all latent groups given $\boldsymbol{q}_{jh}$, and measures a category's overall discriminating power. De la Torre and Chiu (2016) have shown that when the correct Q-matrix is used, the correct q-vector and overspecified q-vectors from the correct one produce the largest GDI, and hence, the q-vector with the largest GDI, but requiring fewest attributes is the correct q-vector. In practice, however, overspecified q-vectors from the correct one have larger GDI than the correct q-vector due to random errors. De la Torre and Chiu (2016) calculated the proportion of variance accounted for (PVAF) by a particular q-vector relative to the maximum for each item, and the q-vector with a PVAF greater than a certain prespecified cutoff, but requiring fewest attributes can be considered correct. This method is very flexible, given that it can be used without any assumption about the form of CDMs. The use of PVAF also provides a way of quantifying the discriminating power of each candidate q-vector. However, this method does not consider the item parameter estimation errors, and determining the cutoff for PVAF a priori could be challenging.

3.4 Attribute Validation Using the Wald Test

Wald test (Wald, 1943) is a widely used hypothesis test in Statistics. In the context of CDMs, it has been used for comparing the G-DINA model and the reduced CDMs that the G-DINA model subsumes (de la Torre, 2011; de la Torre & Lee, 2013; W. Ma, Iaconangelo, & de la Torre, 2016), and detecting differential item functioning (Hou, de la Torre, & Nandakumar, 2014). This section illustrates how the Wald test can be used to evaluate whether or not an attribute that is assumed to be required is statistically necessary in a q-vector involving two or more ones. Specifically, if changing an element one to zero in a q-vector does not lead to a worse model-data fit, the attribute is said to be unnecessary statistically. This allows us to conduct the Q-matrix validation from a perspective of model comparison.

Suppose we want to test whether an element one can be changed to zero in a q-vector $\tilde{\boldsymbol{q}}_{jh} = \{\tilde{q}_{jhk}\}$ for step h of item j, and $\tilde{K}^*_{jh} = \sum_{k=1}^{K} \tilde{q}_{jhk}$. Note that $\tilde{\boldsymbol{q}}_{jh}$ has at least two ones (i.e., $\tilde{K}^*_{jh} \geq 2$), and is not necessarily the same as $\boldsymbol{q}_{jh}$. A $2^{\tilde{K}^*_{jh}-1} \times 2^{\tilde{K}^*_{jh}}$ restriction matrix $\boldsymbol{R}$ is needed for the Wald test so that under the null hypothesis, $\boldsymbol{R} \times \tilde{\boldsymbol{s}}_{jh} = \boldsymbol{0}$, where $\tilde{\boldsymbol{s}}_{jh} = \{\tilde{S}_j(h|\boldsymbol{\alpha}^*_{ljh}, \tilde{\boldsymbol{q}}_{jh})\}$ are the processing functions for category h of item j when $\tilde{\boldsymbol{q}}_{jh}$ is employed. For example, assume $\tilde{K}^*_{jh} = 3$ and $\tilde{\boldsymbol{q}}_{jh} = (1, 1, 1, \ldots)$. To test whether Attribute 1 is required statistically, the null hypothesis is

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \times \begin{bmatrix} \tilde{S}_j(h|000, \tilde{\boldsymbol{q}}_{jh}) \\ \tilde{S}_j(h|100, \tilde{\boldsymbol{q}}_{jh}) \\ \tilde{S}_j(h|010, \tilde{\boldsymbol{q}}_{jh}) \\ \tilde{S}_j(h|001, \tilde{\boldsymbol{q}}_{jh}) \\ \tilde{S}_j(h|110, \tilde{\boldsymbol{q}}_{jh}) \\ \tilde{S}_j(h|101, \tilde{\boldsymbol{q}}_{jh}) \\ \tilde{S}_j(h|011, \tilde{\boldsymbol{q}}_{jh}) \\ \tilde{S}_j(h|111, \tilde{\boldsymbol{q}}_{jh}) \end{bmatrix} = \mathbf{0}.$$

The restriction matrices for testing the necessity of Attribute 2 and 3 are

$$\begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix},$$

and

$$\begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix},$$

respectively. The Wald statistic is defined as

$$W = \left[\boldsymbol{R} \times \tilde{\boldsymbol{s}}_{jh}\right]' \left[\boldsymbol{R} \times \boldsymbol{V}_{jh} \times \boldsymbol{R}'\right]^{-1} \left[\boldsymbol{R} \times \tilde{\boldsymbol{s}}_{jh}\right], \tag{3.1}$$

where $\boldsymbol{V}_{jh}$ is the variance-covariance matrix of $\tilde{\boldsymbol{s}}_{jh}$, which is of dimension $2^{\tilde{K}^*_{jh}} \times 2^{\tilde{K}^*_{jh}}$. The Wald statistic W is asymptotically χ^2 distributed with $2^{\tilde{K}^*_{jh}-1}$ degrees of freedom.

3.5 A Q-matrix Validation Algorithm

The previous section shows how the Wald test can be used to evaluate whether an attribute that is assumed to be necessary is statistically required or not in a q-vector involving two or more ones. This section describes a Q-matrix validation procedure for the sequential G-DINA model using the aforementioned PVAF and Wald statistic. This procedure is implemented category by category and item by item. Specifically, the first required attribute is chosen based on the PVAF, whereas choosing the next required attributes, if any, is based on both the Wald test and the PVAF. The Wald test serves as a hypothesis test, and the PVAF functions as an effect size measure, which can be critical when more than one attribute is deemed necessary based on the Wald test. More specifically, for category h of item j, the algorithm is conducted as follows:

Step 1 : Define $\Omega = \{1, \ldots, K\}$ as a set consisting of the indices for all K attributes. Also, let A be a set consisting of the indices for all the required attributes identified during the validation process, and $B = \Omega \setminus A$. The attributes indexed in set B are called target attributes in that their necessity needs to be examined. Initialize $A = \emptyset$, and thus $B = \{1, \ldots, K\}$. Define a q-vector search bank C consisting of K single-attribute competing q-vectors. Replace the provisional q-vector (i.e., $\boldsymbol{q}_{jh}$ in the $\mathrm{Q_C}$-matrix) with each of the competing q-vectors in C, and calculate their associated PVAFs. The target attribute required by the competing q-vector producing the largest PVAF is defined as a required attribute. Assume this attribute is attribute k', and update set A and B: $A = \{k'\}$ and $B = \Omega \setminus A$.

Step 2 : Check whether the q-vector requiring the attributes indexed in set A has a PVAF greater than 0.95. If yes, the validation process terminates; otherwise, update the search bank C so that each competing q-vector requires all attributes indexed in set A and one target attribute indexed in set B. As a result, there are at least two ones in each competing q-vector in this step. The Wald test is used to examine whether or not the target attribute is statistically necessary for each competing q-vector. If none of the target attributes is required, the validation process terminates; if at least one target attribute is required, the one specified in the competing q-vector with the largest PVAF is assumed to be required, and the associated q-vector is the best among all current competing q-vectors. The index of the target attribute in this q-vector is added to set *A* and removed from set *B*. The necessity of the required attributes except the target one in this competing q-vector is examined using the Wald test as well. If any of them are deemed unnecessary statistically after the target attribute has been included, their indices are removed from set *A* to set *B*. Step 2 is repeated until no new index can be added to or removed from sets *A* and *B*. The flowchart for this validation procedure is given in Figure 3.1.

Step 1 and *Step 2* are implemented for each category of each item. The former aims to determine the first required attribute using the PVAF, and the latter attempts to identify, if any, other required attributes using the Wald test, in conjunction with the PVAF when necessary. After *Step 2* ends, all attributes indexed in set *A* are believed to be required for the studied category. This process is said to be implemented in a stepwise manner in that the necessity of the attributes is evaluated iteratively, similar to the stepwise procedure for model selection in linear regression (Efroymson, 1960). It should be noted that at the beginning of *Step 2*, the PVAF of the current q-vector is calculated and compared with 0.95. This evaluation is not mandatory, but it is useful when sample size is large, in which condition, the hypothesis test tends to reject the null hypothesis and result in over-specified q-vectors.

In addition, the calculations of the GDI and the Wald statistics involve the estimation of the processing functions based on each competing q-vector for the studied category. It is straightforward to recalibrate the data based on each competing q-vector; however, this can be computationally intensive. An alternative solution is the EM-based approximation, similar to de la Torre (2008) and de la Torre and Chiu (2016).

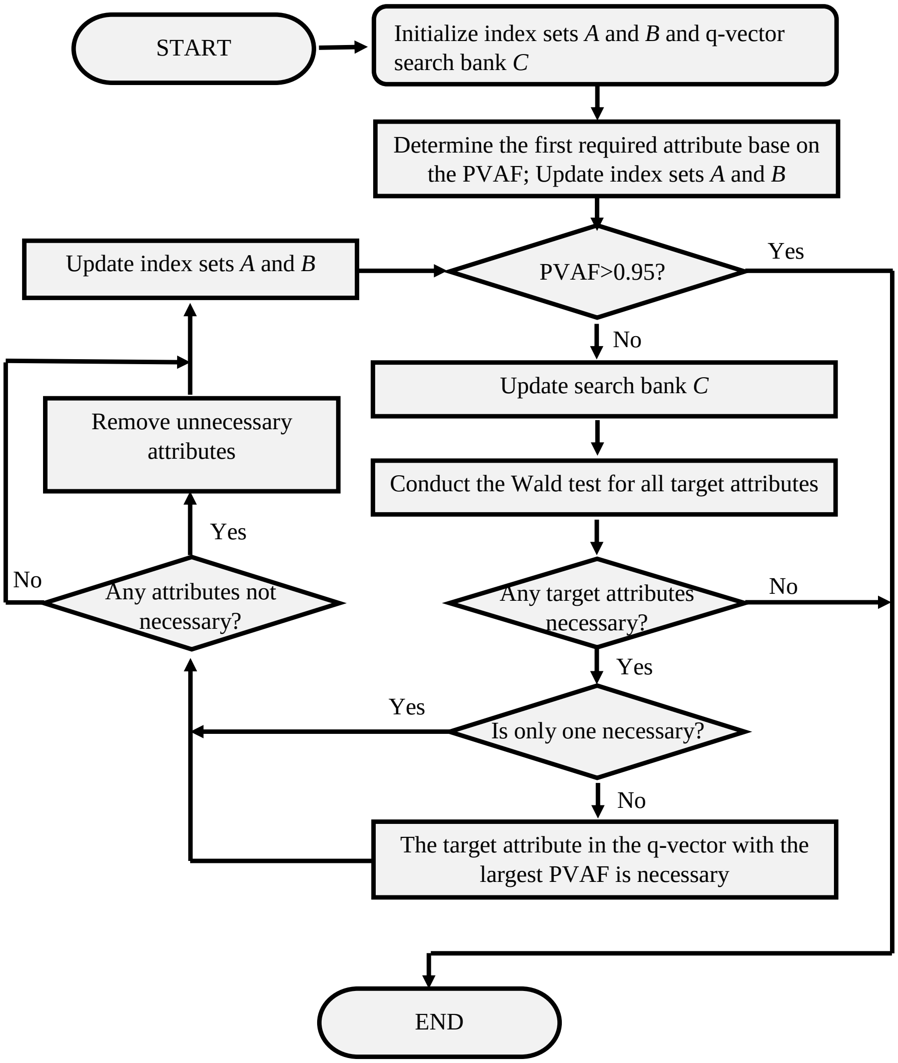

START
Initialize index sets *A* and *B* and q-vector search bank *C*
Determine the first required attribute base on the PVAF; Update index sets *A* and *B*
PVAF>0.95?
Yes
No
Update index sets *A* and *B*
Update search bank *C*
Conduct the Wald test for all target attributes
Remove unnecessary attributes
Yes
No
Any attributes not necessary?
Any target attributes necessary?
No
Yes
Yes
Is only one necessary?
No
The target attribute in the q-vector with the largest PVAF is necessary
END

Figure 3.1: Flowchart of the stepwise Q-matrix validation

The sequential G-DINA model is fitted to the data based on the provisional (i.e., original) Q$_\text{C}$-matrix first. Then, the posterior $P(\boldsymbol{\alpha}_c|\boldsymbol{X}_i)$ can be calculated for each examinee. For a specific competing q-vector for category h of item j, instead of refitting the model, the processing functions are estimated by a one-step EM type of algorithm, as in,

$$\hat{\boldsymbol{S}}_j(h|\boldsymbol{\alpha}^*_{ljh}) = \frac{\tilde{R}^+_h(\boldsymbol{\alpha}^*_{ljh})}{\tilde{R}^+_{h-1}(\boldsymbol{\alpha}^*_{ljh})},$$

where $\tilde{R}^+_h(\boldsymbol{\alpha}^*_{ljh}) = \sum_{i=1}^N \tilde{P}(\boldsymbol{\alpha}^*_{ljh}|\boldsymbol{X}_i)I(X_{ij} \geq h)$, and $\tilde{P}(\boldsymbol{\alpha}^*_{ljh}|\boldsymbol{X}_i)$ is derived from $P(\boldsymbol{\alpha}_c|\boldsymbol{X}_i)$ directly. It should be noted that although the use of $\tilde{P}(\boldsymbol{\alpha}^*_{ljh}|\boldsymbol{X}_i)$ reduces the computational burden dramatically, it is merely an approximation to the posterior when a competing q-vector is adopted. Therefore, the adequacy of this approximation depends on how well the posterior distribution can be estimated based on the original Q$_\text{C}$-matrix.

The details of the Q-matrix validation procedure for category 1 of Item M042303B from the TIMSS 2007 Mathematics assessment is given in Table 3.1 for illustration. This item was analyzed in Section 3.7. The assessment measures 7 attributes, and therefore, at the beginning, $A = \emptyset$ and $B = \{1, 2, \ldots, 7\}$. As shown in Table 3.1, the PVAFs were calculated in Step 1 for the seven candidate single-attribute q-vectors in the initial search bank C. The q-vector measuring α_5 had the highest PVAF, and thus was believed the best single-attribute q-vector. Because the highest PVAF for single-attribute q-vectors was less than 0.95, Step 2 was implemented. In the first round of Step 2, $A = \{5\}$, and the search bank C consists of six candidate q-vectors each measuring two attributes including α_5 and another target attribute. The Wald test was used to evaluate whether the target attributes were statistically necessary. The corresponding p-values were reported as p[entry] in Table 3.1. When α_5 was assumed required, adding any of the target attributes could result in better model-data fit. Because adding α_1 produced the highest PVAF, the q-vector measuring α_1 and α_5 was deemed the best. After the "entry" step, the Wald test was used to evaluate whether α_5 was statistically necessary when α_1 was assumed required in the “removal” step. The associated p-value is denoted as p[removal] in the table. It turned out that α_5 was still statistically necessary when α_1 was assumed required. Because the PVAF was still less than 0.95, Step 2 was repeated after updating sets $A = \{1,5\}$ and $B = \{2,3,4,6,7\}$. In round 2, α_6 was statistically necessary when α_1 and α_5 were assumed required, and the Wald test for α_1 or α_5 was still significant when α_6 was assumed needed. The associated PVAF was less than 0.95 and thus Step 2 was repeated. However, none of the other attributes were significant based on the Wald test. Therefore, the suggested q-vector for this item was 1000110 with a PVAF of 0.922.

3.6 Simulation Studies

Two simulation studies were conducted to evaluate the performance of the proposed stepwise Q-matrix validation method. Simulation Study 1 explored the performance of the proposed method when the processing functions conform to some reduced CDMs, namely, the DINA model, DINO model and A-CDM, whereas the Simulation Study 2 specified the processing function as a more general form (i.e., the G-DINA model). The factors considered in these two studies are summarized in Table 3.2.

3.6.1 Simulation Study 1

3.6.1.1 Design

In this study, the number of items and attributes were fixed to $J = 23$ and $K = 5$, respectively. The sample sizes were N = 1000, 2000 and 4000. Item quality had three levels: $g = s = 0.1$, 0.2 or 0.3 for all categories of all items, representing high, moderate and low quality, where $g = S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{0})$ and $s = 1 - S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{1})$ for category h of item j. When the processing function is the A-CDM, each required attribute contributed equally to the processing function.

Table 3.1: An illustration of the stepwise Q-matrix validation algorithm

Step	Candidate q-vectors in search bank C	PVAF	p[entry]	p[removal]		Decision
				α_5	α_1	
Step 1		$A = \emptyset, B = \{1, \ldots, 7\}$				
	(1000000)	0.260				
	(0100000)	0.123				
	(0010000)	0.333				
	(0001000)	0.065				
	(0000100)	0.369				✓
	(0000010)	0.195				
	(0000001)	0.080				
Step 2						
Round 1		$A = \{5\}, B = \{1,2,3,4,6,7\}$				
	(1000100)	0.818	<.001	<.001		✓
	(0100100)	0.444	<.001			
	(0010100)	0.546	<.001			
	(0001100)	0.424	<.001			
	(0000110)	0.610	<.001			
	(0000101)	0.435	<.001			
Round 2		$A = \{1,5\}, B = \{2,3,4,6,7\}$				
	(1100100)	0.821	0.970			
	(1010100)	0.872	0.070			
	(1001100)	0.822	0.887			
	(1000110)	0.922	0.001	<.001	<.001	✓
	(1000101)	0.822	0.911			
Round 3		$A = \{1,5,6\}, B = \{2,3,4,7\}$				
	(1100110)	0.925	1.000			
	(1010110)	0.924	1.000			
	(1001110)	0.927	0.997			
	(1000111)	0.986	0.375			

Table 3.2: Summary of Factors in simulation studies

Factors	Simulation Study 1	Simulation Study 2
N	1000, 2000, 4000	
(J,K)	(23, 5)	
Misspecification generation	Random	
% of misspecified entries	0%, 10%, 20%	
Attribute distribution	Uniform, Higher-order	
Processing function	DINA/DINO/A-CDM	G-DINA
$\left[S_j(h\|\mathbf{0}),S_j(h\|\mathbf{1})\right]$	$[0.1,0.9]/[0.2,0.8]/[0.3,0.7]$	$[U(0.1,0.3),U(0.7,0.9)]$

Individuals' attribute patterns were generated from two different distributions: the uniform distribution, where all possible attribute patterns are equally likely; and the higher-order distribution (de la Torre & Douglas, 2004), where the probability of mastering attribute k for individual i were defined as:

$$P(\alpha_k = 1|\theta_i, \delta_k) = \frac{\exp(\theta_i - \delta_k)}{1+\exp(\theta_i - \delta_k)},$$

where θ_i represents the ability of examinee i, and was drawn from the standard normal distribution; and δ_k is the difficulty of attribute k, which was randomly drawn from one of the five equal intervals from -1.5 to 1.5.

Table 4.1 gives the $\mathrm{Q_C}$-matrix for the simulation studies. There are five two-category items (i.e., category 0 and 1), 12 three-category items and six four-category items. Misspecified $\mathrm{Q_C}$-matrices were constructed by altering 10% or 20% entries in the correct $\mathrm{Q_C}$-matrix randomly from 0 to 1 or from 1 to 0 with the constraints that each nonzero category measured at least one attribute and that each attribute was required by at least one nonzero category. The processing functions used for data simulation were the DINA model, DINO model, and A-CDM. In each condition, 200 data sets were simulated. The GDINA R package (W. Ma & de la Torre, 2017) was used for data simulation and model estimation, and the stepwise Q-matrix validation was implemented in the R programming environment (R Core Team, 2017).

Table 3.3: $\mathrm{Q_C}$-matrix for simulation studies

Item	Category	α_1	α_2	α_3	α_4	α_5	Item	Category	α_1	α_2	α_3	α_4	α_5
1	1	1	0	0	0	0	13	1	1	0	0	0	0
1	2	0	1	0	0	0	13	2	0	1	0	0	0
2	1	0	0	1	0	0	13	3	0	0	1	0	0
2	2	0	0	0	1	0	14	1	0	0	0	1	0
3	1	0	0	0	0	1	14	2	0	0	1	0	0
3	2	1	0	0	0	0	14	3	0	1	0	0	0
4	1	0	1	0	0	0	15	1	0	0	0	0	1
4	2	0	0	1	0	0	15	2	0	0	1	0	0
5	1	0	0	1	1	0	15	3	1	0	0	1	0
5	2	0	0	0	0	1	16	1	0	0	0	1	0
6	1	1	1	0	0	0	16	2	1	0	0	0	0
6	2	0	0	0	1	0	16	3	0	0	1	0	1
7	1	0	0	1	1	1	17	1	1	0	0	0	0
7	2	0	1	0	0	0	17	2	0	1	0	0	0
8	1	1	1	0	1	0	17	3	0	1	0	1	1
8	2	0	0	0	0	1	18	1	0	1	0	0	0
9	1	0	0	0	0	1	18	2	0	0	0	1	0
9	2	1	0	1	0	0	18	3	1	0	1	0	1
10	1	0	1	0	0	0	19	1	1	0	0	0	0
10	2	1	0	0	0	1	20	1	0	1	0	0	0
11	1	0	0	0	0	1	21	1	0	0	1	0	0
11	2	0	1	1	1	0	22	1	0	0	0	1	0
12	1	0	0	1	0	0	23	1	0	0	0	0	1
12	2	1	0	0	1	1							

An initial recovery rate (IRR) is defined as the percentage of the attributes that are correctly selected after *Step 1* of the Q-matrix validation procedure. An average IRR across all replications for each condition is used to evaluate the performance of the GDI in selecting the first required attribute. To examine the performance of the Q-matrix validation procedure, true positive rate and true negative rate were calculated. The true positive rate is the percentage of misspecified entries that were correctly identified, and the true negative rate is the percentage of correct entries that were correctly retained.

3.6.1.2 Results

The IRRs when items were of low quality are provided in Table 3.4. When item quality were high or moderate, the IRRs were always greater than 99%, which implies that the GDI has excellent performance in selecting the initial required attribute under these conditions and therefore, the results were omitted from the table. When item quality was low, the IRRs were still very good with a minimum value of 95.3%, which occurred under $N = 1000$, 20% misspecifications, higher-order attribute distribution and A-CDM processing function. From Table 3.4, the IRR increased as the sample size increased or the percentage of misspecification decreased. This pattern, however, does not hold under high or moderate item quality probably because of the ceiling effect. Regarding attribute distributions, similar IRRs were observed when there were 10% misspecifications or less; but lower IRRs were observed for the higher-order attribute distribution when there were 20% misspecifications. Generating processing functions do not have much impact on IRRs.

Table 3.4: IRR for reduced models when item quality was low

Processing		Uniform			Higher-Order		
Function	% Misp	N=1000	N=2000	N=4000	N=1000	N=2000	N=4000
DINA	0%	0.997	0.999	1.000	0.999	1.000	1.000
	10%	0.991	0.998	0.999	0.989	0.997	0.999
	20%	0.977	0.994	0.996	0.958	0.972	0.984
DINO	0%	0.998	1.000	1.000	0.998	1.000	1.000
	10%	0.993	0.998	0.999	0.992	0.998	0.999
	20%	0.976	0.989	0.995	0.963	0.982	0.987
ACDM	0%	0.997	1.000	1.000	0.999	1.000	1.000
	10%	0.994	0.999	1.000	0.993	0.998	0.999
	20%	0.980	0.993	0.998	0.953	0.977	0.982

Note: % Misp represents the percentage of misspecifications.

Figure 3.2 gives the true positive rates across sample sizes, item qualities, attribute distributions, percentages of misspecifications, and processing functions. Item quality influences the true positive rates. The average true positive rate was 98.8% with a minimum value of 97.4% when items were of high quality, and 96.2% with a minimum value of 91% when items were of moderate quality. When item quality was low, however, the average true positive rate dropped to 78.9% with a minimum value of 66.4%. The impact of sample sizes, attribute distributions, and percentages of misspecifications was apparent when items were of moderate or low quality. Specifically, the true positive rate increased as the sample size increased, or the percentage of misspecifications decreased. Also, uniformly distributed attributes yielded higher true positive rates than higher-order attributes. When items were of high quality, however, the impact of other factors was not always consistent, which may be caused by ceiling effect in that the range of true positive rate was only 1.9%.

Table 3.5 gives the true negative rates for the stepwise Q-matrix validation method across varied conditions. It can be observed that across all conditions, the validation method performed excellently. Even when item quality was low, the average true negative rate was 97.5% with the minimum values of 95.3%. The average true negative rate for high item quality conditions (i.e., 99.4%) is slightly higher than that for moderate item quality conditions (i.e., 99.3%), but with large sample size and small percentage of misspecification, items of moderate quality could have slightly larger true negative rates than items of high quality. The true negative rates increased as sample size increased. In addition, the uniformly distributed attributes produced higher true negative rates than the higher-order attributes. There was no apparent difference in true negative rate among different percentages of misspecification and different processing functions.

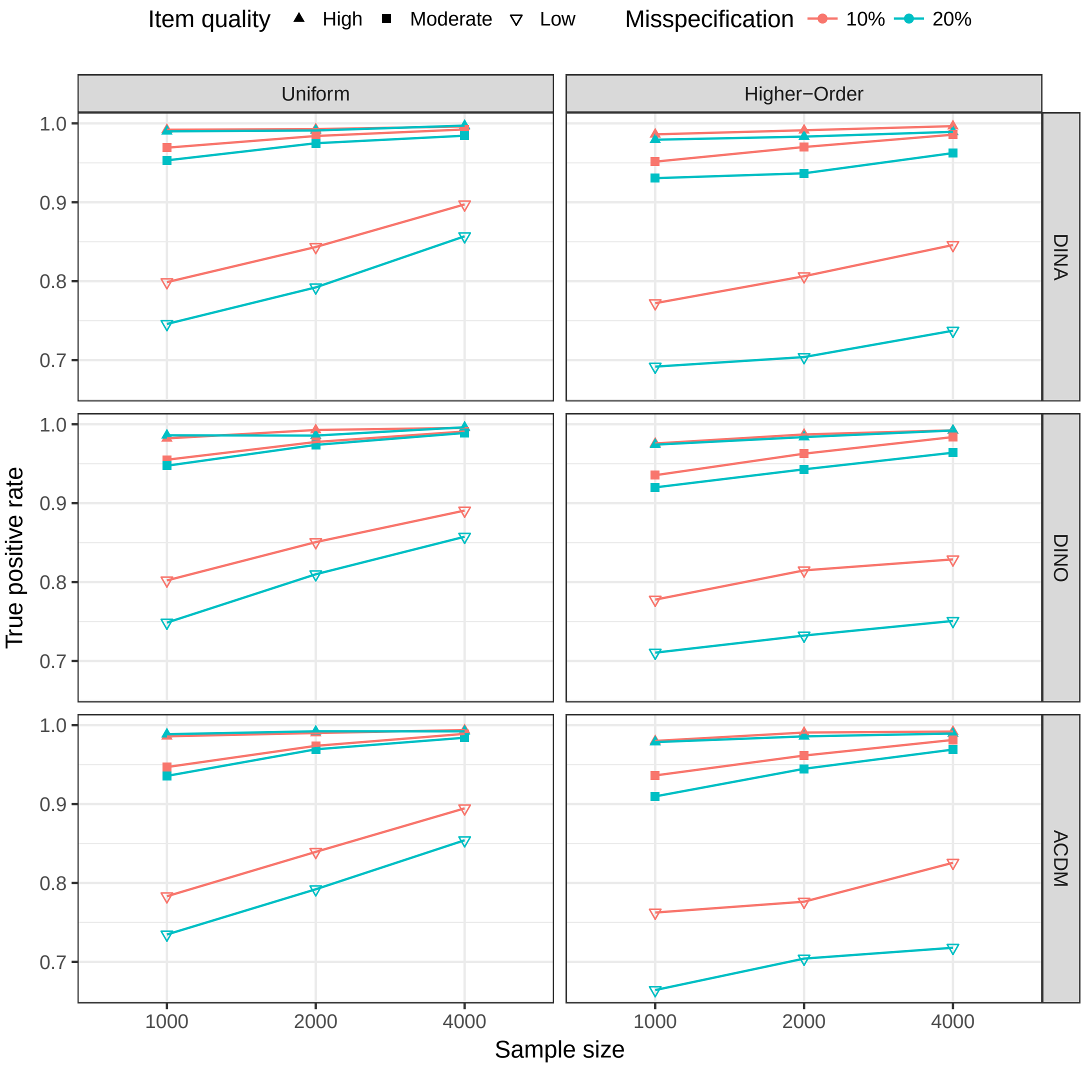


Figure 3.2: True positive rates under reduced processing functions

Table 3.5: True negative for reduced models

			DINA			DINO			ACDM		
Attribute	% Misp	N	High	Moderate	Low	High	Moderate	Low	High	Moderate	Low
Uniform	0%	1000	0.995	0.996	0.963	0.989	0.989	0.962	0.992	0.988	0.956
		2000	0.996	0.999	0.985	0.993	0.995	0.981	0.995	0.998	0.978
		4000	0.998	1.000	0.997	0.998	0.999	0.992	0.997	1.000	0.992
	10%	1000	0.995	0.996	0.962	0.990	0.989	0.962	0.994	0.988	0.958
		2000	0.996	0.999	0.983	0.993	0.994	0.979	0.996	0.997	0.978
		4000	0.999	1.000	0.996	0.998	0.999	0.991	0.998	1.000	0.992
	20%	1000	0.995	0.993	0.959	0.989	0.986	0.955	0.993	0.985	0.956
		2000	0.997	0.997	0.981	0.994	0.993	0.974	0.997	0.996	0.976
		4000	0.999	0.998	0.994	0.999	0.998	0.988	0.998	0.998	0.991
Higher-Order	0%	1000	0.990	0.993	0.963	0.984	0.984	0.959	0.987	0.984	0.954
		2000	0.994	0.998	0.983	0.989	0.993	0.973	0.993	0.995	0.974
		4000	0.997	1.000	0.993	0.995	0.997	0.987	0.996	0.999	0.990
	10%	1000	0.990	0.991	0.962	0.983	0.981	0.957	0.988	0.983	0.956
		2000	0.995	0.997	0.981	0.990	0.992	0.974	0.994	0.995	0.975
		4000	0.998	0.999	0.992	0.996	0.997	0.985	0.997	0.999	0.989
	20%	1000	0.989	0.987	0.959	0.982	0.976	0.954	0.988	0.980	0.953
		2000	0.992	0.992	0.976	0.990	0.988	0.970	0.994	0.993	0.973
		4000	0.995	0.994	0.988	0.994	0.994	0.980	0.996	0.997	0.987

3.6.2 Simulation Study 2

3.6.2.1 Design

In Simulation Study 1, the performance of the stepwise Q-matrix validation method was examined when the processing functions was a reduced CDM. This study used the G-DINA model as the processing function, which relaxes the assumptions about the condensation rule for each category. A more realistic condition for item quality was considered as well, where $S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{0})$ and $S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{1})$ were drawn from $U(0.1, 0.3)$ and $U(0.7, 0.9)$, respectively. When $K^*_{jh} > 1$, the processing functions for latent classes with $\boldsymbol{\alpha}^*_{ljh}$ not equal to $\mathbf{0}$ or $\mathbf{1}$, that is, $S_j(h|\boldsymbol{\alpha}^*_{ljh} \not\subset \{\mathbf{0}, \mathbf{1}\})$, were drawn from the uniform distribution $U[S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{0}), S_j(h|\boldsymbol{\alpha}^*_{ljh} = \mathbf{1})]$. The processing functions were simulated with the monotonic constraint that mastering an additional attribute would not produce a lower processing function. In each condition, 200 data sets were generated. As in the previous study, IRR was used to evaluate the performance of the GDI in identifying the initial required attribute, and the true positive and true negative rates were used to assess the performance of the stepwise validation procedure.

3.6.2.2 Results

Table 3.6 gives the IRRs across sample sizes, attribute distributions and percentages of misspecifications. The GDI has excellent power to identify the first required attribute with a minimum IRR of 98.9%. It can also be observed that the IRR increased as the sample size increased and the percentage of misspecification decreased. Additionally, the IRRs under the uniform attributes were slightly higher than or equal to these under the higher-order attributes with only one exception.

True positive and true negative rates are given in Table 3.7. The stepwise Q-matrix validation procedure performs well in correcting the misspecifications and retaining the correct q-entries in the Q-matrix. Across all conditions, the true positive and true

Table 3.6: IRR for the G-DINA processing function

	Uniform			Higher-Order		
% Misp	N=1000	N=2000	N=4000	N=1000	N=2000	N=4000
0%	0.999	1.000	1.000	0.998	1.000	1.000
10%	0.999	0.999	1.000	0.998	1.000	1.000
20%	0.993	0.996	0.996	0.989	0.991	0.998

negative rates were greater than 91% and 96%, respectively. In addition, both true positive and true negative rates increased as sample sizes increased or the percentage of misspecifications decreased. Compared with the higher-order attribute distribution, the true positive and true negative rates were higher under the uniform distribution.

Table 3.7: Recovery rates for the G-DINA processing function

		Uniform			Higher-Order		
	% Misp	N=1000	N=2000	N=4000	N=1000	N=2000	N=4000
True positive	10%	0.945	0.960	0.979	0.929	0.954	0.969
	20%	0.933	0.956	0.968	0.910	0.924	0.956
True negative	0%	0.979	0.988	0.993	0.973	0.984	0.990
	10%	0.977	0.988	0.992	0.971	0.983	0.989
	20%	0.973	0.985	0.990	0.965	0.978	0.987

3.7 Real Data Analysis

Seventeen items from Block 4 of the Trends in International Mathematics and Science Study (TIMSS) 2007 eigth-grade mathematics assessment were analyzed in this study. The Q-matrix for these items was developed by L. Ma (2014) using multiple regression and the least squares distance method. Note that L. Ma (2014) considers both cognitive process attributes and content attributes and builds attributes at two levels. Nevertheless, for illustration purposes, only seven second level content attributes were used for current analysis, including (α_1) whole numbers and integers, (α_2) fractions, decimals, ratio proportion, and percent, (α_3) algebraic expressions and equations/formulas functions, (α_4) geometric shapes, (α_5) geometric measurement and location movement, (α_6) data organization and representation, and (α_7) data interpretation and chance. Out of 17 items, three are polytomously scored with the maximum point of 2. We assume that for these polytomous items, each nonzero category measures all attributes required by the item, and thus the unrestricted $\mathrm{Q_C}$-matrix (W. Ma & de la Torre, 2016) was created. The $\mathrm{Q_C}$-matrix is given in Table 4.4, where the suffix "-1" and "-2" were added to item numbers to indicate category 1 and 2, respectively, for polytomously scored items. Nonmissing responses of 1328 students from the United States including 448 Massachusetts and Minnesota benchmark students were calibrated using the sequential G-DINA model. Note that for dichotomous responses, the sequential G-DINA model is equivalent to the G-DINA model and the stepwise Q-matrix validation is applicable as well.

Based on the stepwise Q-matrix validation procedure, modifications were suggested to nine categories of eight item as shown in Table 4.4. Take Item 17 in Box 1 as an example. Attributes 2 and 7 were assumed to be required, but the Q-matrix validation method suggested that the Attribute 2 may not be necessary. A close scrutiny of this item reveals that it may be solved by using intuition when students understand the question well, and thus Attribute 2 may not be required. It is worth emphasizing that the stepwise Q-matrix validation method should be used with the intent of providing ancillary information to aid experts judgments rather than to replace the experts in determining the association between attributes and items.

Table 3.8: Q_C-matrix for the TIMSS 2007 data

Item No.	TIMSS Item ID	α_1	α_2	α_3	α_4	α_5	α_6	α_7
1	M042001	1	0	0	0	0	0	$\underline{0}$
2	M042022	1	0	$\underline{0}$	0	0	0	0
3	M042082	1	0	1	0	0	0	0
4	M042088	$\underline{0}$	0	1	0	0	0	0
5	M042304A	1	0	0	0	0	0	$\underline{0}$
6-1	M042304B-1	0	1	1	0	0	0	0
6-2	M042304B-2	0	1	1	0	0	0	0
7	M042304C	0	1	1	0	0	0	0
8-1	M042304D-1	1	0	0	0	0	0	0
8-2	M042304D-2	1	0	0	0	0	0	0
9	M042267	1	0	1	0	0	0	0
10	M042239	1	0	1	0	0	0	0
11	M042238	$\underline{1}$	0	1	0	1	0	0
12	M042279	0	0	0	1	0	0	0
13	M042036	1	0	0	$\underline{1}$	1	0	0
14	M042130	1	0	0	1	1	1	0
15	M042303A	0	1	0	0	0	1	0
16-1	M042303B-1	1	0	0	0	1	1	$\underline{1}$
16-2	M042303B-2	1	0	0	0	1	$\underline{1}$	$\underline{1}$
17	M042222	0	$\underline{1}$	0	0	0	0	1

Note: Modifications were suggested to underlined entries.

The model data fits based on the original and suggested Q-matrices were compared. As shown in Table 3.9, the sequential G-DINA model using the suggested Q_C-matrix had better model data fit in terms of both Akaike's (1974) information criterion and Schwarz's (1978) BIC. This implies that the suggested Q_C-matrix is statistically preferred to the original, though this does not guarantee that the suggested one is correct.

Sophie has a bag in which there are 16 marbles: 8 are red and 8 are black marbles. She draws 2 marbles from the bag and does not put them back. Both marbles are black. She then draws a third marble out of the bag. What can you say about the likely color of this third marble?
A. It is more likely to be red than black.
B. It is more likely to be red than black.
C. It is equally likely to be red or black.
D. You cannot tell if red or black is more likely.

Box 1: Item M042222 from TIMSS 2007 Assessment

The estimated proportions of ten dominant latent classes based on the sequential

Table 3.9: Comparison between original and suggested Q_C-matrices

Sequential G-DINA model	-2 log Likelihood	AIC	BIC
Model with original Q_C-matrix	26967.857	27441.857	28672.226
Model with suggested Q_C-matrix	26866.526	27296.526	28412.684

G-DINA model using the original and suggested Q_C-matrices are given in Figure 3.3. In addition to the first two dominant latent classes (i.e., 1111111 and 1101011), there were another three latent classes common to the two Q-matrices. Furthermore, the model based on the suggested Q_C-matrix classified more examinees to extreme latent classes (i.e., 1111111 and 0000000).

Discussion

The importance of a correctly specified Q-matrix has been recently recognized by many researchers, but research on the Q-matrix validation mainly centers on dichotomous responses. The stepwise Q-matrix validation procedure developed in this study can be used to validate the association between attributes and problem-solving steps of polytomously scored items empirically based on a recently developed CDM - the sequential G-DINA model. It can also be applied to dichotomous response data directly without the assumption about the specific forms of the CDMs involved, as long as they are special cases of the G-DINA model. The stepwise Q-matrix validation method incorporates a formal hypothesis test with an effect size measure. Specifically, the procedure intends to identify all statistically required attributes for each nonzero category based on the Wald test, which takes the item parameter estimation errors into account. The GDI or PVAF from de la Torre and Chiu (2016) is used as an effect size measure to exclude the attributes that are identified as statistically necessary but without substantial contributions.

The GDI was used to identify the first required attribute and under most conditions,

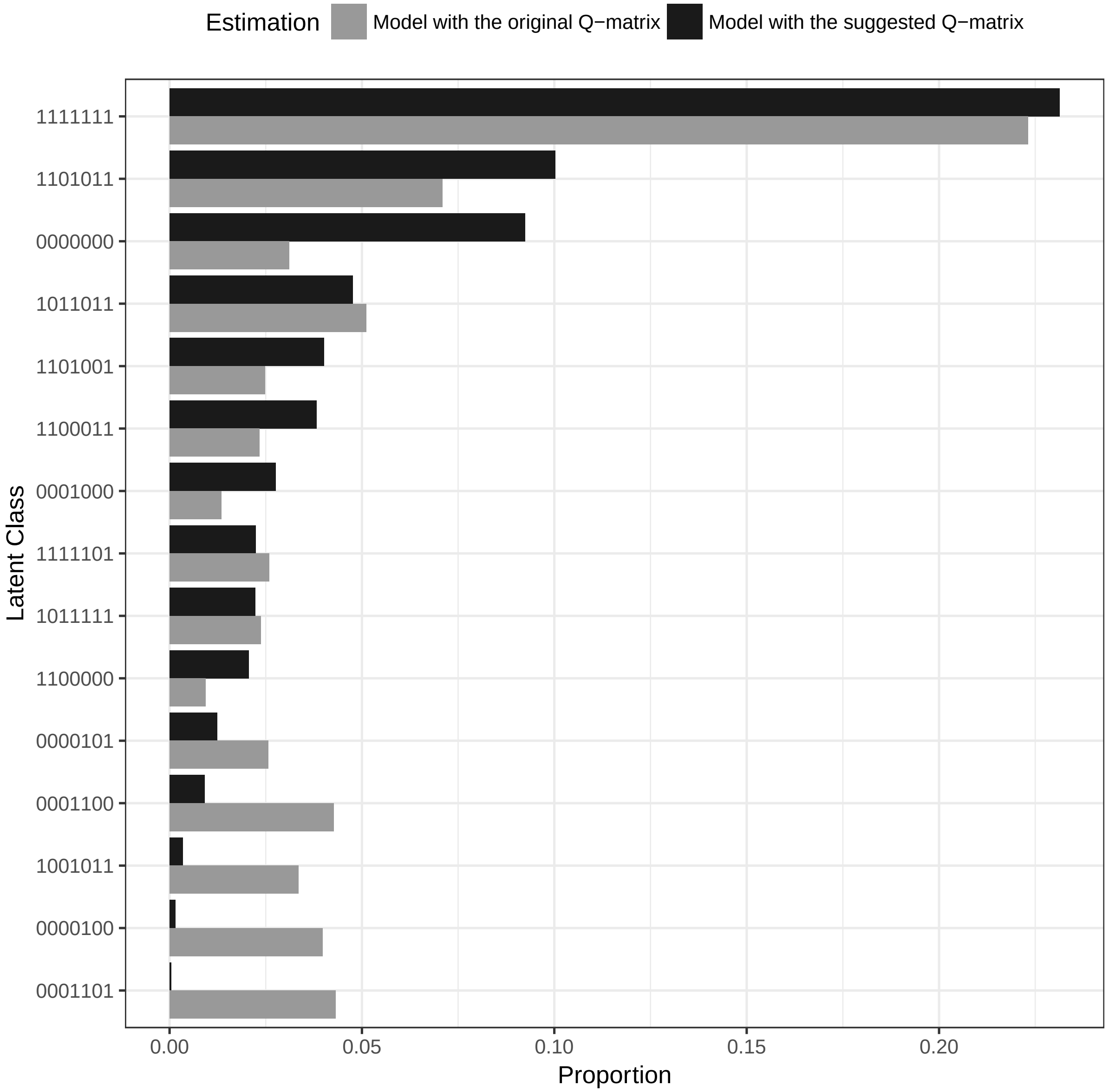


Figure 3.3: Latent class proportion estimation

it performed excellently. We also conducted a simulation study to evaluate whether the stepwise validation method could be further improved if the first required attribute was always selected correctly. It turns out that under the condition $N = 1000$, 20% misspecifications, low item quality, higher-order attribute distribution, and A-CDM processing function, where the GDI had the worst performance in the first step, the stepwise Q-matrix validation can be improved by only 3.3% in terms of the true positive rate, and by 0.3% in terms of the true negative rate. Under other conditions, where the GDI had better performance, the improvements were smaller.

Despite the promising results from the simulated and real data analyses, additional research is needed along this line. First, the effectiveness of the stepwise Q-matrix validation method relies upon the reliable estimation of item parameters, but many factors that could degrade the estimation accuracy have not been investigated yet. For example, the number of attributes measured by the assessment is assumed to be known correctly, but if it is not the case, item parameters may not be estimated accurately. Therefore, the impact of missing one or more required attributes on the stepwise Q-matrix validation procedure is worth investigating. Second, the proposed approach assumes a provisional Q-matrix, which needs to be largely correct. This can be satisfied if a Q-matrix has been developed by domain experts. However, it would be interesting to explore how the proposed methods can be extended to generate Q-matrix without a provisional Q-matrix as in Liu, Xu, and Ying (2012) and Liu, Xu, and Ying (2013). In addition, although the sequential G-DINA can be used for both ordinal and nominal responses, the proposed validation methods are only suitable for ordinal response data. Exploring how to extend the proposed methods for nominal response data is an important direction to consider. Lastly, as a method that can be used for dichotomous response data, it would be interesting to compare it with other Q-matrix validation methods.

3.8 References

Agresti, A. (2013). *Categorical data analysis*. New York: John Wiley & Sons.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Cen, H., Koedinger, K., & Junker, B. (2005). Learning factors analysis: A general method for cognitive model evaluation and improvement. In M. Ikeda, K. Ashley, & T. Chan (Eds.), *Intelligent tutoring systems: 8th International conference* (pp. 164–175). Berlin, Germany: Springer.

Chen, J. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement*. doi: 0146621616686021

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

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8–26.

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. (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.

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*, 355–373.

de la Torre, J., van der Ark, L. A., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*. doi: 10.1177/0748175615569110

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). A review of cognitively diagnostic assessment and a summary of psychometric models. In R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics* (Vol. 26, pp. 979–1030). Amsterdam, Netherlands: Elsevier.

Efroymson, A. (1960). Multiple regression analysis. In A. Ralston & H. S. Wilf (Eds.), *Mathematical methods for digital computers* (pp. 191–203). New York: Wiley.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321.

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

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*, 98–125.

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

Liu, J., Xu, G., & Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli*(5A), 1790–1817.

Ma, L. (2014). *Validation of the item-attribute matrix in TIMSS: Mathematics using multiple regression and the LSDM* (Unpublished doctoral dissertation). University of Denver.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, *69*, 253–275.

Ma, W., & de la Torre, J. (2017). *GDINA: The generalized DINA model framework.* [Computer software version 1.4.2]. Retrieved from `https://CRAN.R-project.org/package=GDINA`

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*, 200–217.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*, 91–100.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the dina model. *Educational and Psychological Measurement*, *68*, 78–96.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, *19*, 506–532.

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

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 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.

Tutz, G. (1997). Sequential models for ordered responses. In W. J. van der Linden

& R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 139–152). New York: Springer-Verlag.

Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A step model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–138). New York: Springer-Verlag.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, *54*, 426–482.

Chapter 4
Category-Level Model Selection for the Sequential G-DINA Model

Abstract

Solving a problem usually requires performing a sequence of steps successfully. Each step may measure different skills, and the required skills may contribute to the performance in each step in various ways. The sequential generalized deterministic noisy "and" gate model is a general cognitive diagnosis model (CDM) for polytomously scored/graded response items of this type. Despite a host of dichotomous CDMs that may be used to parameterize the success probability for each step, specifying the most appropriate model remains challenging. However, if the model for each step is not in accordance with the underlying cognitive processes, the validity of the inference could be questionable. This study aims to evaluate whether several hypothesis tests, namely, the Wald test using various variance-covariance matrices, the likelihood ratio test and the likelihood ratio test using approximated parameters, can be used to select appropriate CDMs for each step of a graded response item. Simulation studies are conducted to examine the Type I error and power of the hypothesis tests under varied conditions. A data set from the TIMSS 2007 Mathematics assessment is analyzed as an illustration.

4.1 Introduction

Cognitive diagnosis models (CDMs) have attracted considerable attention recently within the field of educational measurement. CDMs are multidimensional psychometric models aiming to uncover individuals' profiles on a set of latent variables from their observed item responses in an assessment. The latent variables are referred to as attributes, and represent the skills or cognitive processes of interest in educational assessment. Typically, although not always, the latent variables are assumed to be binary indicating either mastery or nonmastery of the attributes.

A number of cognitive diagnosis models (CDMs) have been developed (for reviews, see DiBello, Roussos, & Stout, 2007). To understand these models, the so-called condensation rule (Maris, 1999) is critical. A condensation rule specifies how latent variable responses are "condensed" to produce a manifest item response. For example, based on a conjunctive condensation rule, the deterministic inputs, noisy "and" gate (DINA; Haertel, 1989) model assumes that individuals are expected to answer an item correctly only when they possess all required attributes; whereas based on the disjunctive condensation rule, the deterministic inputs, noisy "or" gate (DINO; Templin & Henson, 2006) model assumes that mastering at least one required attribute can yield a high success probability. The *additive* model (*A*-CDM; de la Torre, 2011), which has an additive condensation rule, assumes that each required attribute contributes to the success probability independently and uniquely. Aside from these specific models, some general CDM frameworks have also been developed, namely the generalized DINA (G-DINA; de la Torre, 2011) model, the log-linear CDM (LCDM; Henson, Templin, & Willse, 2009), and the general diagnostic model (GDM; von Davier, 2008). Note that the G-DINA model and LCDM consider all main effects of latent variables and all possible interactions among them. By setting appropriate constraints, specific models with conjunctive, disjunctive or additive condensation rules can be obtained as special cases.

Specifying the condensation rule for each item is largely based on experts' judgment, and thus could be subjective. A misspecification in the condensation rules produces the use of inapporiate CDMs, which then results in a model-data misfit (Kunina-Habenicht, Rupp, & Wilhelm, 2012; Liu, Tian, & Xin, 2016) and could call into question the validity of inferences. For example, Rojas, de la Torre, and Olea (2012) have shown that fitting the conjunctive model to the data generated from the disjunctive model, or vice versa, can lead to poor attribute estimation. With the development of the general CDMs, some may argue that the general models should be preferred to the reduced models, such as the DINA model, DINO model and A-CDM, because they can provide better model-data fit in terms of the likelihood. However, as noted by W. Ma, Iaconangelo, and de la Torre (2016), the reduced models may still be more appropriate for several reasons. For example, the reduced CDMs usually have more straightforward interpretations because of the corresponding condensation rules. In addition, due to fewer item parameters involved, the reduced models need a smaller sample for accurate parameter estimation. Lastly, W. Ma et al. (2016) have found that the appropriate reduced models can provide better person attribute estimation than the saturated models, especially when the sample size is small.

As emphasized by von Davier (2014), it is important to consider other alternatives prior to committing to using one particular model. A few studies along this line can be found in literature. For example, Chen, de la Torre, and Zhang (2013), Henson et al. (2009), and Sinharay and Almond (2007) evaluated and compared different models using Akaike's (1974) information criterion, Schwarz's (1978) BIC, and deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) at the test level. A limitation of model comparison at the test level is that all items are typically assumed to conform to the same model, which, most likely, is not the case in practice. At the item level, Henson et al. (2009) provided a way to determine the appropriate reduced model by visual inspection of the estimates of LCDM, and Sinharay and Almond (2007) checked the item fit plots generated from a residual analysis. De la Torre (2011) proposed to use the Wald test (Wald, 1943) to evaluate whether the reduced models subsumed by the G-DINA model can be used in place of the saturated G-DINA model without a significant loss in model-data fit. The Type I error and power of the Wald test for comparing the G-DINA model and the DINA model, DINO model and *A*-CDM were examined in de la Torre and Lee (2013), and the performance of the Wald test in comparing the G-DINA model with the logistic linear model (Maris, 1999) and reduced reparameterized unified model (Hartz, 2002) was later investigated in W. Ma et al. (2016). Although the model selected by the Wald test can provide better person classification, the Type I error rates of the Wald test are found to be inflated, especially when sample size is small or item quality is low (de la Torre & Lee, 2013; W. Ma et al., 2016), which may be, among other things, caused by the underestimation in the variance-covariance matrix of item parameters.

Research on model comparison, or condensation rule selection, mainly focus on dichotomous responses, though graded response data are very common in educational measurement because of the widely use of the constructed-response items. In this paper, we consider polytomously scored items that are solved in a sequential manner. For this type of items, it is not possible for students to perform a step successfully unless all previous steps have been completed correctly. The sequential G-DINA model (W. Ma & de la Torre, 2016) is a general CDM suitable for items of this type, and steps involved in a problem-solving sequence are modeled separately using the dichotomous G-DINA model (de la Torre, 2011). However, different steps may require different attributes and involve different condensation rules. Using the saturated G-DINA model may not always be the optimal choice due to the same reasons in dichotomous responses. This study aims to evaluate the performance of the Wald test and likelihood ratio (LR) test in selecting appropriate condensation rules for each step of the graded response items based on the sequential G-DINA model.

4.2 Overview of the Sequential G-DINA Model

Suppose a test measuring K attributes has J items, and also suppose that item j has $H_j + 1$ response categories (i.e., category 0, 1, $\ldots, H_j$). A binary q-vector $\boldsymbol{q}_{jh} = \{q_{jhk}\}$ is associated with nonzero category h of item j, where $h \in (1, \ldots, H_j)$. Element $q_{jhk} = 1$ indicates that the attribute k is required by step h of item j, and $q_{jhk} = 0$ otherwise. A collection of $\boldsymbol{q}_{jh}$ produces a category level Q-matrix, or $\mathrm{Q_C}$-matrix, which is a $\sum_{j=1}^{J} H_j \times K$ binary matrix. If all items are scored dichotomously, the $\mathrm{Q_C}$-matrix is equivalent to the traditional Q-matrix (Tatsuoka, 1983). Individuals can be grouped into 2^K latent classes because of the K attributes involved in the assessment. Individuals in the same latent class have the same attribute pattern. For latent class c, denote the attribute pattern as $\boldsymbol{\alpha}_c = (\alpha_{c1}, \ldots, \alpha_{cK})$, where $c = 1, \ldots, 2^K$. Element $\alpha_{ck} = 1$ indicates attribute k is mastered by individuals in the latent class, and $\alpha_{ck} = 0$ indicates attribute k is not mastered.

The sequential G-DINA model (W. Ma & de la Torre, 2016) assumes that solving an item involves a sequence of steps, and that individuals cannot complete a step unless they have already performed the previous step successfully. Let X_{ij} be the response of individual i to item j, and $s_{jh}(\boldsymbol{\alpha}_c) = P(X_{ij} \geq h | X_{ij} \geq h-1, \boldsymbol{\alpha}_c)$ be the *processing function* (W. Ma & de la Torre, 2016), or the probability of performing step h correctly given that step $h-1$ has been completed successfully. The conditional probability of obtaining score h on item j can be written as

$$P(X_{ij} = h | \boldsymbol{\alpha}_c) = [1 - s_{j,h+1}(\boldsymbol{\alpha}_c)] \prod_{x=0}^{h} s_{jx}(\boldsymbol{\alpha}_c),$$

where $s_{j0}(\boldsymbol{\alpha}_c) \equiv 1$ and $s_{j,H_j+1}(\boldsymbol{\alpha}_c) \equiv 0$.

To model the processing function, the G-DINA model (de la Torre, 2011) is used. Specifically, for step h of item j, let $\boldsymbol{\alpha}^*_{ljh}$ be the reduced attribute pattern consisting of the required attributes for this step only. Without loss of generality, the first K^*_{jh} attributes are assumed to be required, that is, $l = 1, \cdots, 2^{K^*_{jh}}$. The processing function can be expressed as

$$g\left[s_{jh}(\boldsymbol{\alpha}^*_{ljh})\right] = \phi_{jh0} + \sum_{k=1}^{K^*_{jh}} \phi_{jhk}\alpha_{lk} + \sum_{k'=k+1}^{K^*_{jh}} \sum_{k=1}^{K^*_{jh}-1} \phi_{jhkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \phi_{jh12\cdots K^*_{jh}} \prod_{k=1}^{K^*_{jh}} \alpha_{lk},$$

where $g[\cdot]$ is the identity, log or logit link function. By setting appropriate constraints to the identity link model as in de la Torre (2011), the DINA model, DINO model or A-CDM can be used as the processing function, and different models can be used at different steps within a single item. Specifically, the DINA model is obtained when all main effects and interaction terms except the highest-order interaction are set to be 0:

$$s_{jh}(\boldsymbol{\alpha}^*_{ljh}) = \phi_{jh0} + \phi_{jh12\cdots K^*_{jh}} \prod_{k=1}^{K^*_{jh}} \alpha_{lk}.$$

The processing function based on the DINO model is given by

$$s_{jh}(\boldsymbol{\alpha}^*_{ljh}) = \phi_{jh0} + \phi_{jhk}\alpha_{lk},$$

where $\phi_{jhk} = -\phi_{jhk'k''} = \cdots = (-1)^{K^*_{jh}+1}\phi_{jh12\cdots K^*_{jh}}$, for $k = 1, \cdots, K^*_{jh}$, $k^{'} = 1, \cdots, K^*_{jh} - 1$, and $k^{''} > k^{'}, \cdots, K^*_{jh}$. The A-CDM processing function is the constrained identity-link G-DINA model without the interaction terms. It can be formulated as

$$s_{jh}(\boldsymbol{\alpha}^*_{ljh}) = \phi_{jh0} + \sum_{k=1}^{K^*_{jh}} \phi_{jhk}\alpha_{lk}.$$

4.3 Category-Level Model Comparison

If a response category only requires one attribute, the G-DINA model and other reduced CDMs (e.g., DINA model, DINO model, A-CDM) are not distinguishable, which implies that all condensation rules are equivalent. Therefore, model comparison is only necessary for categories requiring two or more attributes, which are referred to as multi-attribute categories. This section introduces how the LR test and Wald test can be used for model comparisons.

4.3.1 Likelihood Ratio Test

Let $\boldsymbol{s}_j = \{\boldsymbol{s}_{jh}\}$ be a vector of processing functions of all categories of item j, where $\boldsymbol{s}_{jh} = \{s_{jh}(\boldsymbol{\alpha}^*_{ljh})\}$, and $\boldsymbol{s} = \{\boldsymbol{s}_j\}$ for all items. Also, let $\boldsymbol{\pi} = \{\pi_c | c = 2, 3, \ldots, 2^K\}$ be free latent class proportion parameters, $\pi_1 = 1 - \sum_{c=2}^{2^K} \pi_c$, and $\boldsymbol{\psi} = (\boldsymbol{s}, \boldsymbol{\pi})$. The log marginalized likelihood of response vector $\boldsymbol{X}_i$ for individual i is

$$\ell(\boldsymbol{\psi}; \boldsymbol{X}_i) = \log \sum_{c=1}^{2^K} \pi_c P(\boldsymbol{X}_i | \boldsymbol{\alpha}_c),$$

and the log marginalized likelihood of responses for all individuals is

$$\ell(\boldsymbol{\psi}; \boldsymbol{X}) = \sum_{i=1}^{N} \ell(\boldsymbol{\psi}; \boldsymbol{X}_i).$$

The LR test has been widely used to compare two nested models: A compact model and an augmented model. Let $\ell(\boldsymbol{\psi}_C; \boldsymbol{X})$ and $\ell(\boldsymbol{\psi}_A; \boldsymbol{X})$ be the log likelihood of the compact and augmented models, respectively. The LR statistic can be written as

$$LR = -2\left[\ell(\boldsymbol{\psi}_C; \boldsymbol{X}) - \ell(\boldsymbol{\psi}_A; \boldsymbol{X})\right],$$

which follows a χ^2 distribution with the degrees of freedom equal to the difference in the number of parameters estimated for the two models.

In this study, the LR test is conducted category by category, and item by item. For the augmented model, the G-DINA model is used as the processing functions for all categories of all items, whereas for the compact model, the G-DINA model is used as the processing functions for all categories except the studied category, for which, a reduced model is used as the processing function. The augmented model only needs to be calibrated once. Given that the reduced models to be tested for each category in this study include the DINA model, DINO model and A-CDM, the data needs to be calibrated $1+3\sum_{j=1}^{J}\sum_{h=1}^{H_j} I(K_{jh}^{*}>1)$ times.

The LR test could be time-consuming, and thus we also consider an EM-based approximation, which is referred to as two-step LR test (Sorrel, de la Torre, Abad, & Olea, in press). Specifically, the processing functions under a reduced model are estimated using a one-step EM algorithm based on the estimates under the G-DINA processing functions directly without recalibrating the data. When the DINA or DINO model is used as the processing function, some reduced latent groups are equivalent in that they have the same processing function. We use a vector of length $2^{K_{jh}^{*}}$, $\boldsymbol{\omega}_{jh}$, to denote the equivalent reduced latent groups for category h of item j, where $\omega_{ljh}=g$ if $\boldsymbol{\alpha}_{ljh}^{*}$ is in the gth set of the equivalent reduced latent groups. For example, suppose $\boldsymbol{\alpha}_{jh}^{*}=(00,10,01,11)$. $\boldsymbol{\omega}_{jh}=(1,2,3,4)$ for the G-DINA processing function, $(1,1,1,2)$ for the DINA processing function, and $(1,2,2,2)$ for the DINO processing function. It can be shown that the marginal maximum likelihood estimates of $s_{jh}(\boldsymbol{\alpha}_{ljh}^{*})$ when $h\geq 1$ is given by

$$\hat{s}_{jh}(\boldsymbol{\alpha}_{ljh}^{*})=\frac{R_{h}^{+}(\boldsymbol{\alpha}_{ljh}^{*})}{R_{h-1}^{+}(\boldsymbol{\alpha}_{ljh}^{*})}.$$

where $R_{h}^{+}(\boldsymbol{\alpha}_{ljh}^{*})$ is the expected number of examinees in reduced latent group l and other equivalent groups getting at least a score of h, and can be calculated as,

$$R_h^+(\boldsymbol{\alpha}_{ljh}^*) = \sum_{i=1}^{N} \sum_{\{l':\omega_{l'jh}=\omega_{ljh}\}} P(\boldsymbol{\alpha}_{l'jh}^*|\boldsymbol{X}_i) I(X_{ij} \geq h).$$

Note that $P(\boldsymbol{\alpha}_{l'jh}^*|\boldsymbol{X}_i)$ is calculated based on the augmented model. When the processing function is the A-CDM, the following log likelihood function is maximized for a studied category while the processing functions of other categories are hold constant,

$$\sum_{l=1}^{2^{K_j^*}} \sum_{h=1}^{H_j} \bar{r}_{ljh} \log \left[P(X_{ij} = h|\boldsymbol{\alpha}_{lj}^*) \right].$$

4.3.2 The Wald Test

To use the Wald test to examine whether a reduced model can be used in place of the G-DINA model as the processing function for a multi-attribute category, category h of item j, a $(2^{K_{jh}^*} - m) \times 2^{K_{jh}^*}$ restriction matrix $\boldsymbol{R}$ needs to be set up so that under the null hypothesis, $\boldsymbol{R} \times \boldsymbol{s}_{jh} = \mathbf{0}$, where m is the number of parameters involved in this category when a reduced CDM is used as the processing function, and $\boldsymbol{s}_{jh}$ is a vector consisting of the processing functions for category h when the G-DINA model is used. For example, when $K_{jh}^* = 3$, the null hypothesis for the A-CDM is

$$\begin{bmatrix} 1 & -1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 0 & 1 & 0 \\ -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \end{bmatrix} \times \begin{bmatrix} s_{jh}(000) \\ s_{jh}(100) \\ s_{jh}(010) \\ s_{jh}(001) \\ s_{jh}(110) \\ s_{jh}(101) \\ s_{jh}(011) \\ s_{jh}(111) \end{bmatrix} = \mathbf{0}.$$

Examples of the restriction matrices for the DINA and DINO models can be found in de la Torre and Lee (2013). The Wald statistic can be calculated as

$$W = \left[\boldsymbol{R} \times \hat{\boldsymbol{s}}_{jh}\right]' \left[\boldsymbol{R} \times \boldsymbol{V}(\hat{\boldsymbol{s}}_{jh}) \times \boldsymbol{R}'\right]^{-1} \left[\boldsymbol{R} \times \hat{\boldsymbol{s}}_{jh}\right],$$

where $\boldsymbol{V}(\hat{\boldsymbol{s}}_{jh})$ is the variance-covariance matrix of the processing functions for category h of item j. The Wald statistic W is asymptotically χ^2 distributed with $2^{K_{jh}^*} - m$ degrees of freedom.

The (observed) Fisher information can be approximated using the outer product of gradients of the log marginalized likelihood as follows:

$$\mathscr{I}_{\hat{\psi}} = \frac{1}{N} \sum_{i=1}^{N} \left[\left(\frac{\partial \ell(\boldsymbol{\psi}; \boldsymbol{X}_i)}{\partial \boldsymbol{\psi}} \right) \left(\frac{\partial \ell(\boldsymbol{\psi}; \boldsymbol{X}_i)}{\partial \boldsymbol{\psi}} \right)' \right] \Bigg|_{\boldsymbol{\psi} = \hat{\boldsymbol{\psi}}}.$$

The score function for the processing function $\boldsymbol{s}$ is a vector with elements

$$\frac{\partial \ell(\boldsymbol{\psi}; \boldsymbol{X}_i)}{\partial s_{jh}(\boldsymbol{\alpha}_{ljh}^*)} = P(\boldsymbol{\alpha}_{ljh}^* | \boldsymbol{X}_i) \left[\frac{I(X_{ij} \geq h)}{s_{jh}(\boldsymbol{\alpha}_{ljh}^*)} - \frac{I(X_{ij} = h-1)}{1 - s_{jh}(\boldsymbol{\alpha}_{ljh}^*)} \right].$$

The score function for the latent class proportion parameters $\boldsymbol{\pi}$ has elements

$$\frac{\partial \ell(\boldsymbol{\psi}; \boldsymbol{X}_i)}{\partial \pi_c} = \frac{P(\boldsymbol{X}_i | \boldsymbol{\alpha}_c) - P(\boldsymbol{X}_i | \boldsymbol{\alpha}_1)}{P(\boldsymbol{X}_i)}.$$

The variance-covariance matrix of $\boldsymbol{\psi}$ can be written as

$$\boldsymbol{V}(\hat{\boldsymbol{\psi}}) = \begin{bmatrix} \boldsymbol{V}(\hat{\boldsymbol{s}}) & \boldsymbol{V}(\hat{\boldsymbol{s}}, \hat{\boldsymbol{\pi}}) \\ \boldsymbol{V}(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{s}}) & \boldsymbol{V}(\hat{\boldsymbol{\pi}}) \end{bmatrix},$$

where $\boldsymbol{V}(\hat{\boldsymbol{s}})$ and $\boldsymbol{V}(\hat{\boldsymbol{\pi}})$ are variance-covariance matrices for processing functions and latent class proportion parameters, respectively. $\boldsymbol{V}(\hat{\boldsymbol{s}})$ is of dimension $\sum_{j=1}^{J} \sum_{h=1}^{H_j} 2^{K_{jh}^*} \times$

$\sum_{j=1}^{J}\sum_{h=1}^{H_j}2^{K_{jh}^*}$ and $\boldsymbol{V}(\hat{\boldsymbol{\pi}})$ is of dimension $(2^K-1)\times(2^K-1)$. $\text{Cov}(\hat{\boldsymbol{s}},\hat{\boldsymbol{\pi}})$ is the covariance between processing functions and latent class proportions. $\boldsymbol{V}(\hat{\boldsymbol{\psi}})$ can be calculated by inverting the information matrix, as in, $\boldsymbol{V}(\hat{\boldsymbol{\psi}})=\mathscr{I}_{\hat{\boldsymbol{\psi}}}^{-1}$, and $\boldsymbol{V}(\hat{\boldsymbol{s}}_{jh})$ used to calculate the Wald statistic is a submatrix of $\boldsymbol{V}(\hat{\boldsymbol{\psi}})$ associated with category h of item j. Because all free parameters are considered in $\mathscr{I}_{\hat{\boldsymbol{\psi}}}$, it is referred to as the complete information matrix (Philipp, Strobl, de la Torre, & Zeileis, 2016).

In practice, only item parameters are typically of interest, and thus, some previous studies (e.g., de la Torre, 2008) calculate the variance-covariance matrix of item parameters by inverting the information matrix for each item separately. For the sequential G-DINA model, this can be obtained by inverting the following $\sum_{h=1}^{H_j}2^{K_{jh}^*}\times\sum_{h=1}^{H_j}2^{K_{jh}^*}$ itemwise information matrix,

$$\mathscr{I}_{\hat{\boldsymbol{s}}_j}=\frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{\partial\ell(\boldsymbol{\psi};\boldsymbol{X}_i)}{\partial\boldsymbol{s}_j}\right)\left(\frac{\partial\ell(\boldsymbol{\psi};\boldsymbol{X}_i)}{\partial\boldsymbol{s}_j}\right)'\right]\Bigg|_{\boldsymbol{s}_j=\hat{\boldsymbol{s}}_j}.$$

Calculating covariance matrix based on this itemwise information matrix is based on an implicit assumption that parameters are independent among items, and between items and latent classes. Although inverting the itemwise information matrix can be much faster than inverting the complete information, the Wald test based on the itemwise information matrix has been shown to have inflated Type I errors under some conditions for model comparison and differential item functioning for dichotomous responses (de la Torre & Lee, 2013; Hou, de la Torre, & Nandakumar, 2014; W. Ma et al., 2016).

In between the complete information matrix and the itemwise information matrix is an incomplete information matrix, $\mathscr{I}_{\hat{\boldsymbol{s}}}$, which does not consider the latent class proportion parameters (Philipp et al., 2016). It is of dimension $\sum_{j=1}^{J}\sum_{h=1}^{H_j}2^{K_{jh}^*}\times\sum_{j=1}^{J}\sum_{h=1}^{H_j}2^{K_{jh}^*}$. This incomplete information matrix considers the associations among parameters of different items, and its size does not increase exponentially with the number of attributes as the complete information matrix. For example, when there are 15 attributes as in Lee, Park, and Taylan (2011), $\mathscr{I}_{\hat{\pi}}$ is of dimension $2^{15} \times 2^{15}$, or 32768×32768, which may be problematic when calculating the inverse. However, as Philipp et al. (2016) showed, both itemwise and incomplete information matrix produced underestimated standard errors, and their impact on the Wald test needs further investigations.

4.4 Simulation Study

The goal of this simulation study is to systematically evaluate the performance of the LR tests, and the Wald test using different information matrices for category-level model selection in the context of the sequential G-DINA model. The Type I error and power of these statistical tests were examined under varied conditions.

4.4.1 Design

The number of items and attributes were fixed to $J = 23$ and $K = 5$, respectively. The sample sizes were N = 1000, 2000, and 4000. The processing functions for the generating models were the DINA model, DINO model and A-CDM, representing the conjunctive, disjunctive and additive condensation rules. Note that all categories had the same condensation rule for data generation in each condition. Item quality had three levels: $g = 0.1$, 0.2 or 0.3 for all categories of all items, representing high, moderate and low quality, where $g = s_{jh}(\boldsymbol{\alpha}_{ljh}^{*} = \mathbf{0}) = 1 - s_{jh}(\boldsymbol{\alpha}_{ljh}^{*} = \mathbf{1})$ for category h of item j. When the A-CDM was used as the processing function for data generation, each required attribute was assumed to contribute equally to the processing function. The $\mathrm{Q_C}$-matrix is given in Table 4.1, where each attribute was measured 13 times. The $\mathrm{Q_C}$-matrix consists of six two-attribute response categories and six three-attribute response categories distributed uniformly at categories 1, 2 and 3. Attribute patterns were generated from the uniform distribution. Under each condition, 1000 data sets were simulated. The GDINA package (W. Ma & de la Torre, 2017) was used for data simulation and model estimation. The code for model comparison was written in the R programming environment (R Core Team, 2017). The Type I error rates of the Wald and LR tests were investigated at 10 significance levels: .01, .02, .03, .04, .05, .06, .07, .08, .09 and .10.

Table 4.1: Q_C-matrix for the simulation study

Item	Cat	α_1	α_2	α_3	α_4	α_5	Item	Cat	α_1	α_2	α_3	α_4	α_5
1	1	1	0	0	0	0	13	1	1	0	0	0	0
1	2	0	1	0	0	0	13	2	0	1	0	0	0
2	1	0	0	1	0	0	13	3	0	0	1	0	0
2	2	0	0	0	1	0	14	1	0	0	0	1	0
3	1	0	0	0	0	1	14	2	0	0	1	0	0
3	2	1	0	0	0	0	14	3	0	1	0	0	0
4	1	0	1	0	0	0	15	1	0	0	0	0	1
4	2	0	0	1	0	0	15	2	0	0	1	0	0
5	1	0	0	1	1	0	15	3	1	0	0	1	0
5	2	0	0	0	0	1	16	1	0	0	0	1	0
6	1	1	1	0	0	0	16	2	1	0	0	0	0
6	2	0	0	0	1	0	16	3	0	0	1	0	1
7	1	0	0	1	1	1	17	1	1	0	0	0	0
7	2	0	1	0	0	0	17	2	0	1	0	0	0
8	1	1	1	0	1	0	17	3	0	1	0	1	1
8	2	0	0	0	0	1	18	1	0	1	0	0	0
9	1	0	0	0	0	1	18	2	0	0	0	1	0
9	2	1	0	1	0	0	18	3	1	0	1	0	1
10	1	0	1	0	0	0	19	1	1	0	0	0	0
10	2	1	0	0	0	1	20	1	0	1	0	0	0
11	1	0	0	0	0	1	21	1	0	0	1	0	0
11	2	0	1	1	1	0	22	1	0	0	0	1	0
12	1	0	0	1	0	0	23	1	0	0	0	0	1
12	2	1	0	0	1	1							

Note: Cat represents the level of response category.

4.4.2 Results

4.4.2.1 Type I Error

Type I error or false positive occurs when a hypothesis test concludes that the G-DINA processing function is statistically better than the generating processing function. For each of the multi-attribute response categories, the (observed) Type I error rate or false positive rate is the percentage of times that the hypothesis test makes the Type I error out of the 1000 replications under a specific significance level. The Type I error rates were averaged across categories with the same K_{jh}^* and the level of the response category. Figures 4.1 to 4.3 give the Type I error rates of the Wald test, two-step LR test, and LR test when the processing function was the A-CDM. The Type I error rates may not be equal to the significance level exactly due to the sampling errors, even when the tests conform well to the nominal level p. However, the Type I error rates are expected to have a probability of 95% of falling within $p \pm 1.96\sqrt{p(1-p)/n}$ under a nominal level of p with n replications. This region is established for different significance levels, and shown as the gray ribbon in Figures 4.1 to 4.3. The black lines in the figures are reference lines. The Type I error rates when the processing function was the DINA or DINO model are given in Appendix 4.7.

Regarding the Wald test, across all conditions, using the itemwise information yielded the largest Type I error rates, whereas using the complete information produced the smallest Type I error rates. This was expected because inverting a submatrix of the complete information produces underestimated variances, as showed in Philipp et al. (2016). From Figure 4.1 where $N = 1000$, the Wald test using the incomplete and complete information matrices tended to be conservative when item quality was high, but the Wald test using the itemwise information performed well. However, when the processing function was the DINA or DINO model, the Wald test using the itemwise information performed poorly when items were of high quality, $N = 1000$ and $K_{jh}^* = 3$.

More specifically, the test tended to be inflated for lower category levels, but overly conservative for the higher category level. When items were of moderate quality and $K^*_{jh} = 3$, the Wald test using the itemwise information showed inflated type I error regardless the generating processing functions. Additionally, when items were of low quality, the Wald test had inflated type I error regardless of the information matrices used and the generating processing functions, but the inflation was out of control when using the itemwise information especially when $K^*_{jh} = 3$.

The two-step LR test and the LR test performed similarly well when items were of high or moderate quality. When items were of low quality, both the LR test and two-step LR test produced inflated type I error, but the inflation is much more severe for the two-step LR test. Under the low item quality condition, the LR test performed similarly as the Wald test using the complete or incomplete information matrix, whereas the two-step LR test performed similarly as the Wald test using the itemwise information matrix. Last, the impact of sample size was apparent. With larger samples, the Type I error rates for all hypothesis tests were closer to the nominal levels.

4.4.2.2 Power

Receiver Operating Characteristic Curves. True positive occurs when a reduced processing function is rejected correctly by a hypothesis test. The true positive rate is defined as the percentage of times that a hypothesis test rejects a reduced processing function correctly out of the 1000 replications under a specific significance level. Like the false positive rate, the true positive rates were also averaged across categories with the same K^*_{jh} and the level of the response category. The previous section shows that different hypothesis tests have different false positive rates, which makes it difficult to compare their true positive rates directly. As a result, the receiver operating characteristic (ROC) curves were drawn by plotting the true positive rates against the false positive rates at varied nominal levels.

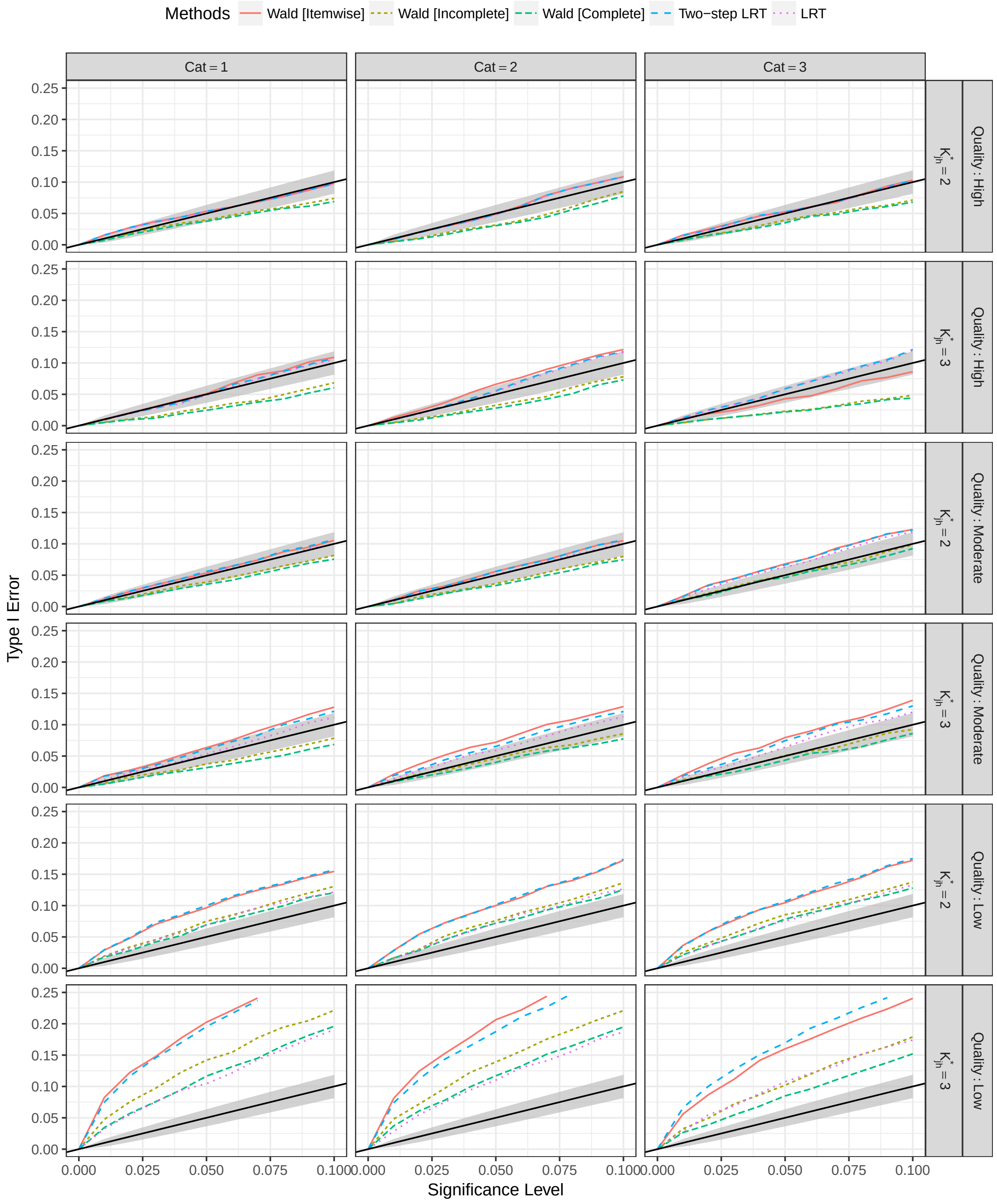


Figure 4.1: Type I error for the *A*-CDM under $N = 1000$

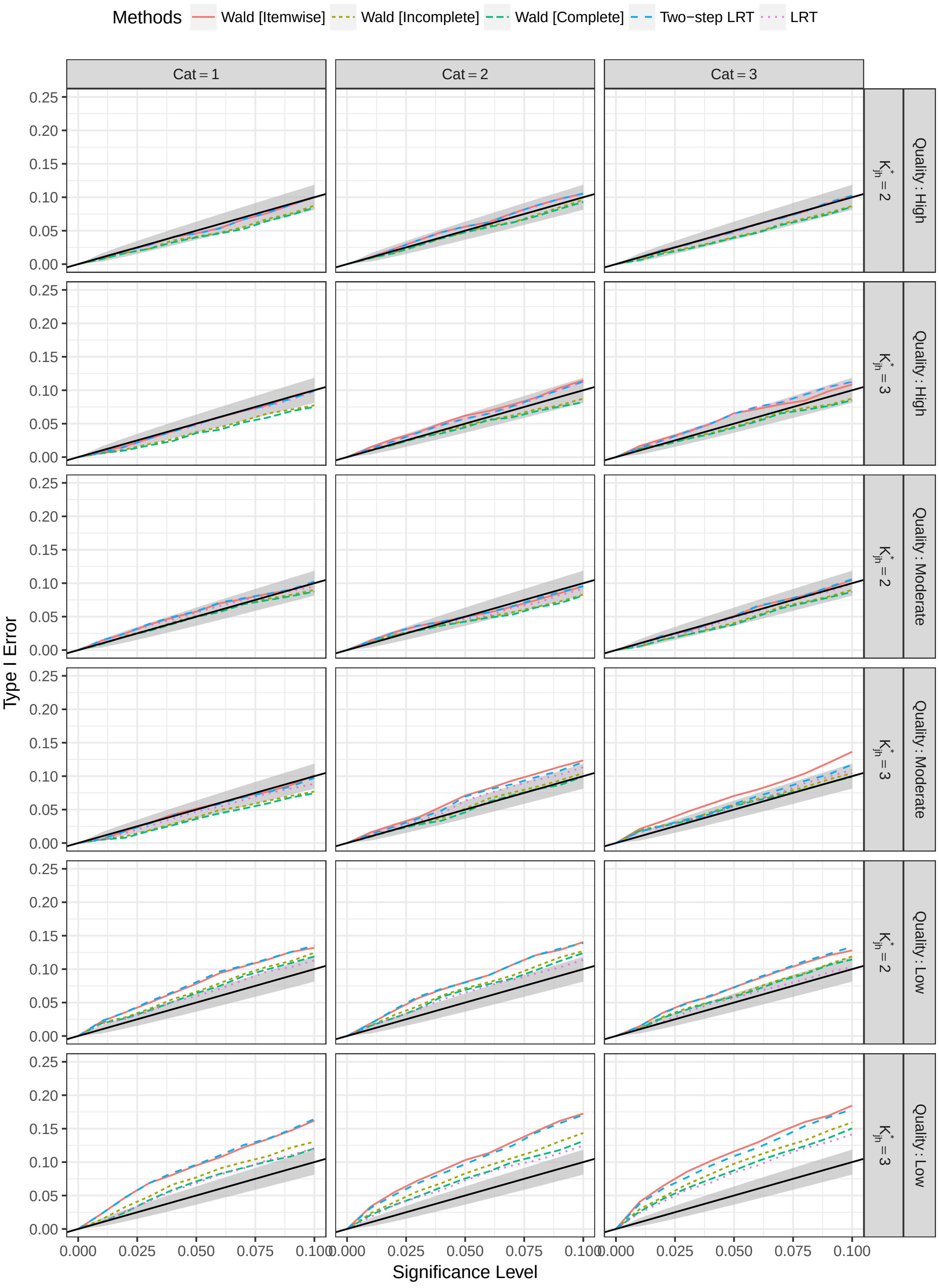


Figure 4.2: Type I error for the A-CDM under $N = 2000$

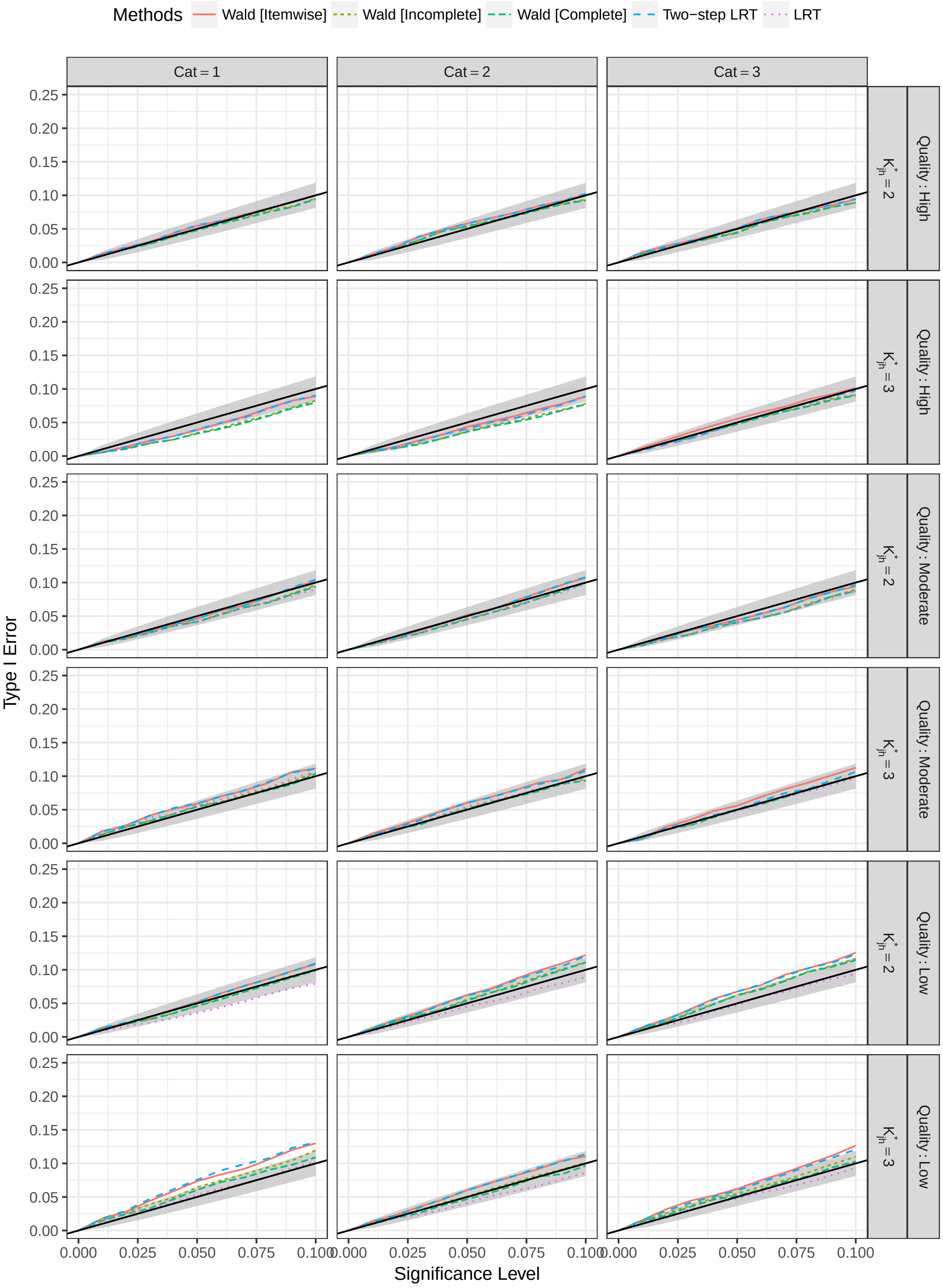


Figure 4.3: Type I error for the A-CDM under $N = 4000$

Figures 4.4 and 4.5 give the ROC curves for the DINA and DINO processing functions when the true processing function is the A-CDM and items were of low quality. The ROC curves under other conditions can be found in Appendix 4.7. The ROC curves showed the relationship between the true positive and false positive for each hypothesis test. A hypothesis test is more accurate if its ROC curve is closer to the left-hand and upper borders, and less accurate if closer to the diagonal line. The area under the ROC curve is an overall measure of the classification accuracy. From Figures 4.4 and 4.5, all hypothesis tests had similar ROC curves when sample size was large or the level of category was low. When sample size was small or the level of category was high, the two-step LRT and the Wald test using itemwise information matrix tended to have larger areas under the ROC curve.

Empirical Power Rates. Statistical power, which is the same as the true positive, is used more commonly in the context of hypothesis testing. Although the ROC curves can provide an overall measure of the power, it is more common in practice to conduct a hypothesis test under a nominal level of 0.05, and therefore, examining the corresponding power is necessary. To compare statistical power rates, all hypothesis tests should have the same observed Type I error rate. However, this is not the case as shown in the previous section. As a result, the empirical power rates calculated from the empirical distributions under the null hypothesis were reported instead. Specifically, when the generating model was fitted to the data, the 5th percentile of the p-values for each hypothesis test was calculated and used as the empirical cutoff for each condition. The empirical power rate, which was calculated for each hypothesis test under each condition, is defined as the percentage of p-values that were less than the empirical cutoff under the same condition. Like the Type I error rate, the empirical power rates were also averaged across categories with the same K_{jh}^{*} and the level of the response category. As in de la Torre and Lee (2013), a test power of 0.80 or higher is considered adequate, and of 0.90 or higher excellent. Tables 4.2 and 4.3 give the empirical power rates of the Wald and LR tests for the DINA and DINO processing functions when the generating processing function was the A-CDM across sample sizes, item qualities, response category levels and K^*_{jh}. The empirical power rates when the generating processing functions were the DINA and DINO models are given in the Appendix 4.7.

From Tables 4.2 and 4.3, the empirical power rates of the Wald test and LR test increased as the sample size increased, K^*_{jh} decreased, items quality improved, or the category level decreased. Specifically, the power rates for all tests were excellent when items were of high quality, with a minimum value of 0.931 occurring when the category level was 3, $N = 1000$, and $K^*_{jh} = 3$. When items were moderate quality, the power rates were higher than 0.958 when $N = 2000$ or higher, but can be as low as 0.726 when $N = 1000$. When items were of low quality, the power rates can be very low especially when the category level was high, and the sample size was small. For example, the power rate for the LR test to distinguish the DINA model and A-CDM was merely 0.154 when the category level was 3, $N = 1000$ and $K^*_{jh} = 3$. However, increasing the sample size could improve the power substantially. For example, the empirical power rate for the Wald test using the complete information was improved from 0.2 to 0.708 when N increased from 1000 to 4000, given that the category level was 3, item quality was low and $K^*_{jh} = 3$.

Similar patterns can be observed when the generating processing function was the DINA or DINO model. The power rates were high under the favorable conditions (i.e., higher item quality, larger sample size, lower level of category and smaller K^*_{jh}), but dropped considerably under some unfavorable conditions. In addition, when the processing function was the DINA model, the power rates for the A-CDM were lower than those for the DINO model under all conditions; and when the processing function was the DINO model, the power rates for the A-CDM were lower than those for the DINA model under all conditions. These results imply that distinguishing the conjunctive and disjunctive models is easier than distinguishing them from the additive model.

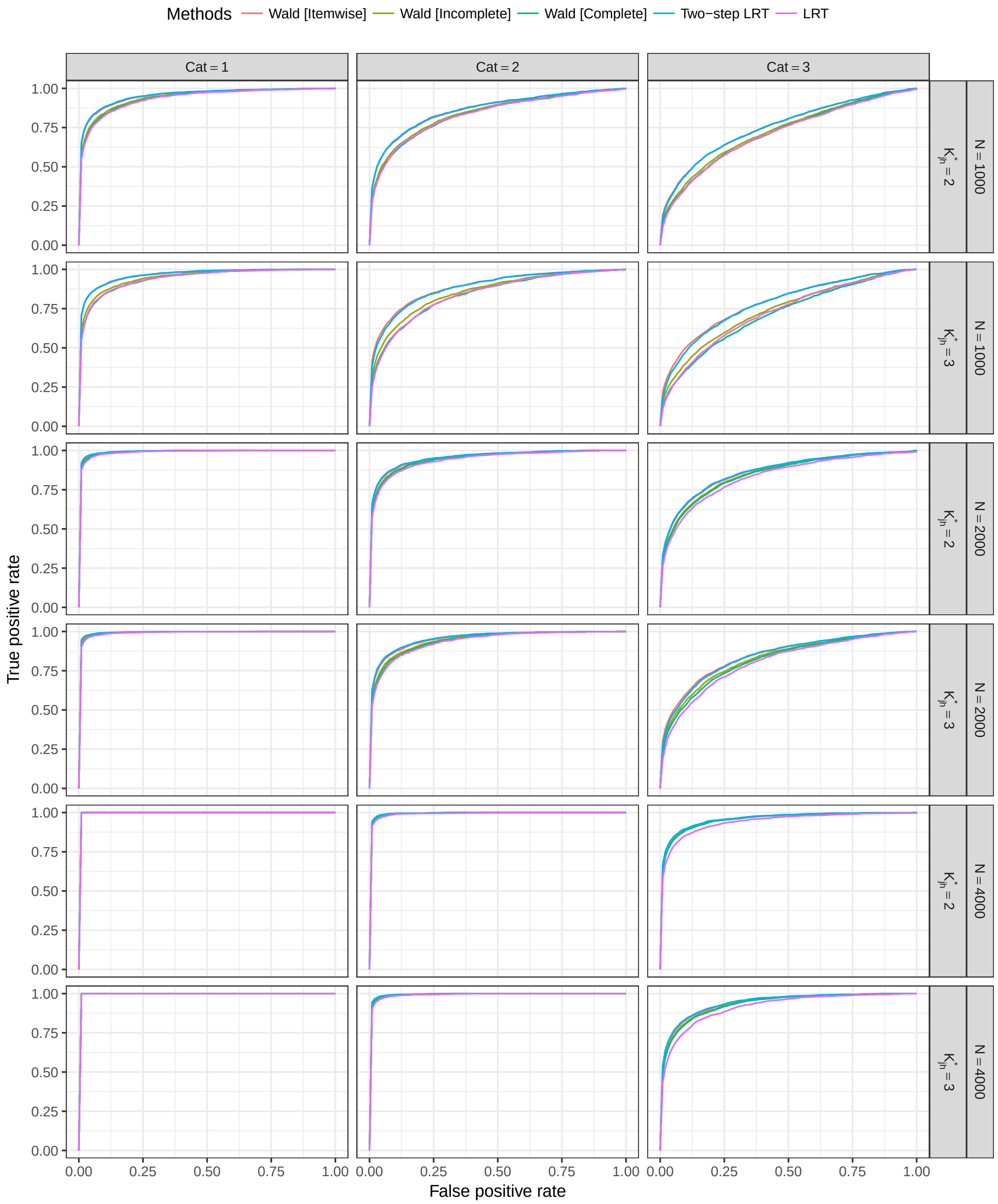


Figure 4.4: ROC curves for the DINA processing function: *A*-CDM generated processing function under low item quality

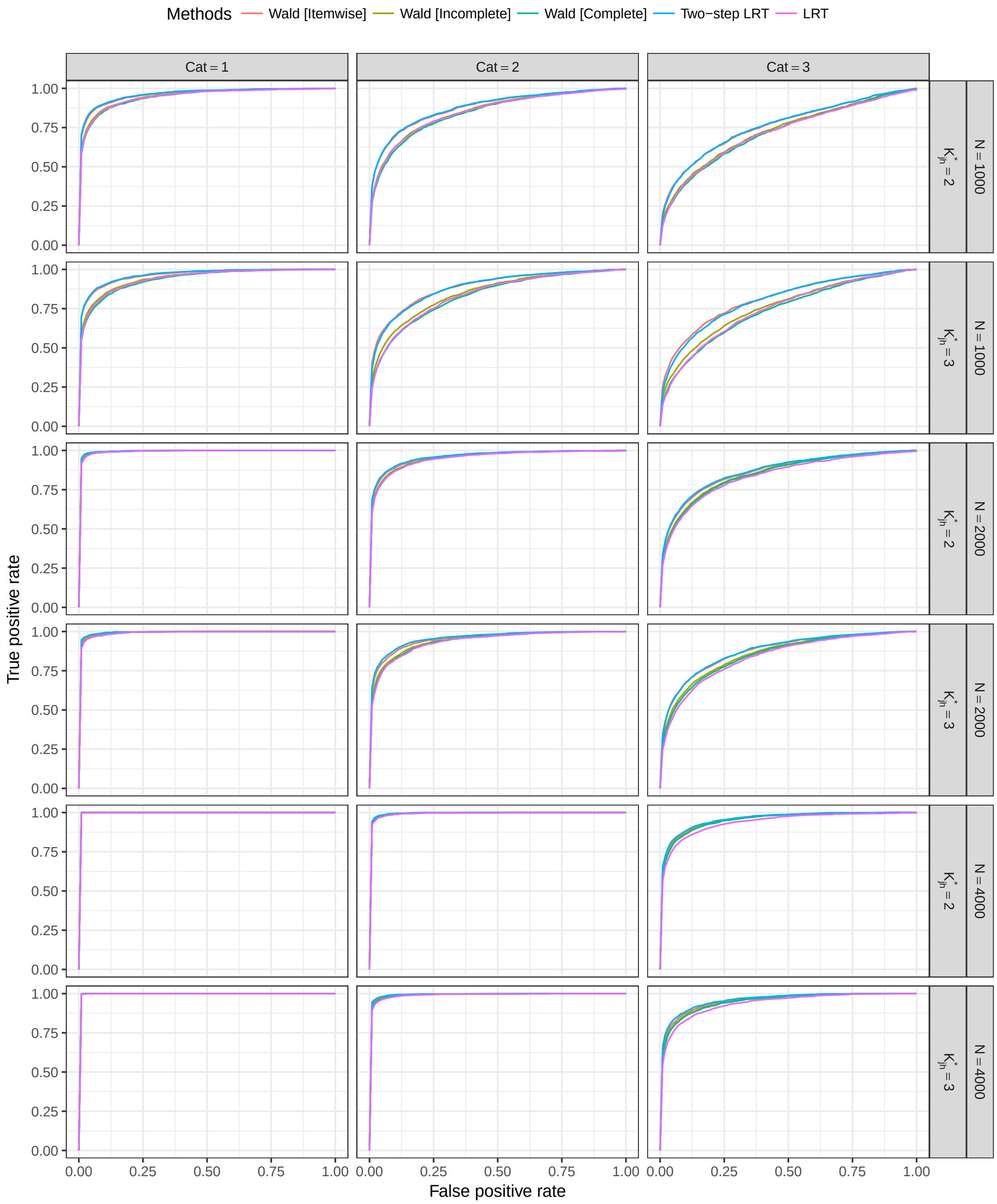


Figure 4.5: ROC curves for the DINO processing function: *A*-CDM generated processing function under low item quality

Table 4.2: Power for the DINA processing function: ACDM-generated data

N	Quality	K_{jh}^*	Wald [Itemwise]			Wald [Incomplete]			Wald [Complete]			Two-step LRT			LRT		
			Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3
1000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	0.940	1.000	1.000	0.936	1.000	1.000	0.931	1.000	1.000	0.950	1.000	1.000	0.951
	Moderate	2	1.000	0.995	0.874	1.000	0.995	0.868	1.000	0.995	0.866	1.000	0.994	0.854	1.000	0.994	0.854
		3	1.000	0.988	0.768	1.000	0.987	0.763	1.000	0.986	0.760	1.000	0.985	0.726	1.000	0.985	0.730
	Low	2	0.729	0.441	0.226	0.722	0.434	0.218	0.718	0.428	0.212	0.736	0.437	0.210	0.707	0.424	0.210
		3	0.610	0.335	0.218	0.610	0.332	0.214	0.621	0.321	0.200	0.618	0.314	0.152	0.624	0.322	0.154
2000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	0.997	1.000	1.000	0.996	1.000	1.000	0.996	1.000	1.000	0.998	1.000	1.000	0.998
	Moderate	2	1.000	1.000	0.995	1.000	1.000	0.995	1.000	1.000	0.994	1.000	1.000	0.994	1.000	1.000	0.994
		3	1.000	1.000	0.960	1.000	1.000	0.958	1.000	1.000	0.958	1.000	1.000	0.958	1.000	1.000	0.958
	Low	2	0.959	0.761	0.462	0.958	0.756	0.454	0.956	0.754	0.454	0.962	0.769	0.470	0.944	0.736	0.440
		3	0.966	0.688	0.344	0.966	0.686	0.339	0.964	0.682	0.335	0.968	0.698	0.322	0.960	0.674	0.299
4000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Moderate	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Low	2	1.000	0.981	0.804	1.000	0.981	0.802	1.000	0.980	0.801	1.000	0.984	0.812	1.000	0.974	0.788
		3	1.000	0.979	0.708	1.000	0.979	0.708	1.000	0.978	0.708	1.000	0.984	0.726	1.000	0.974	0.672

Note: Cat represents the level of response category.

Table 4.3: Power for the DINO processing function: ACDM-generated data

N	Quality	K_{jh}^*	Wald [Itemwise]			Wald [Incomplete]			Wald [Complete]			Two-step LRT			LRT		
			Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3
1000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	0.999	1.000	1.000	0.998	1.000	1.000	0.995	1.000	1.000	1.000	1.000	1.000	1.000
	Moderate	2	1.000	0.998	0.880	1.000	0.997	0.866	1.000	0.997	0.864	1.000	0.998	0.864	1.000	0.998	0.866
		3	1.000	0.992	0.871	1.000	0.989	0.859	1.000	0.988	0.854	1.000	0.990	0.844	1.000	0.990	0.846
	Low	2	0.759	0.456	0.240	0.746	0.431	0.227	0.738	0.421	0.224	0.765	0.456	0.227	0.746	0.442	0.225
		3	0.611	0.326	0.255	0.600	0.307	0.240	0.612	0.306	0.228	0.624	0.312	0.184	0.626	0.314	0.180
2000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Moderate	2	1.000	1.000	0.997	1.000	1.000	0.997	1.000	1.000	0.997	1.000	1.000	0.997	1.000	1.000	0.997
		3	1.000	1.000	0.996	1.000	1.000	0.996	1.000	1.000	0.995	1.000	1.000	0.996	1.000	1.000	0.996
	Low	2	0.976	0.773	0.479	0.975	0.765	0.464	0.974	0.762	0.461	0.981	0.782	0.488	0.970	0.762	0.457
		3	0.956	0.696	0.388	0.954	0.688	0.379	0.954	0.684	0.376	0.962	0.717	0.394	0.953	0.675	0.352
4000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Moderate	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Low	2	1.000	0.978	0.778	1.000	0.978	0.770	1.000	0.977	0.770	1.000	0.982	0.798	1.000	0.976	0.770
		3	1.000	0.970	0.780	1.000	0.968	0.778	1.000	0.968	0.778	1.000	0.978	0.811	1.000	0.966	0.766

Note: Cat represents the level of response category.

When the generating processing function was the A-CDM, as shown in Tables 4.2 and 4.3, the Wald test using the itemwise information had higher power rates than that using the incomplete information, both of which had higher power rates than that using the complete information, with only two exceptions occurring under low quality items, $N = 1000$ and $K^*_{jh} = 3$. In addition, the two-step LR test tended to have higher powers than the LR test when $N = 2000$ or above. Overall, no one method outperformed others consistently when the processing function was the A-CDM, and the same can be said for the DINA or DINO generated processing functions. However, it is worth emphasizing that when the generating processing function is the DINO model, the Wald test regardless of the information matrix produced much lower power rates than the LR test and two-step LR test when the sample size was small, $K^*_{jh} = 3$ and the category level was 3. For example, the power rates of the Wald test using different information matrices for the A-CDM ranged from 0.634 to 0.638 when $N = 1000$, $K^*_{jh} = 3$ and the category level was 3. In contrast, under the same condition, the power rates of the LR and two-step LR tests were 0.93.

4.5 Real Data Analysis

Responses of 1328 students from the United States to 17 items from the block 4 of the Trends in International Mathematics and Science Study (TIMSS) 2007 eigth-grade mathematics assessment were analyzed in this study. The attributes measured by these items were identified by L. Ma (2014), who considered both cognitive process attributes and content attributes, and built attributes at two levels. However, for illustration purposes, only seven second level content attributes were considered in this study, namely (α_1) whole numbers and integers, (α_2) fractions, decimals, ratio proportion, and percent, (α_3) algebraic expressions and equations/formulas functions, (α_4) geometric shapes, (α_5) geometric measurement and location movement, (α_6) data organization and representation, and (α_7) data interpretation and chance. L. Ma (2014) also developed the Q-matrix for these items using multiple regression and the least squares distance method, and the Q_C-matrix, given in Table 4.4, was created based on L. Ma's (2014) work by assuming that for each polytomously scored item, all required attributes are measured by each step of the item. The sequential G-DINA model was fitted to the data. The Wald test using the itemwise, incomplete and complete information matrices, the LR test and the two-step LR test were conducted to examine whether the saturated G-DINA model can be replaced by the DINA model, DINO model and *A*-CDM.

Table 4.4: Q_C-matrix for the TIMSS 2007 data

Item No.	TIMSS Item ID	α_1	α_2	α_3	α_4	α_5	α_6	α_7
1	M042001	1	0	0	0	0	0	0
2	M042022	1	0	0	0	0	0	0
3	M042082	1	0	1	0	0	0	0
4	M042088	0	0	1	0	0	0	0
5	M042304A	1	0	0	0	0	0	0
6-1	M042304B-1	0	1	1	0	0	0	0
6-2	M042304B-2	0	1	1	0	0	0	0
7	M042304C	0	1	1	0	0	0	0
8-1	M042304D-1	1	0	0	0	0	0	0
8-2	M042304D-2	1	0	0	0	0	0	0
9	M042267	1	0	1	0	0	0	0
10	M042239	1	0	1	0	0	0	0
11	M042238	1	0	1	0	1	0	0
12	M042279	0	0	0	1	0	0	0
13	M042036	1	0	0	1	1	0	0
14	M042130	1	0	0	1	1	1	0
15	M042303A	0	1	0	0	0	1	0
16-1	M042303B-1	1	0	0	0	1	1	1
16-2	M042303B-2	1	0	0	0	1	1	1
17	M042222	0	1	0	0	0	0	1

For 13 multi-attribute response categories from nine dichotomous items and two polytomous items, the DINA model was never selected by any of the hypothesis tests. The DINO model was selected only once for Item 17 by the Wald test using the complete information matrix with a p-value of 0.51; whereas for this item, *A*-CDM was believed appropriate by all hypothesis tests. Because the DINA and DINO models were not selected for any other items, only the p-values for A-CDM were given in Table 4.5. The results from all hypothesis tests were consistent for 8 item response categories. Specifically, the G-DINA model was deemed appropriate for Items 3, 6-1, 10 and 11, whereas the A-CDM was considered as good as the G-DINA model for Item 7, 9, 13 and 17. The Wald test using the itemwise information and the two-step LRT reached the same conclusion, but all other methods identified more item response categories where A-CDM can be used in place of the G-DINA model. Based on the Wald test using the complete information matrix, the A-CDM was appropriate for nine item response categories.

Table 4.5: P-values of the Wald and LR tests for the A-CDM processing function

Item No.	Wald [Itemwise]	Wald [Incomplete]	Wald [Complete]	Two-step LRT	LRT
3					
6-1					
6-2			0.083		
7	0.205	0.282	0.481	0.140	0.684
9	0.438	0.501	0.543	0.440	0.543
10					
11					
13	0.544	0.635	0.723	0.304	0.570
14			0.216		
15			0.081		
16-1		0.137	0.136		
16-2		0.092	0.343		0.275
17	1.000	1.000	1.000	1.000	1.000

Note: p-values less than 0.05 were omitted.

According to the models suggested by each hypothesis test, the data were recalibrated, and the AIC and BIC of each fitted model were calculated. The LR test was also implemented at test level to evaluate whether the suggested models were as good as the saturated sequential G-DINA model. Based on the test-level LR test, the models suggested by the Wald test using the complete information matrix were significantly worse than the sequential G-DINA model ($\chi^2 = 80.417, df = 42, p < 0.001$), whereas the models suggested by other hypothesis tests were all statistically as good as the sequential G-DINA model. From Table 4.6, if not taking the Wald test using the complete information into consideration, the models based on the Wald test using the incomplete information had the smallest AIC and BIC, followed by those suggested by the LR test.

Table 4.6: AIC and BIC for models selected by the Wald and LR tests

	Wald [Itemwise]	Wald [Incomplete]	Wald [Complete]	Two-step LRT	LRT
AIC	27431.159	**27419.739**	27438.274	27431.159	27420.636
BIC	28625.188	28499.556	**28450.602**	28625.188	28557.559

Note: Lowest AIC and BIC were shown in boldface. AIC and BIC for the sequential G-DINA model were 27441.857 and 28672.226, respectively.

Discussion

It has been said that no model is true, but some are more useful than others. A psychometric model should be in line with the underlying cognitive processes to provide a good approximation to the reality. The condensation rule is a central component for many cognitive diagnosis models, and in this study, we examined the Type I error and power of the Wald and likelihood ratio tests in determining the appropriate condensation rules for each response category of a polytomously scored item. This is achieved by comparing whether the reduced models can be used in place of the G-DINA model without a significant loss in model-data fit.

Previous research (e.g., de la Torre & Lee, 2013; W. Ma et al., 2016) on dichotomous responses has revealed that the Type I error of the Wald test using the itemwise information matrix can be inflated under certain conditions. Similar findings have been observed in this study for polytomous response data. Furthermore, this study has shown that using the complete and incomplete information could control the inflated type I error for the Wald test, but only to some extent under unfavorable test conditions. A potential issue associated with the use of the complete and incomplete information matrices is that the the resulting Wald test tended to be conservative when items were of high quality and sample size was small. This finding is partially consistent with Liu, Xin, Li, Tian, and Liu (2016), who examined the type I error of the Wald test for the DINA model using the complete information matrix in detecting differential item functioning for dichotomous responses and found that the Wald test tended to be conservative under small sample sizes regardless of item quality.

Despite not involving an estimated variance-covariance matrix, the LR test also produced inflated Type I error under some unfavorable conditions, similar to the Wald test using the complete or incomplete information matrix. However, unlike the Wald test, the LR test did not tend to be conservative under high item quality conditions. Although the two-step LR test performed as well as the LR test under most conditions, it can yield much more inflated Type I error than the LR test when items were of low quality and sample size was small.

In terms of the computation time, the LR test can be very expensive if the data calibration takes time or the number of categories is very large. For the real data analyzed in this study, the Wald test took only about 0.25 seconds to compare the G-DINA processing function with the DINA, DINO and *A*-CDM for all multi-attributes categories. In contrast, the LR test and two-step LR test took around 16 minutes and 12 seconds, respectively. It should be noted that the code for the LR test was written in R by the author, and faster speeds can be expected by using a program written in a lower-level language such as C.

Under most conditions, the studied methods have similar power rates, and no single method performed best across all conditions. Despite excellent power rates under favorable conditions, their power can drop substantially under unfavorable conditions. This study also found that the power rate decreased as the category level increased. Based on the closed-form solution for item parameter estimation, the processing function for category h is the ratio of the expected number of examinees given a particular attribute pattern obtaining a score of h or higher to the expected number of examinees given the attribute pattern obtaining a score of $h-1$ or higher. As a result, the number of examinees who get at least a score of $h-1$ can be viewed as the "effective" sample size for category h and thus for a higher category, the "effective" sample size is smaller, yielding a poorer power.

Although the Wald test and the LR test may be used for model selection at category level for polytomously scored items, the findings of this study should be used with cautions for several reasons. First, the test length and the number of attributes were fixed and the Q-matrix was assumed known. Guo, Ma, and de la Torre (2017) found that with misspecified Q-matrix, the standard errors of item parameters estimated using the outer product of gradient can be problematic, which may further influence the performance of the Wald test. It would be important to explore the performance of the Wald test using variance-covariance matrix calculated in other ways, such as the observed information matrix (Louis, 1982), the supplemented EM (Meng & Rubin, 1991) and the numerical differential methods (Jamshidian & Jennrich, 2000). Also, all required attributes of the additive model were assumed to contribute equally to the processing functions, which could be relaxed in the future studies.

4.6 References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*, 123–140.

de la Torre, J. (2008). 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., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*, 355–373.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). A review of cognitively diagnostic assessment and a summary of psychometric models. In R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics* (Vol. 26, pp. 979–1030). Amsterdam, Netherlands: Elsevier.

Guo, W., Ma, W., & de la Torre, J. (2017). *Standard error estimation using bootstrap approaches for cognitive diagnosis models.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Antonio, TX.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321.

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

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

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*, 98–125.

Jamshidian, M., & Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*, 257–270.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59–81.

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*, 144–177.

Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics*, *41*, 3–26.

Liu, Y., Xin, T., Li, L., Tian, W., & Liu, X. (2016). An improved method for differential item functioning detection in cognitive diagnosis models: An application of Wald statistic based on observed information matrix. *Acta Psychologica Sinica*, *48*(5), 588–598.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *44*, 226–233.

Ma, L. (2014). *Validation of the item-attribute matrix in TIMSS: Mathematics using multiple regression and the LSDM* (Unpublished doctoral dissertation). University of Denver.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, *69*, 253–275.

Ma, W., & de la Torre, J. (2017). *GDINA: The generalized DINA model framework.* [Computer software version 1.4.2]. Retrieved from `https://CRAN.R-project.org/package=GDINA`

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*, 200–217.

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

Meng, X.-L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, *86*, 899–909.

Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2016). *On the estimation of standard errors in cognitive diagnosis models* (Working Papers). Faculty of Economics and Statistics, University of Innsbruck. Retrieved from `http://EconPapers.repec.org/RePEc:inn:wpaper:2016-25`

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rojas, G., de la Torre, J., & Olea, J. (2012). *Choosing between general and specific cognitive doagnosis models when the sample size is small.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, Vancouver, British Columbia.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement*, *67*, 239–257.

Sorrel, M. A., de la Torre, J., Abad, F. J., & Olea, J. (in press). Two-step likelihood ratio test for model comparison in cognitive diagnosis models. *Methodology*.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 583–639.

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

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

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.

von Davier, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology*, *67*, 49–71.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, *54*, 426–482.

4.7 Appendix

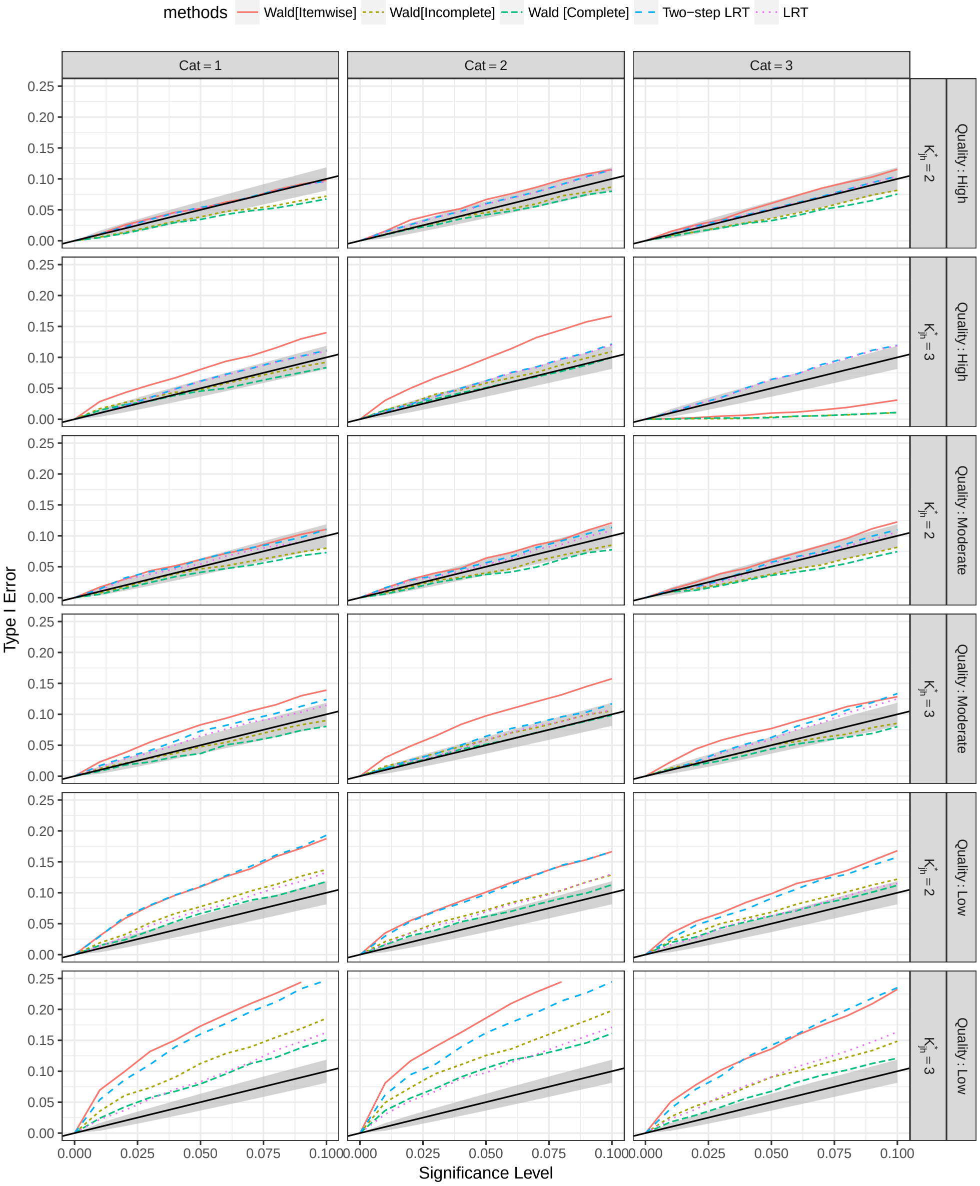


Figure 4.6: Type I error for the DINA processing function under $N = 1000$

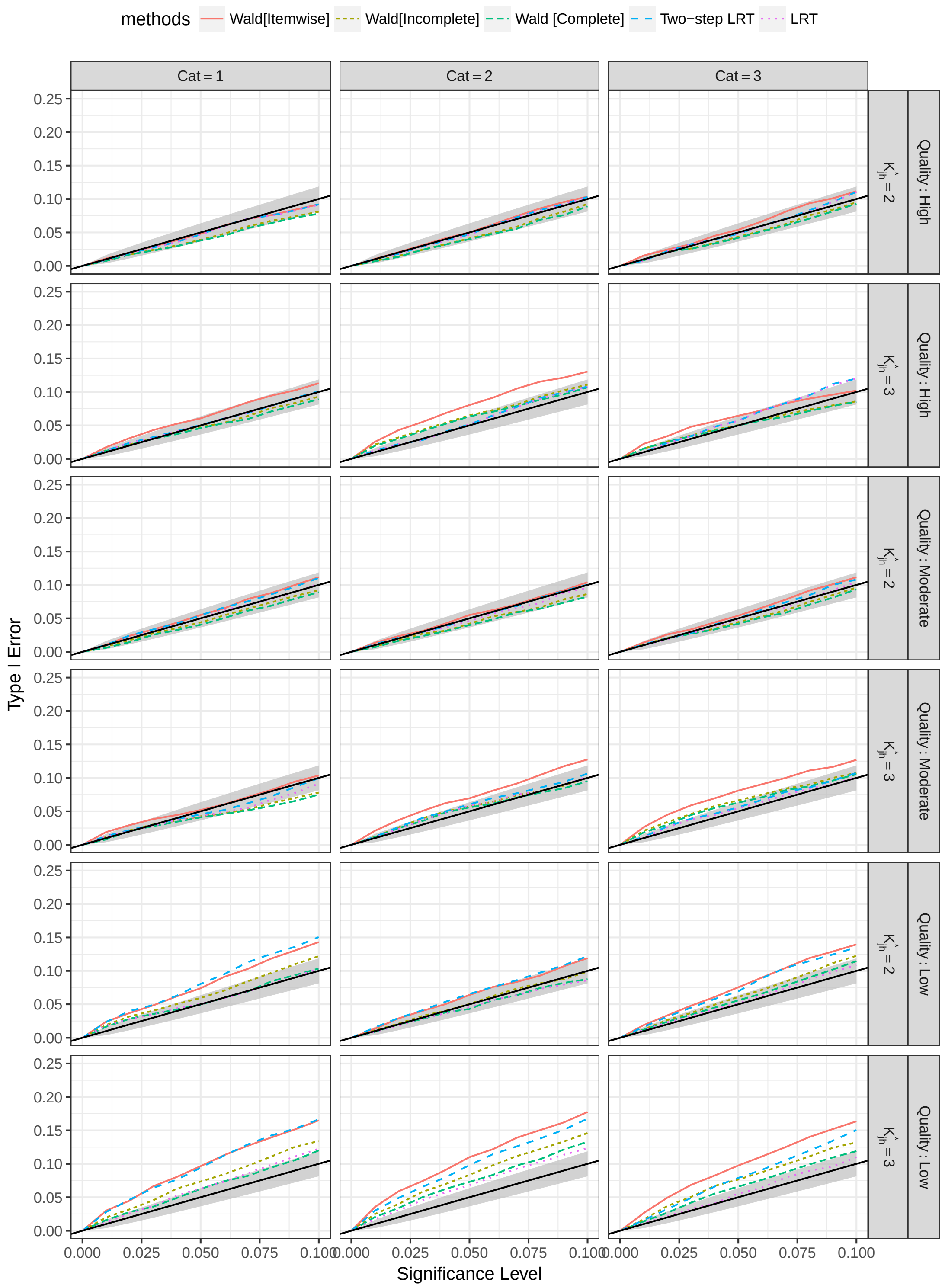


Figure 4.7: Type I error for the DINA processing function under $N = 2000$

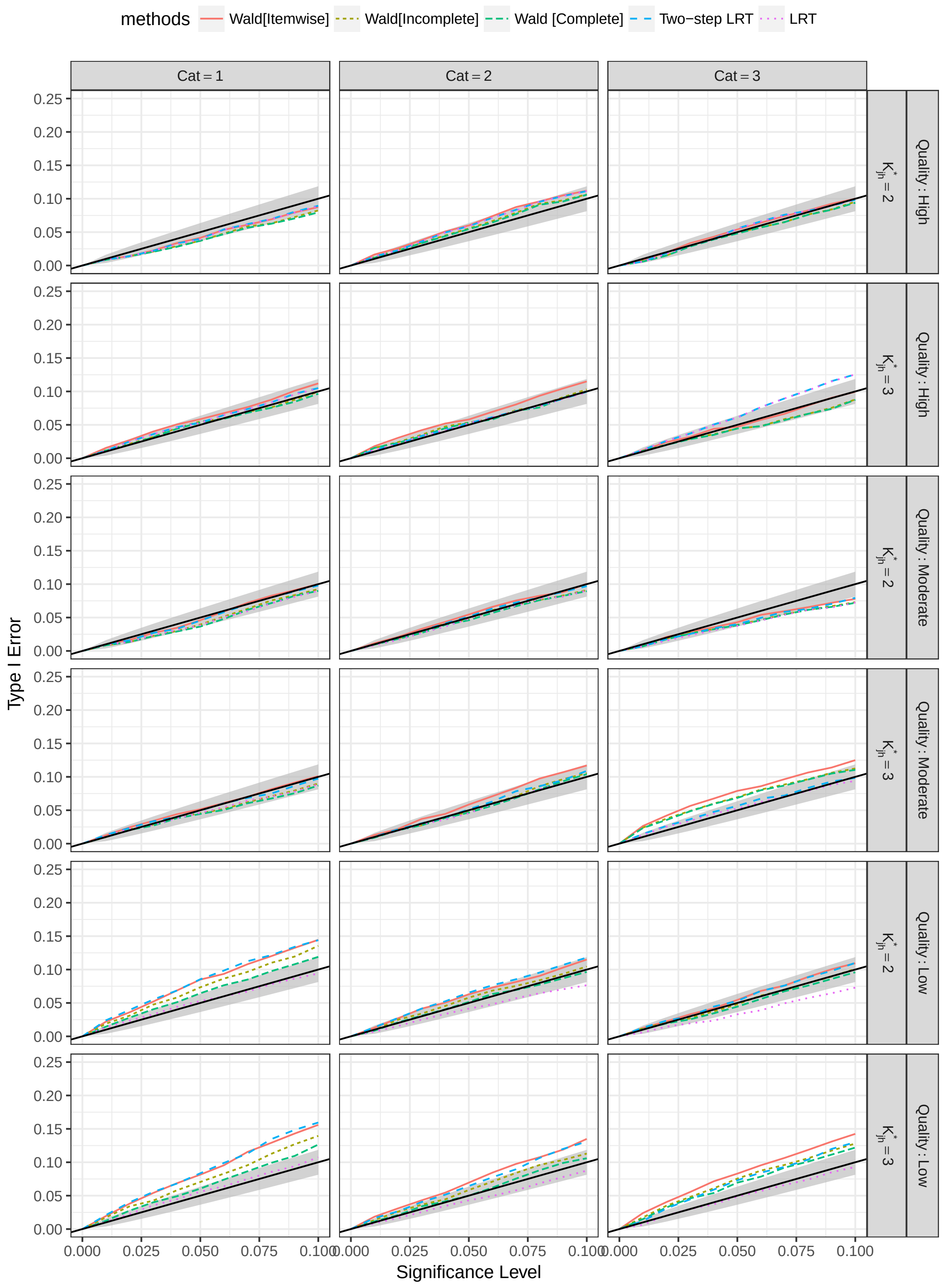


Figure 4.8: Type I error for the DINA processing function under $N = 4000$

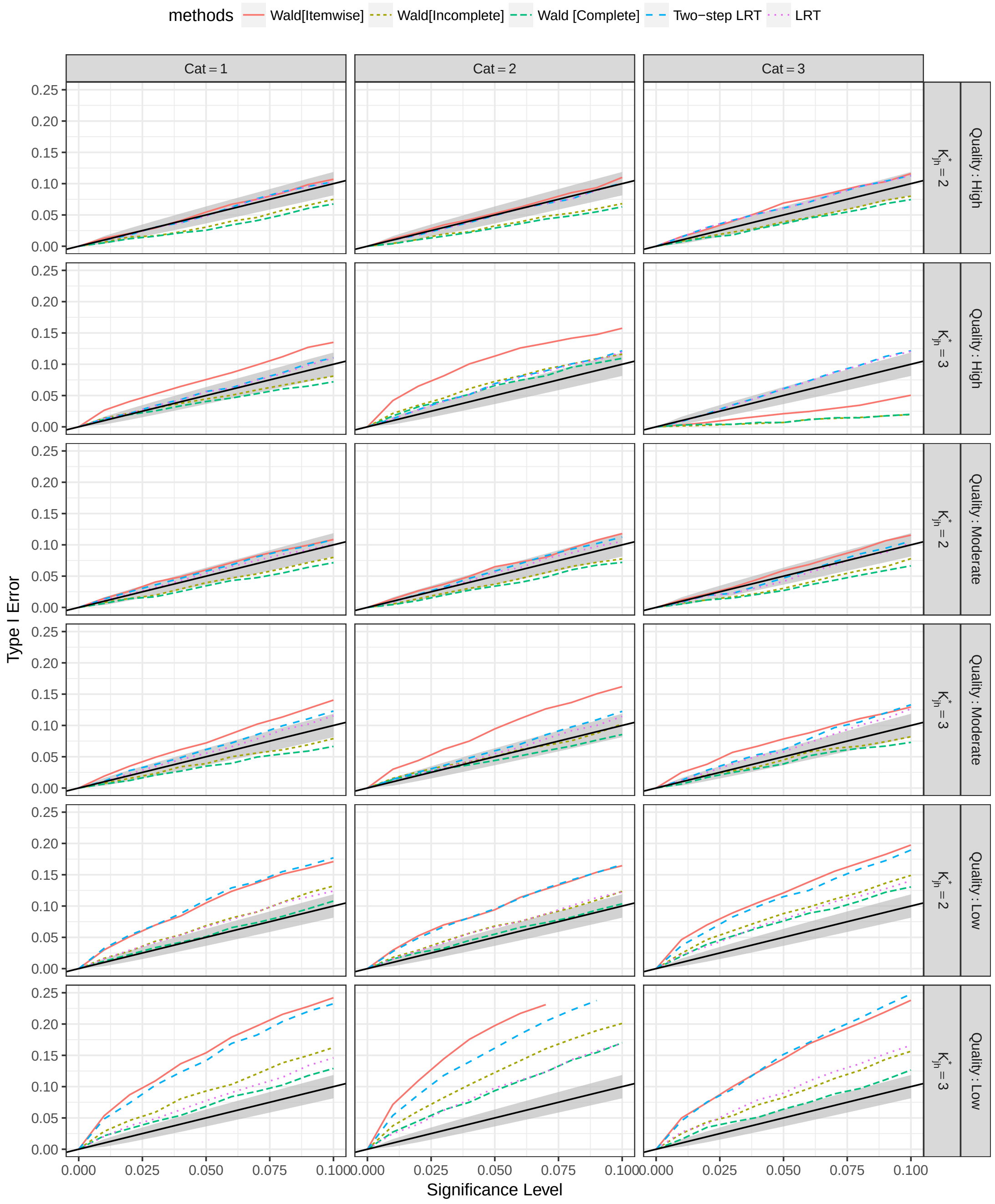


Figure 4.9: Type I error for the DINO processing function under $N = 1000$

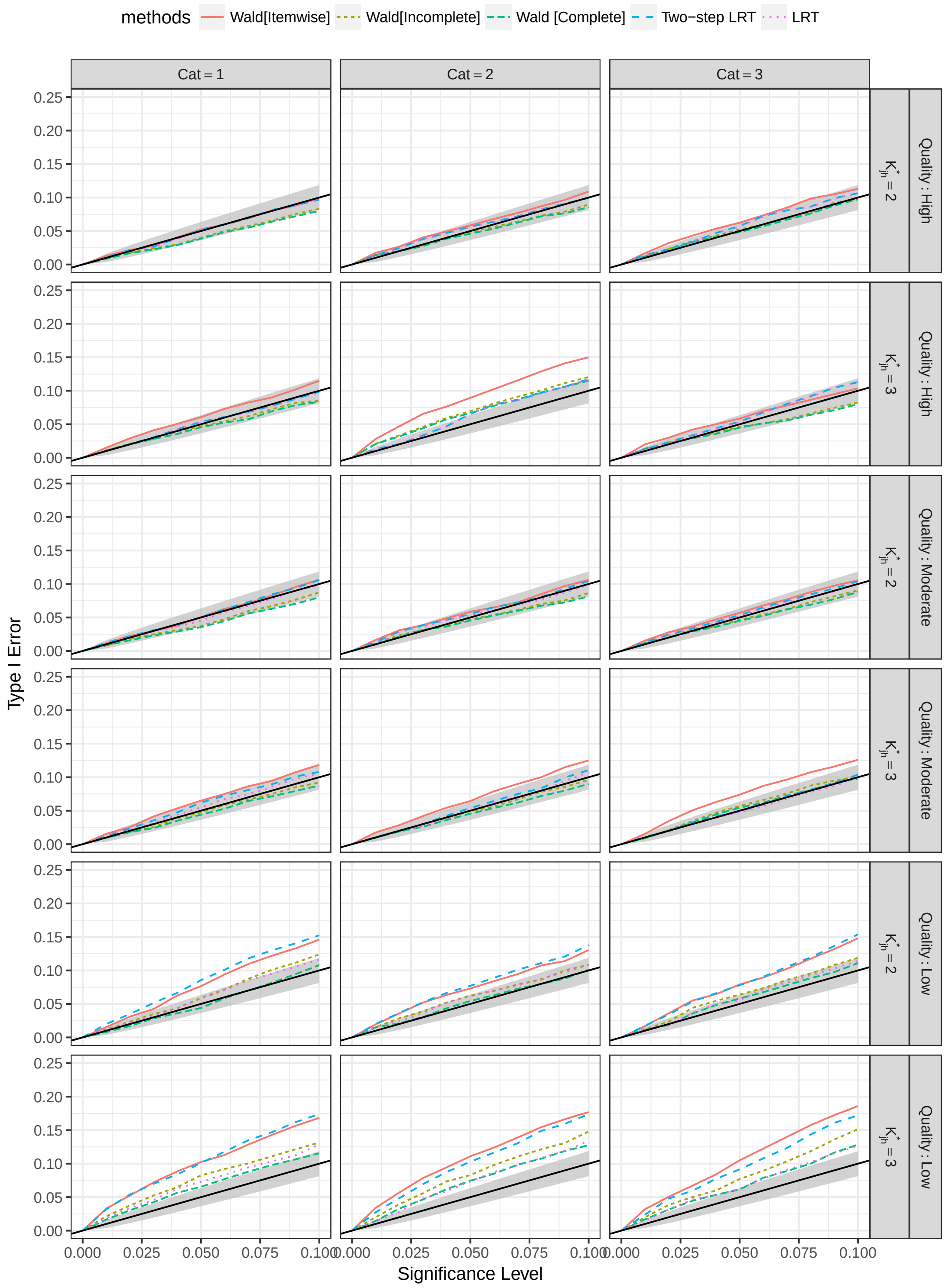


Figure 4.10: Type I error for the DINO processing function under $N = 2000$

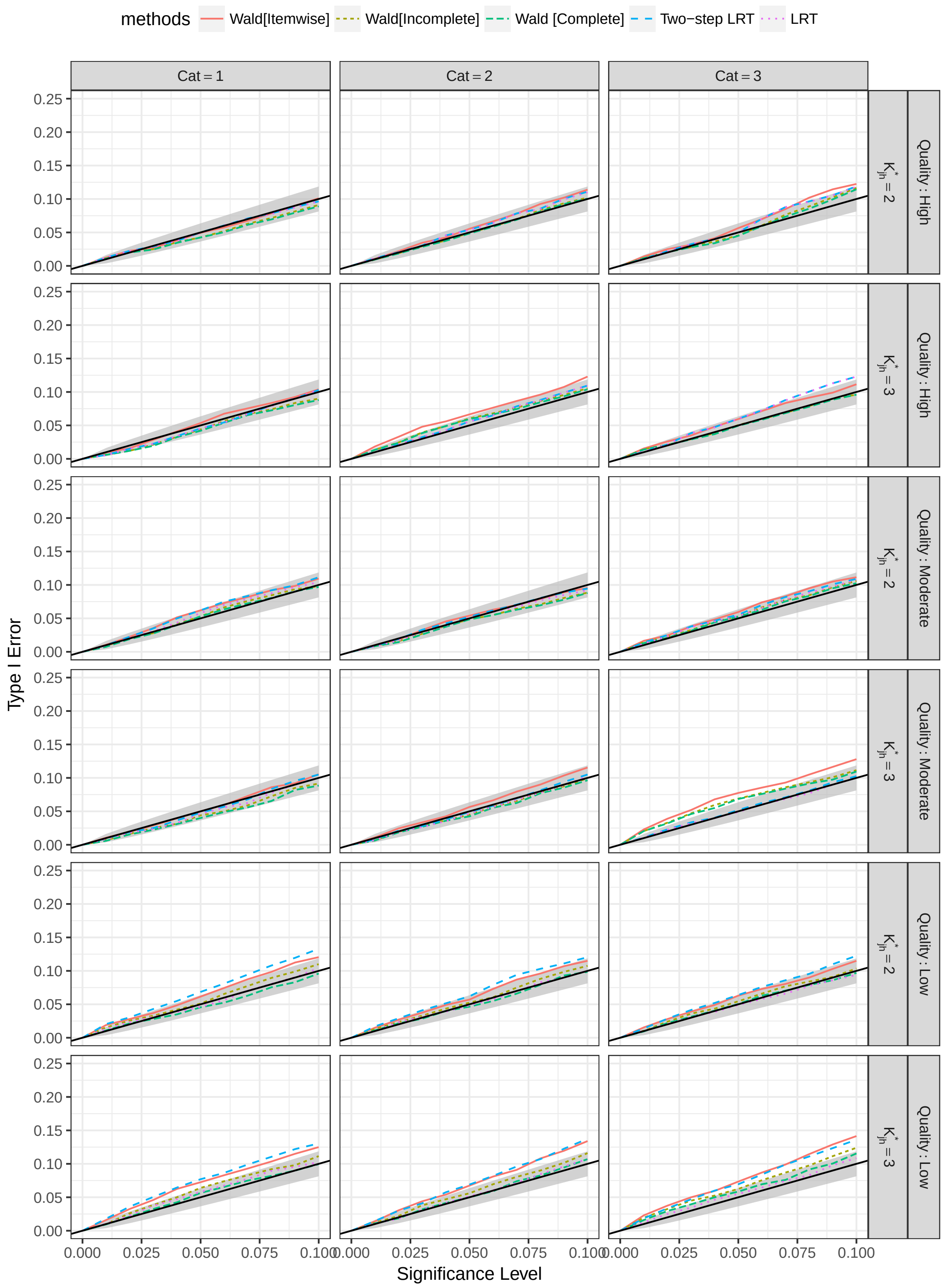


Figure 4.11: Type I error for the DINO processing function under $N = 4000$

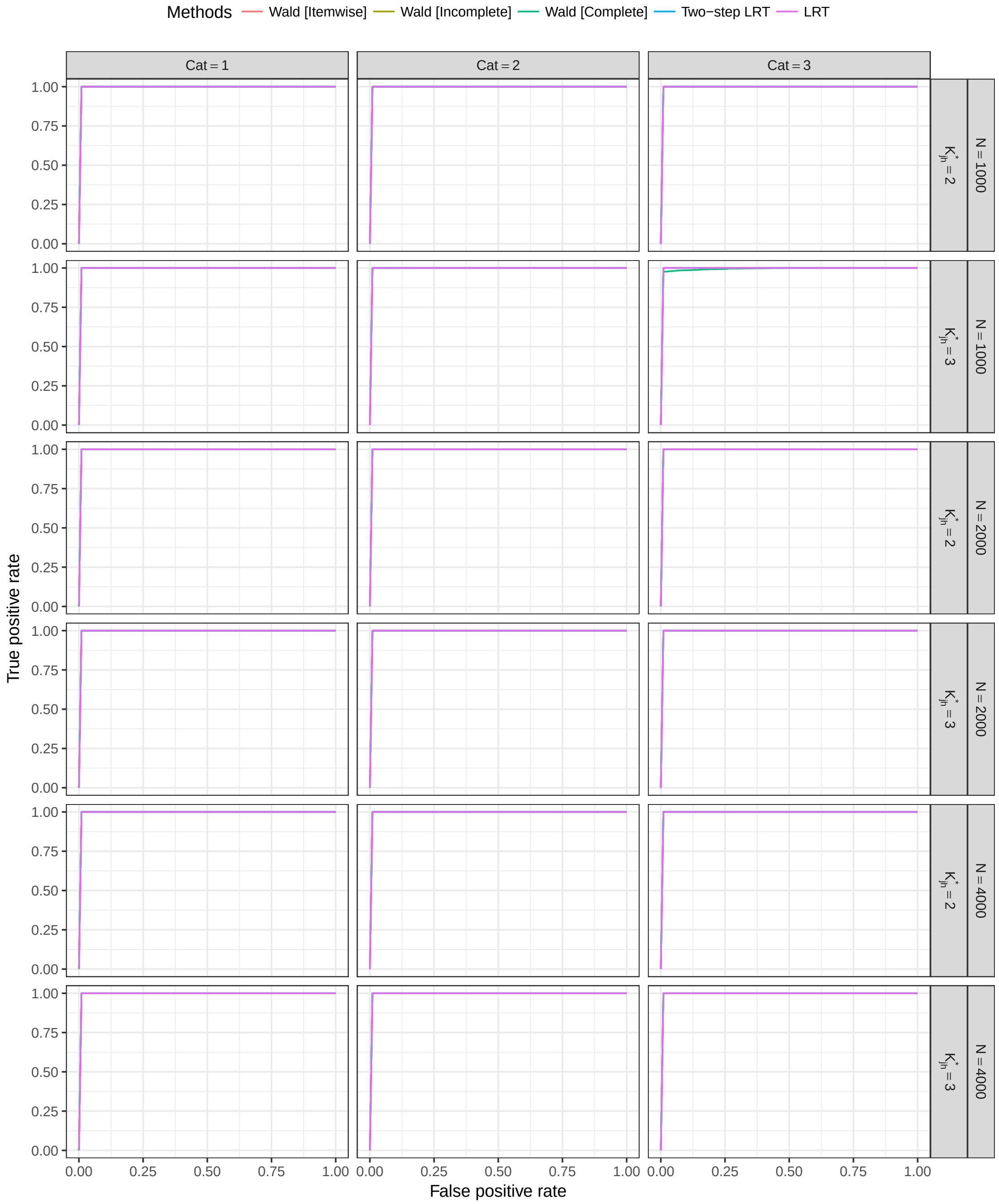


Figure 4.12: ROC curves for the DINO processing function: DINA-generated processing function under high item quality

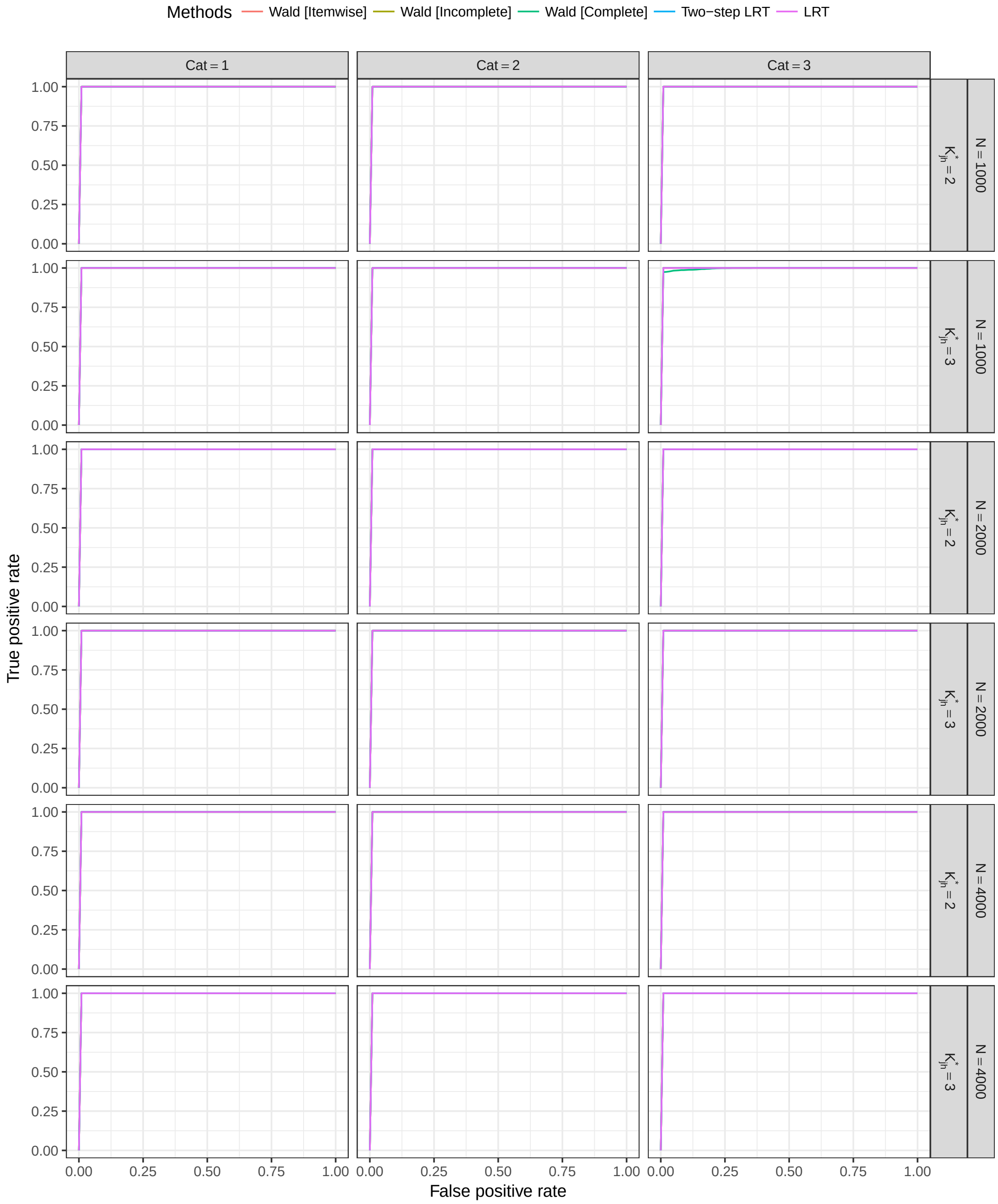


Figure 4.13: ROC curves for the *A*-CDM processing function: DINA-generated processing function under high item quality

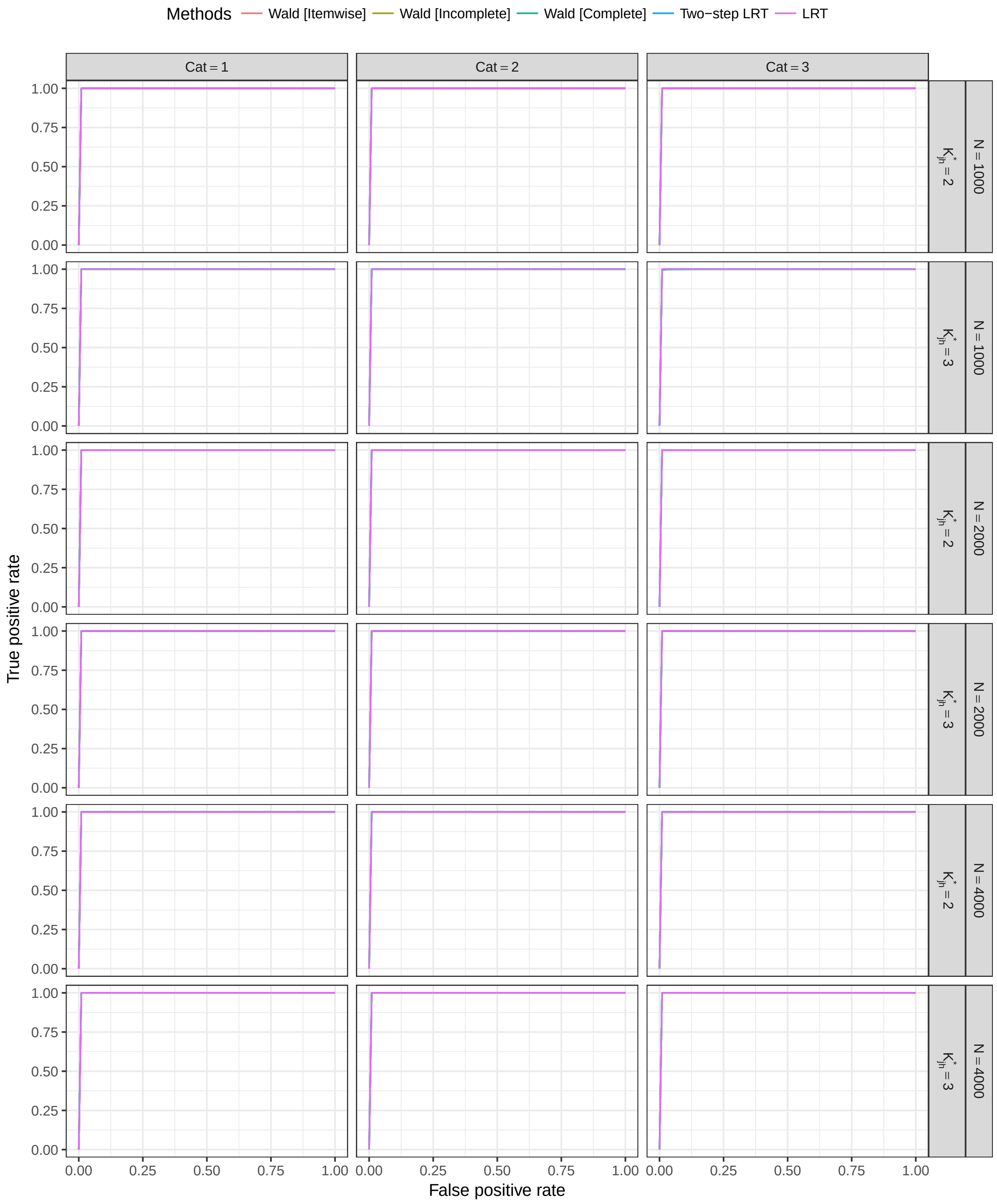


Figure 4.14: ROC curves for the DINO processing function: DINA-generated processing function under moderate item quality

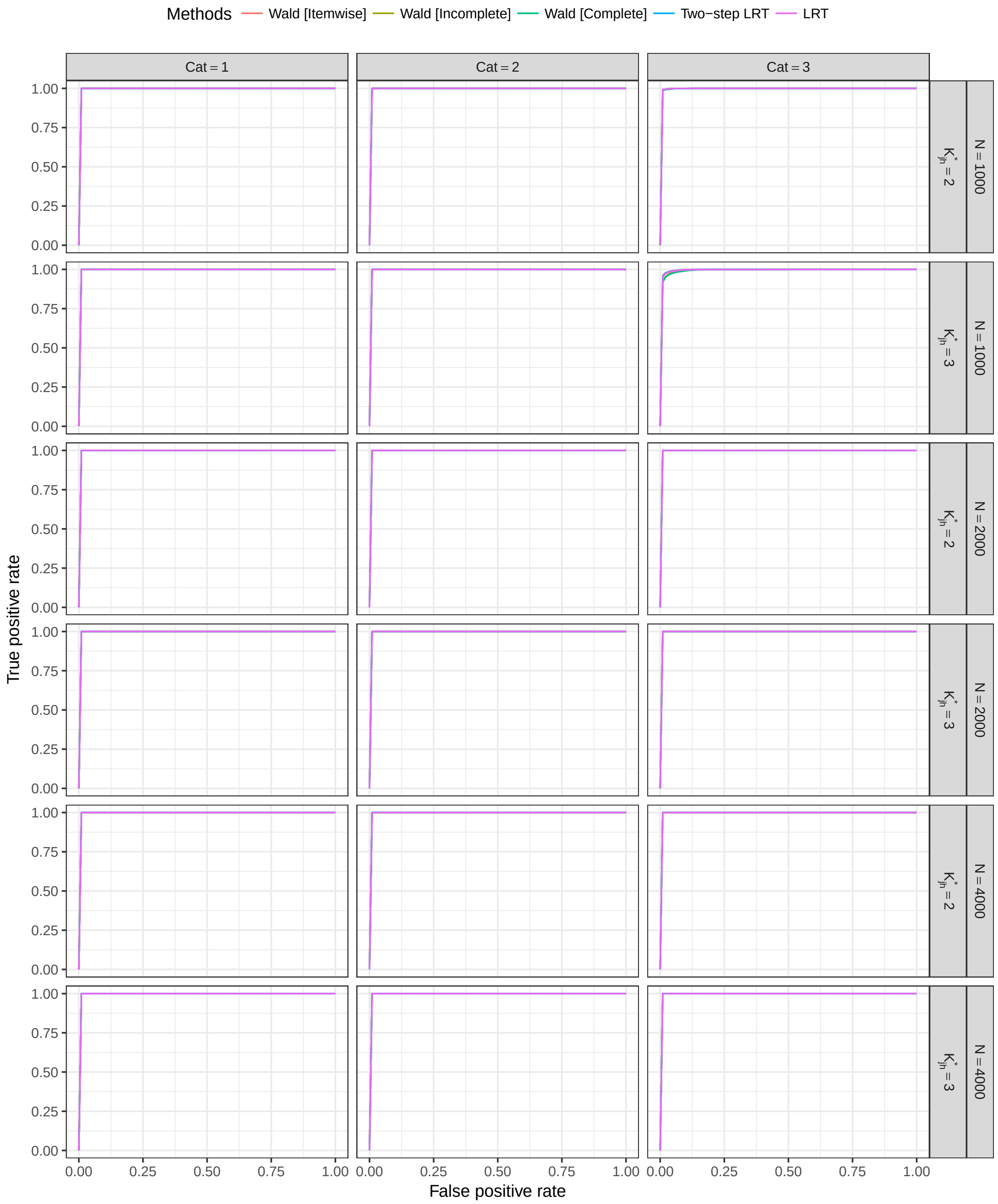


Figure 4.15: ROC curves for the *A*-CDM processing function: DINA-generated processing function under moderate item quality

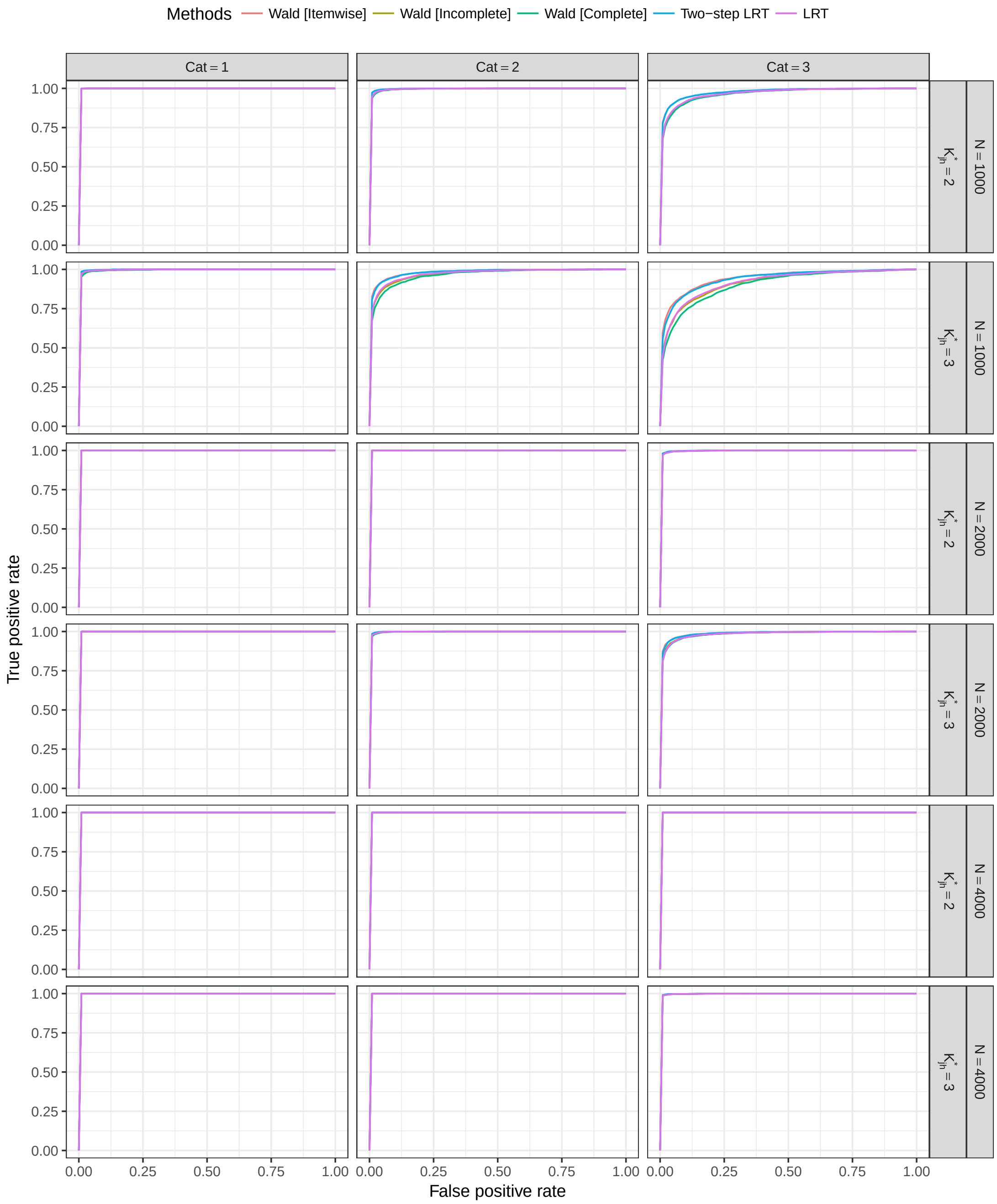


Figure 4.16: ROC curves for the DINO processing function: DINA-generated processing function under low item quality

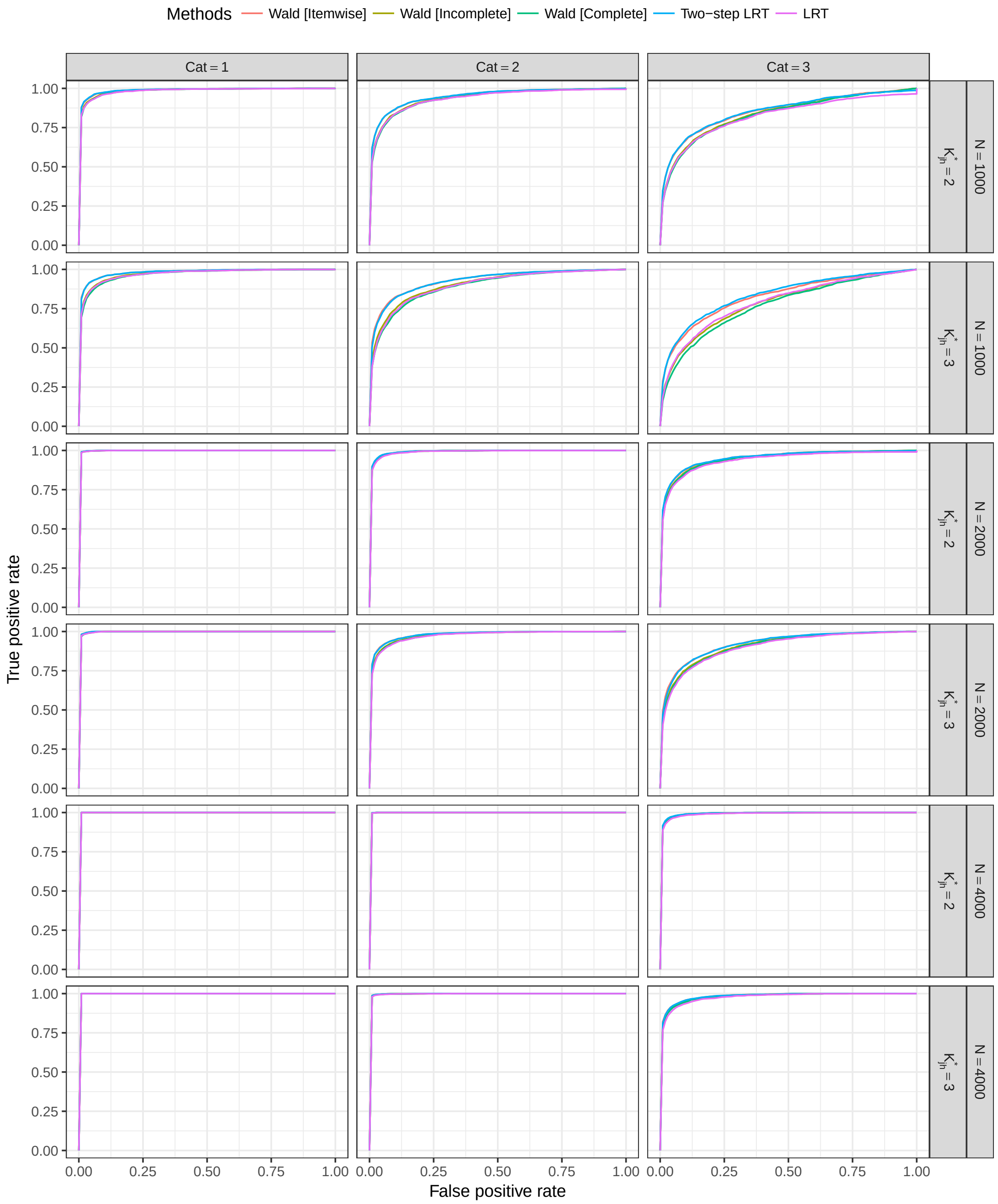


Figure 4.17: ROC curves for the *A*-CDM processing function: DINA-generated processing function under low item quality

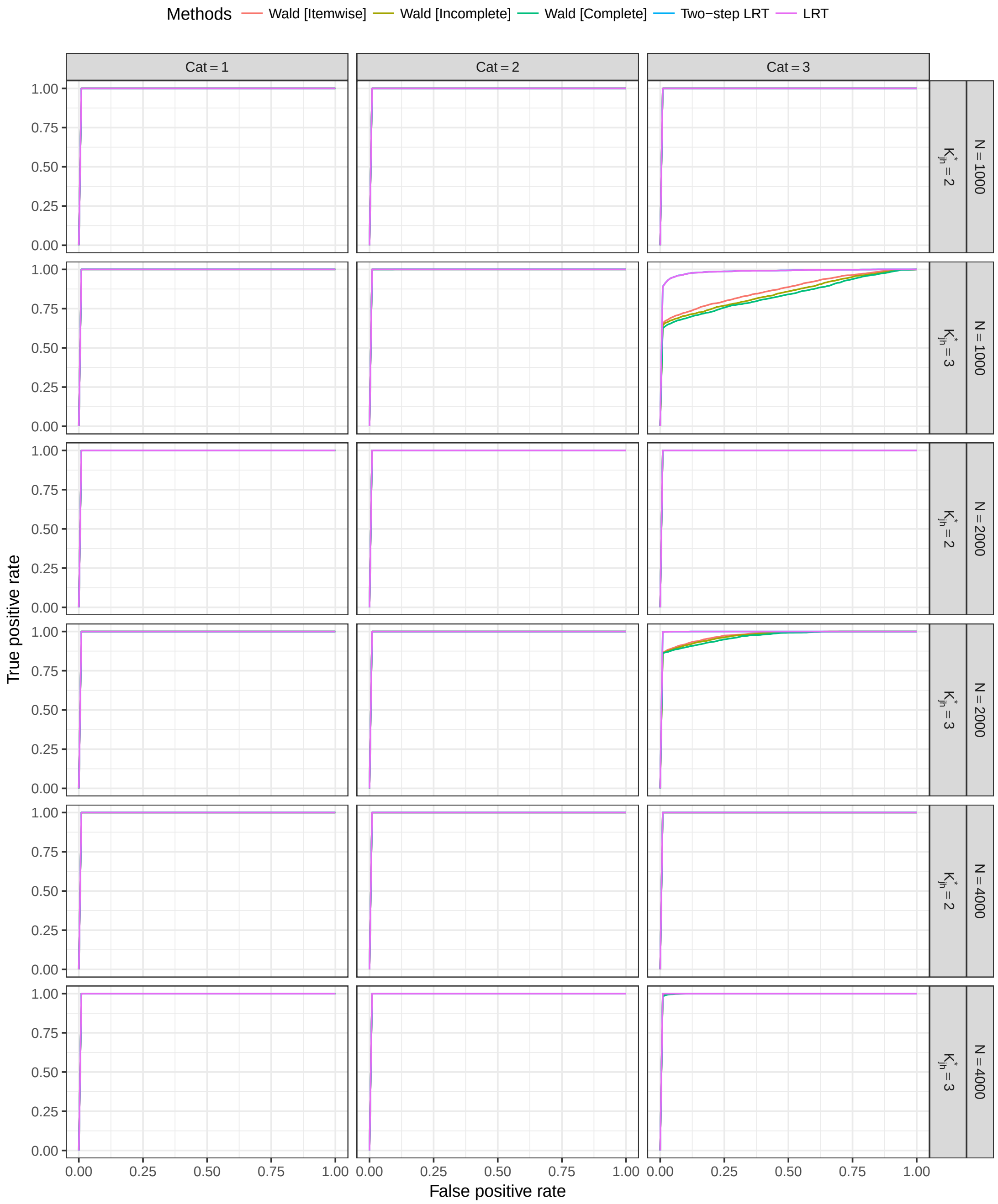


Figure 4.18: ROC curves for the DINA processing function: DINO-generated processing function under high item quality

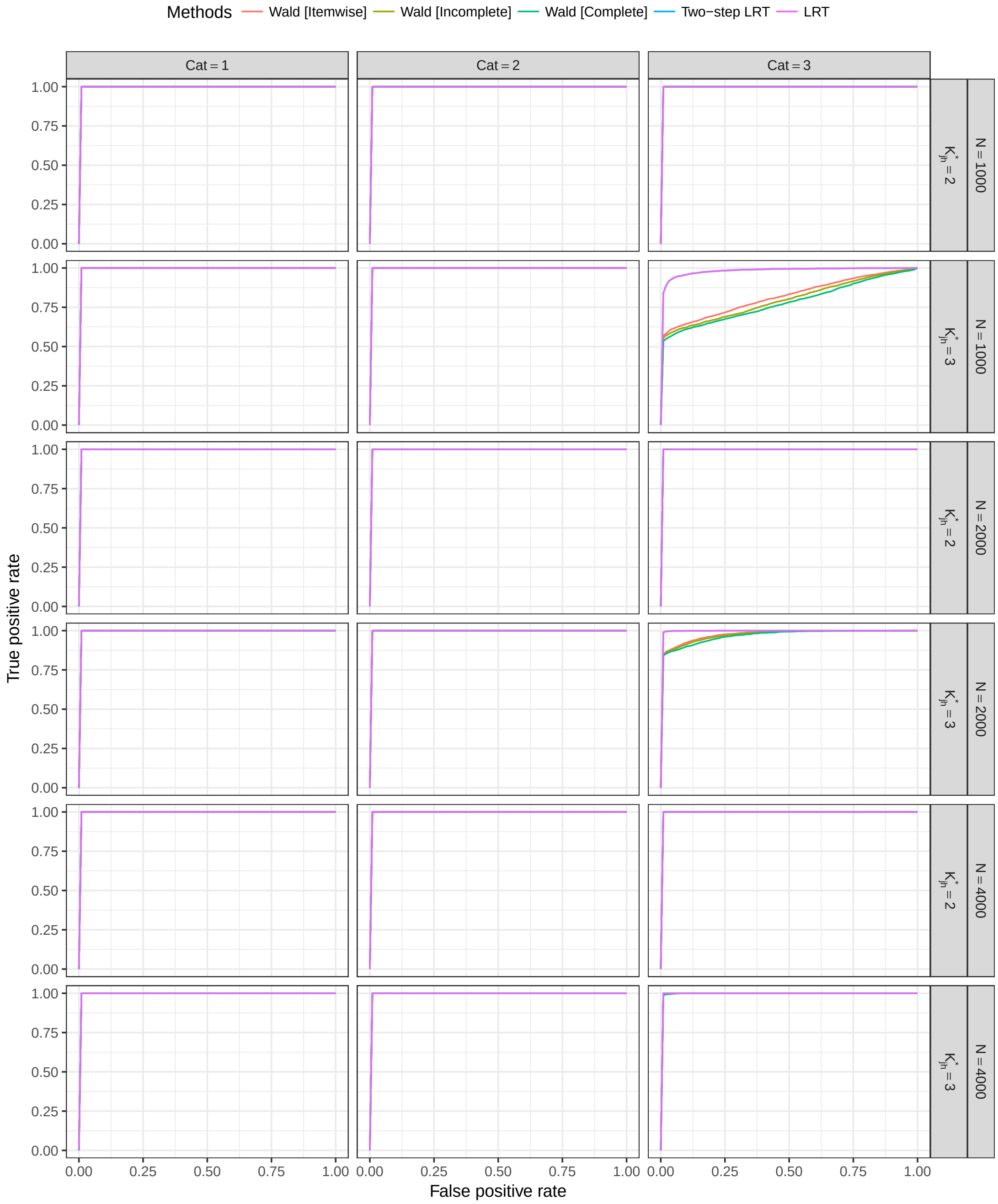


Figure 4.19: ROC curves for the *A*-CDM processing function: DINO-generated processing function under high item quality

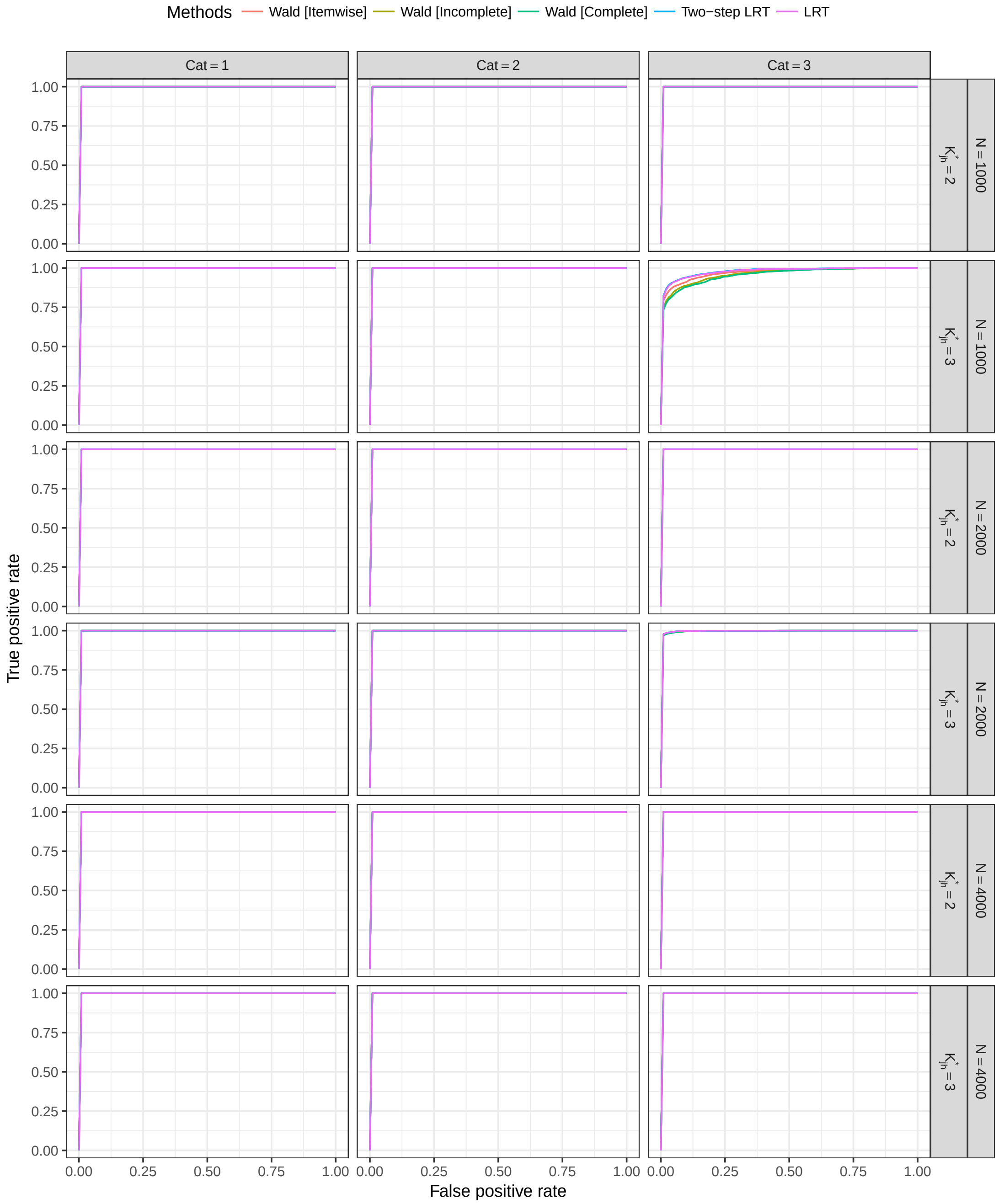


Figure 4.20: ROC curves for the DINA processing function: DINO-generated processing function under moderate item quality

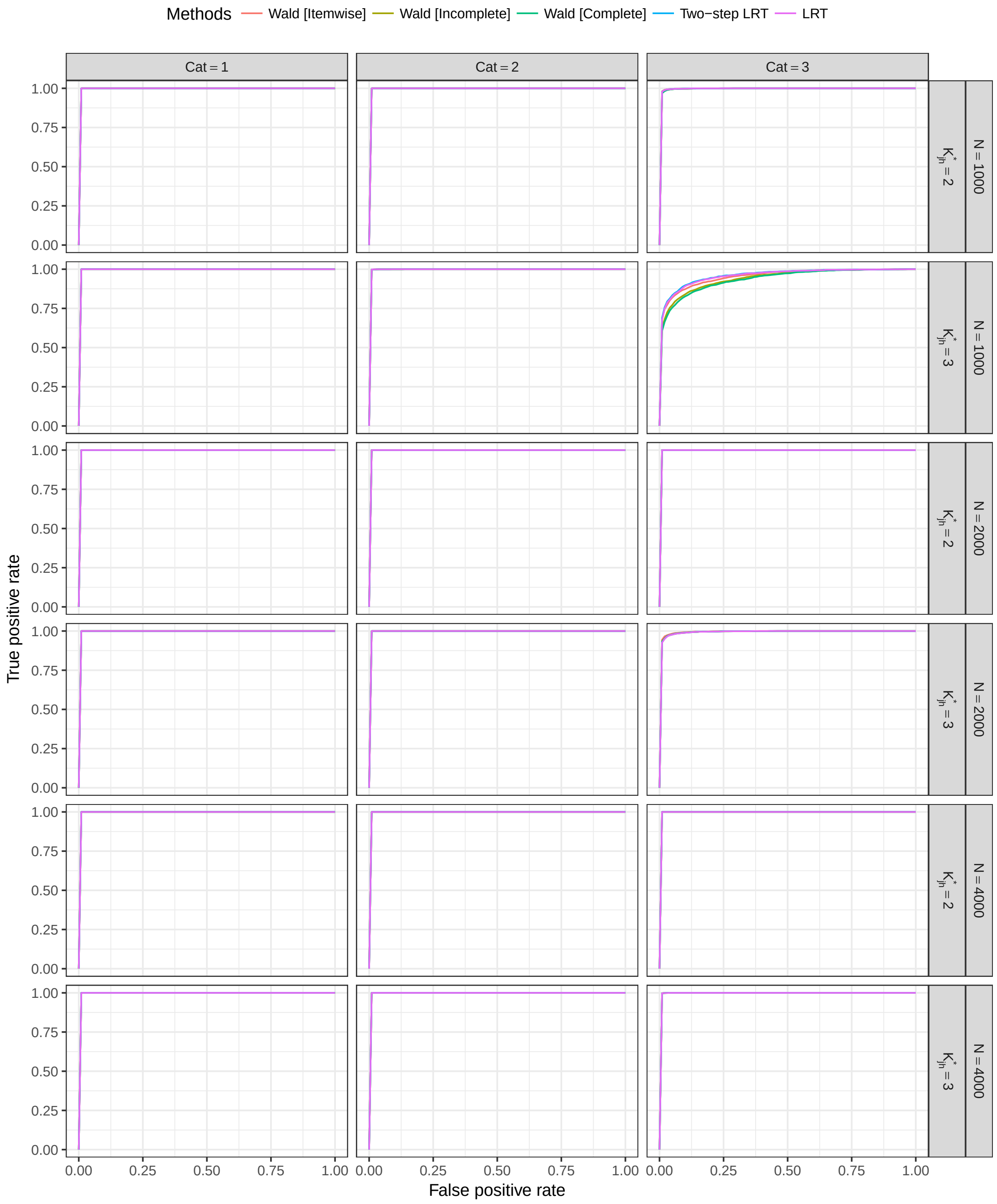


Figure 4.21: ROC curves for the *A*-CDM processing function: DINO-generated processing function under moderate item quality

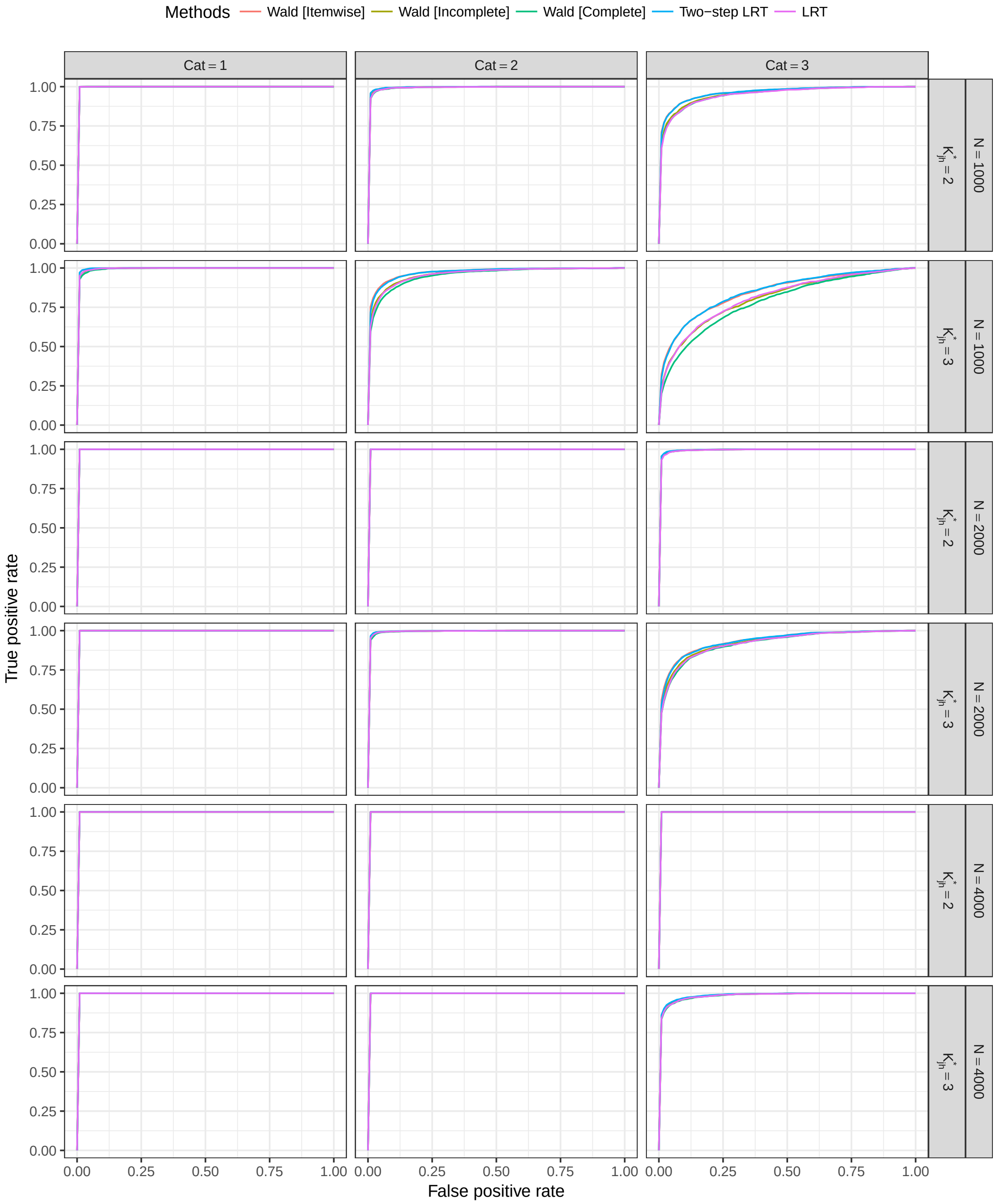


Figure 4.22: ROC curves for the DINA processing function: DINO-generated processing function under low item quality

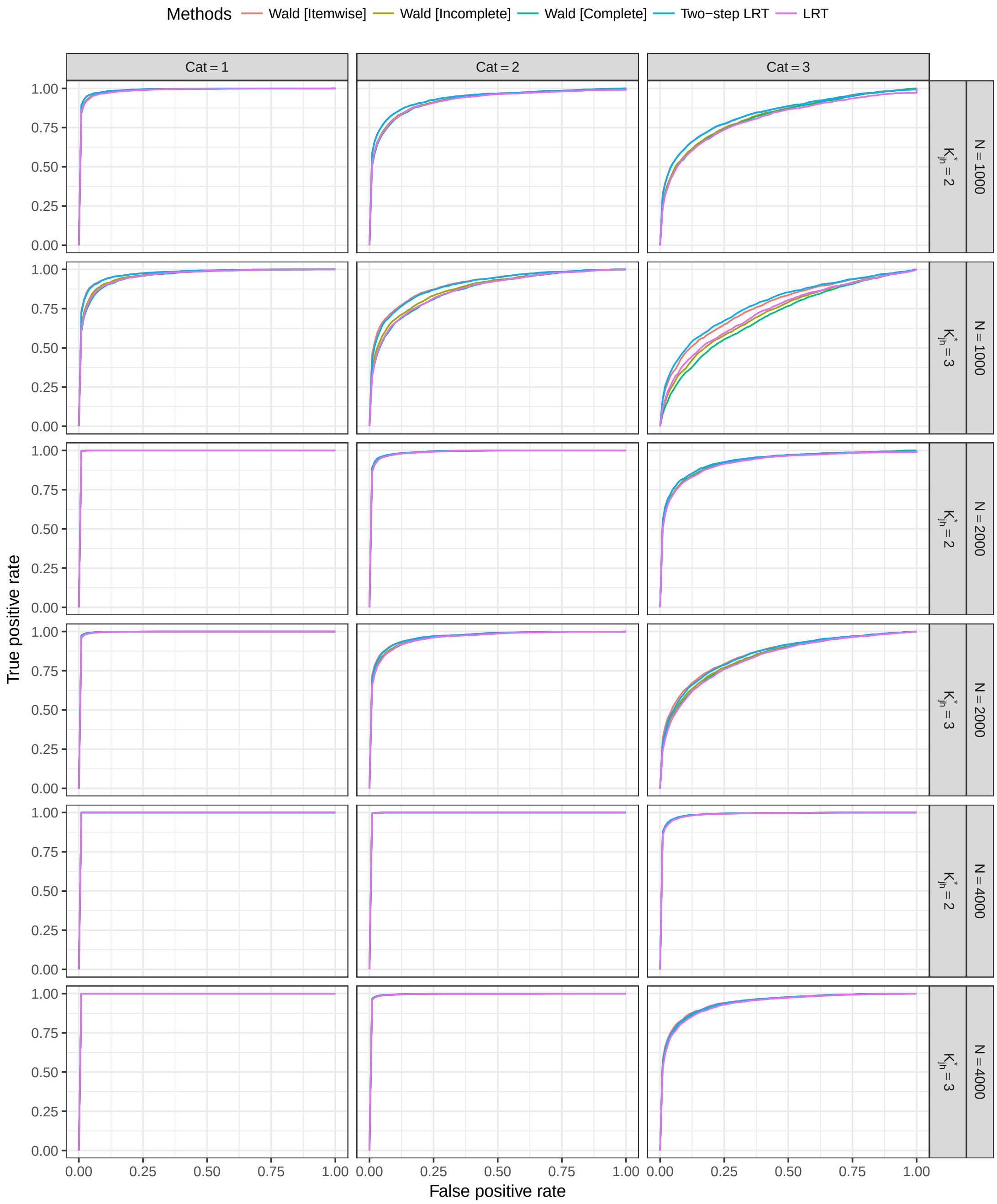


Figure 4.23: ROC curves for the *A*-CDM processing function: DINO-generated processing function under low item quality

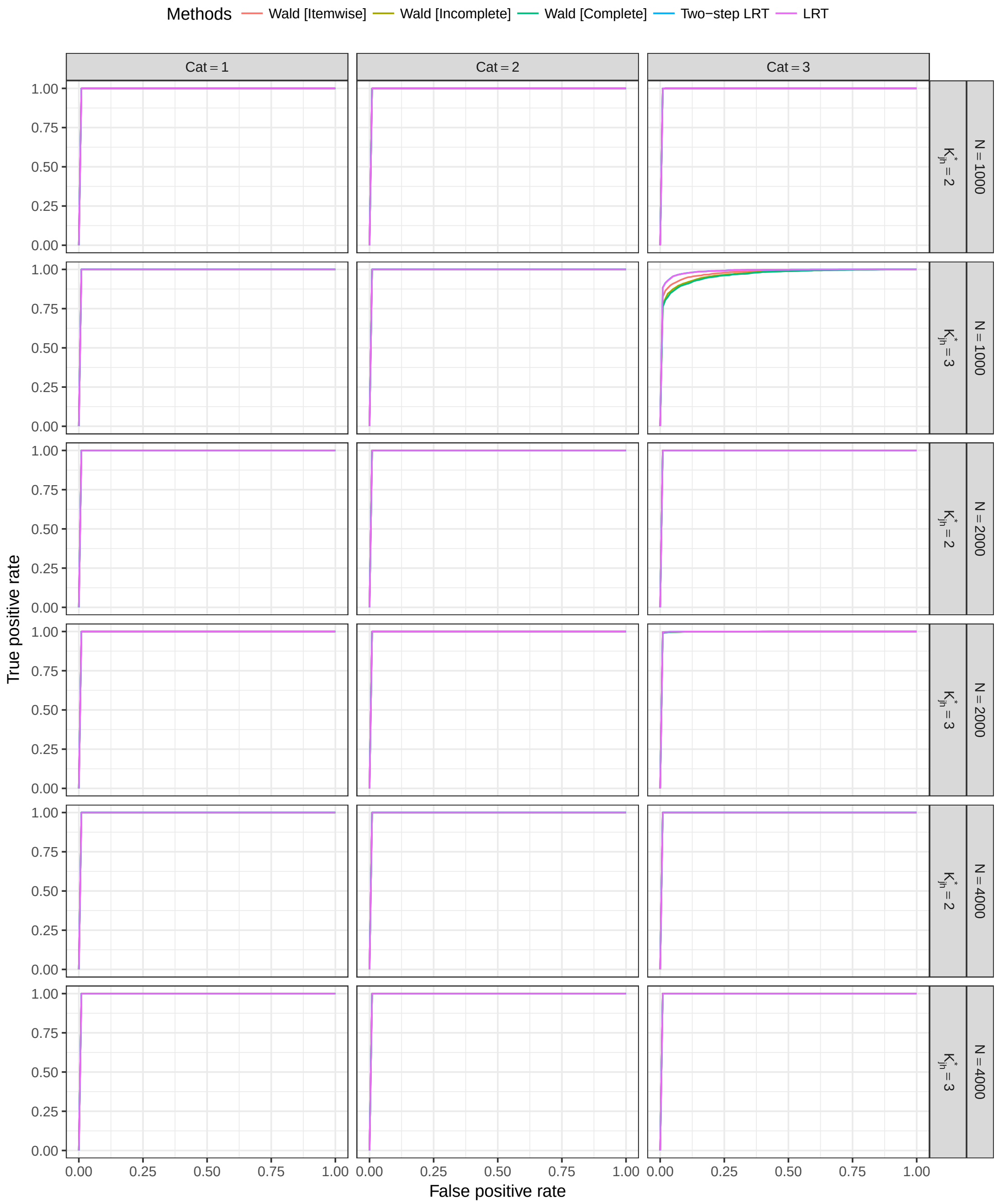


Figure 4.24: ROC curves for the DINA processing function: *A*-CDM-generated processing function under high item quality

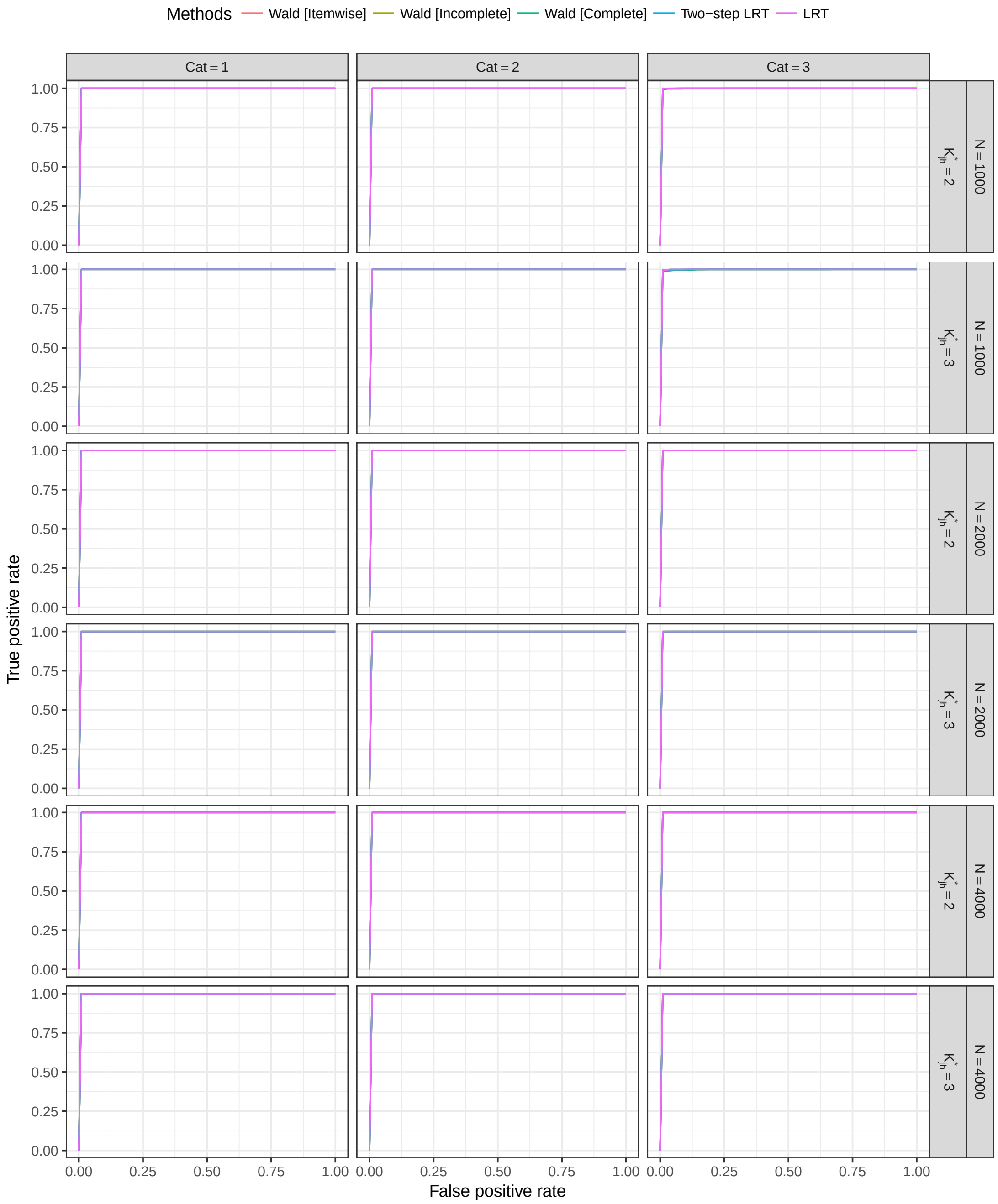


Figure 4.25: ROC curves for the DINO processing function: *A*-CDM-generated processing function under high item quality

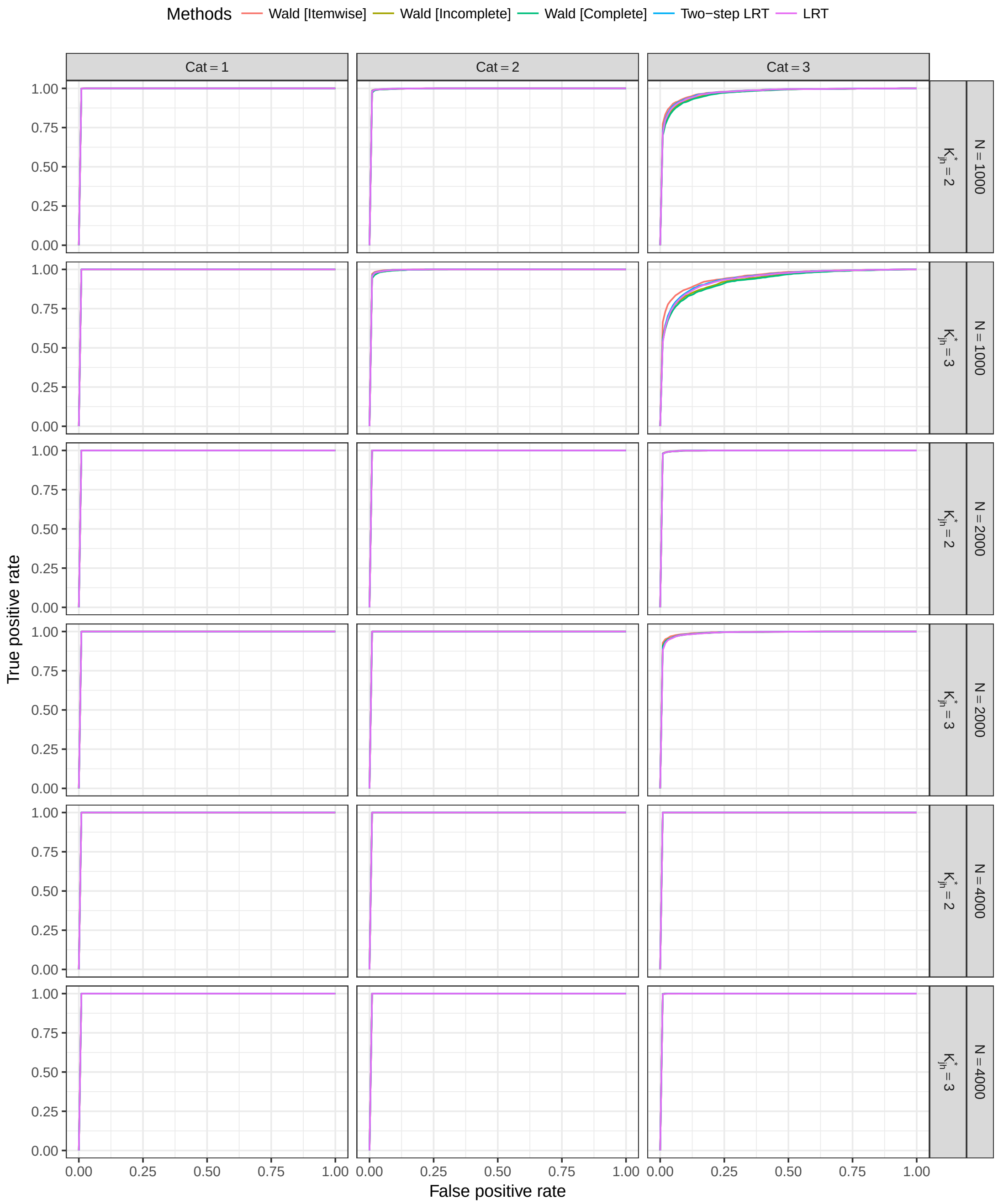


Figure 4.26: ROC curves for the DINA processing function: *A*-CDM-generated processing function under moderate item quality

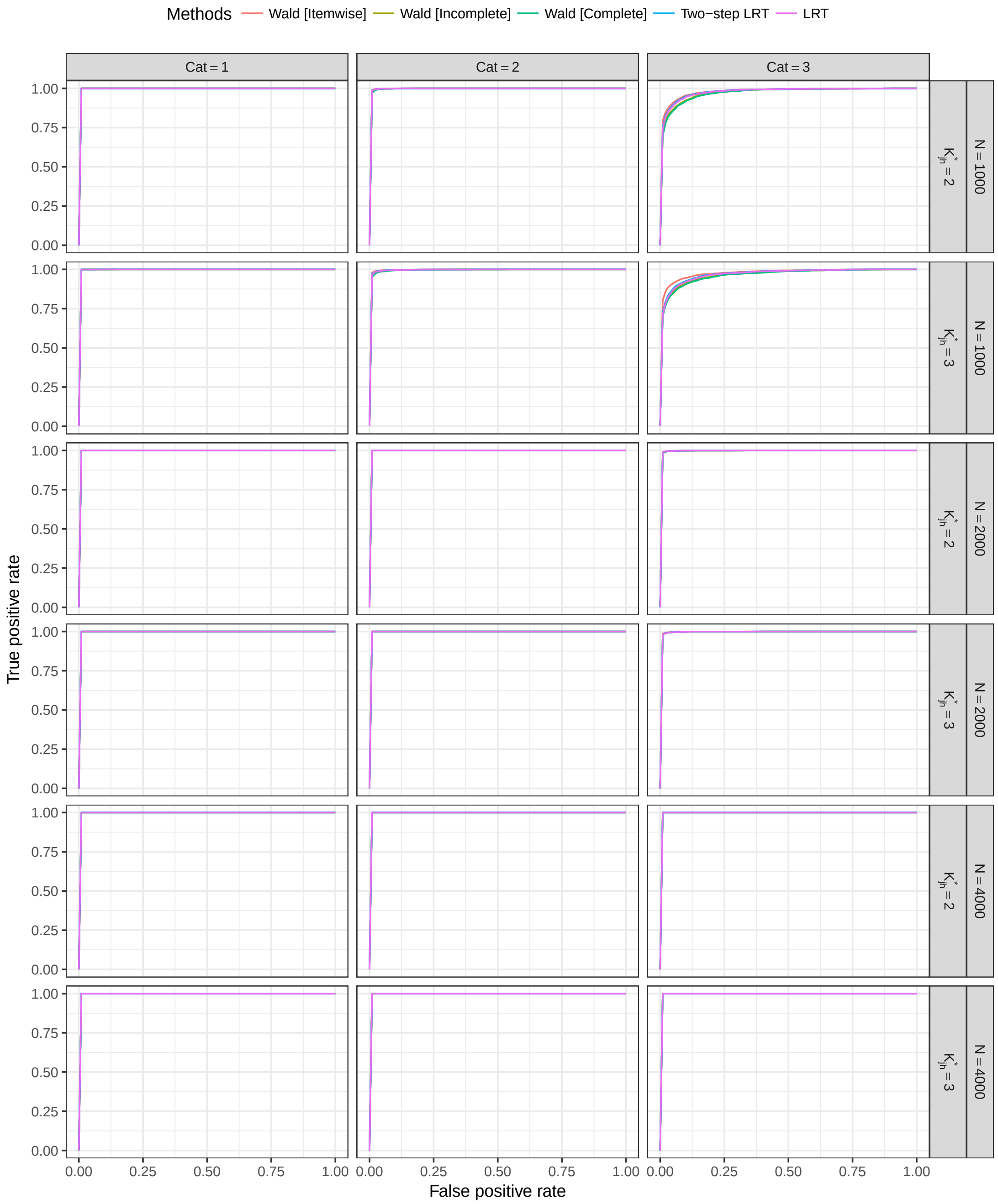


Figure 4.27: ROC curves for the DINO processing function: *A*-CDM-generated processing function under moderate item quality

Table 4.7: Power for the DINO processing function: DINA-generated data

N	Quality	K^*_{jh}	Wald [Itemwise]			Wald [Incomplete]			Wald [Complete]			Two-step LRT			LRT		
			Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3
1000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.992	1.000	1.000	1.000	1.000	1.000	1.000
	Moderate	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000	1.000
	Low	2	1.000	0.982	0.828	1.000	0.980	0.813	1.000	0.981	0.810	1.000	0.982	0.838	1.000	0.981	0.829
		3	0.981	0.762	0.602	0.976	0.733	0.582	0.976	0.740	0.584	0.985	0.784	0.588	0.986	0.789	0.568
2000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Moderate	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Low	2	1.000	1.000	0.993	1.000	1.000	0.991	1.000	1.000	0.992	1.000	1.000	0.994	1.000	1.000	0.992
		3	1.000	0.988	0.916	1.000	0.986	0.908	1.000	0.987	0.908	1.000	0.992	0.930	1.000	0.992	0.925
4000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Moderate	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Low	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	0.996	1.000	1.000	0.996	1.000	1.000	0.996	1.000	1.000	0.998	1.000	1.000	0.998

Table 4.8: Power for the ACDM processing function: DINA-generated data

N	Quality	K_{jh}^*	Wald [Itemwise]			Wald [Incomplete]			Wald [Complete]			Two-step LRT			LRT		
			Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3
1000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	1.000	1.000	1.000	1.000	1.000	1.000
	Moderate	2	1.000	1.000	0.999	1.000	1.000	0.999	1.000	1.000	0.999	1.000	1.000	0.999	1.000	1.000	0.999
		3	1.000	1.000	0.980	1.000	1.000	0.979	1.000	1.000	0.979	1.000	1.000	0.986	1.000	1.000	0.986
	Low	2	0.894	0.686	0.418	0.898	0.697	0.422	0.908	0.712	0.428	0.898	0.694	0.446	0.902	0.692	0.446
		3	0.774	0.432	0.289	0.768	0.417	0.294	0.794	0.448	0.304	0.806	0.461	0.313	0.820	0.458	0.286
2000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Moderate	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Low	2	0.996	0.962	0.757	0.996	0.965	0.758	0.997	0.968	0.764	0.996	0.963	0.770	0.996	0.962	0.761
		3	0.989	0.834	0.584	0.988	0.831	0.584	0.990	0.838	0.587	0.992	0.852	0.624	0.990	0.851	0.614
4000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Moderate	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Low	2	1.000	1.000	0.974	1.000	1.000	0.974	1.000	1.000	0.974	1.000	1.000	0.972	1.000	1.000	0.970
		3	1.000	0.994	0.889	1.000	0.994	0.884	1.000	0.995	0.888	1.000	0.996	0.902	1.000	0.996	0.899

Table 4.9: Power for the DINA processing function: DINO-generated data

N	Quality	K^*_{jh}	Wald [Itemwise]			Wald [Incomplete]			Wald [Complete]			Two-step LRT			LRT		
			Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3
1000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	0.720	1.000	1.000	0.726	1.000	1.000	0.715	1.000	1.000	0.946	1.000	1.000	0.946
	Moderate	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	0.858	1.000	1.000	0.863	1.000	1.000	0.851	1.000	1.000	0.898	1.000	1.000	0.898
	Low	2	1.000	0.974	0.718	1.000	0.976	0.722	1.000	0.978	0.730	1.000	0.976	0.740	1.000	0.976	0.740
		3	0.968	0.686	0.326	0.972	0.708	0.341	0.968	0.700	0.336	0.974	0.706	0.318	0.978	0.736	0.328
2000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	0.922	1.000	1.000	0.927	1.000	1.000	0.912	1.000	1.000	1.000	1.000	1.000	1.000
	Moderate	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	0.987	1.000	1.000	0.988	1.000	1.000	0.986	1.000	1.000	0.990	1.000	1.000	0.990
	Low	2	1.000	1.000	0.980	1.000	1.000	0.979	1.000	1.000	0.982	1.000	1.000	0.982	1.000	1.000	0.982
		3	1.000	0.980	0.644	1.000	0.981	0.656	1.000	0.980	0.644	1.000	0.985	0.625	1.000	0.986	0.642
4000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Moderate	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Low	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	0.914	1.000	1.000	0.916	1.000	1.000	0.916	1.000	1.000	0.926	1.000	1.000	0.922

Table 4.10: Power for the ACDM processing function: DINO-generated data

N	Quality	K^*_{jh}	Wald [Itemwise]			Wald [Incomplete]			Wald [Complete]			Two-step LRT			LRT		
			Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3	Cat=1	Cat=2	Cat=3
1000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	0.638	1.000	1.000	0.646	1.000	1.000	0.634	1.000	1.000	0.930	1.000	1.000	0.930
	Moderate	2	1.000	1.000	0.996	1.000	1.000	0.996	1.000	1.000	0.996	1.000	1.000	0.996	1.000	1.000	0.996
		3	1.000	1.000	0.778	1.000	1.000	0.790	1.000	1.000	0.778	1.000	1.000	0.807	1.000	1.000	0.806
	Low	2	0.926	0.661	0.333	0.934	0.676	0.348	0.945	0.696	0.358	0.926	0.664	0.360	0.931	0.674	0.372
		3	0.739	0.390	0.166	0.756	0.414	0.182	0.766	0.422	0.202	0.736	0.406	0.190	0.760	0.430	0.198
2000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	0.925	1.000	1.000	0.933	1.000	1.000	0.916	1.000	1.000	0.997	1.000	1.000	0.997
	Moderate	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	0.974	1.000	1.000	0.977	1.000	1.000	0.977	1.000	1.000	0.974	1.000	1.000	0.974
	Low	2	1.000	0.944	0.686	1.000	0.948	0.692	1.000	0.951	0.705	1.000	0.944	0.685	1.000	0.946	0.704
		3	0.984	0.774	0.402	0.984	0.780	0.412	0.986	0.786	0.414	0.984	0.774	0.374	0.984	0.784	0.400
4000	High	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Moderate	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		3	1.000	1.000	0.999	1.000	1.000	0.999	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000
	Low	2	1.000	1.000	0.951	1.000	1.000	0.952	1.000	1.000	0.953	1.000	0.999	0.949	1.000	1.000	0.954
		3	1.000	0.987	0.720	1.000	0.988	0.726	1.000	0.988	0.728	1.000	0.986	0.717	1.000	0.988	0.728

Chapter 5
Discussion

Cognitively diagnostic assessment (CDA; de la Torre & Minchen, 2014) has attracted increasing attentions in recent years because of its potential to provide detailed information about students' strengths and weaknesses. To fulfil its potential, a well-developed CDA should be a minimum requirement. However, for many recently developed CDAs (e.g., Bradshaw, Izsák, Templin, & Jacobson, 2014; Tjoe & de la Torre, 2014), multiple-choice items are still predominantly used, though constructed-response items have been shown to be more informative for diagnostic purposes (Birenbaum & Tatsuoka, 1987; Birenbaum, Tatsuoka, & Gutvirtz, 1992). The overemphasis on multiple-choice items can be attributed to several reasons, and one of the reasons may be the constraints of the current development of psychometric models and procedures for the constructed-response items.

CDMs play a critical role in making valid inference about students' attribute patterns based on their observed item responses. Despite the large number of CDMs available, most of them are designed for dichotomous responses. This constrains the use of polytomously scored items such as constructed-response items. More importantly, developing a psychometric model is merely the initial step in a complete data analysis process, and a set of procedures, including, among others, Q-matrix validation, model-data fit and item fit evaluation, condensation rule selection, and differential item functioning detection, may also need to be developed. A body of research can be found in literature on these topics, but most of them are related to CDMs for dichotomous responses. This results in another challenge in the use of polytomously scored items in

CDAs.

This dissertation developed a psychometric model for polytomously scored items, as well as associated statistical procedures for Q-matrix validation and condensation rule determination. Specifically, in Chapter 2, the sequential G-DINA model, which is particularly suitable for items that involve a sequence of steps in the problem-solving process, was proposed. This model can utilize the information about the association between attributes and steps when it is available. If the step and attribute association is unknown, the sequential G-DINA model is still applicable, but is equivalent to nominal response model. Simulation studies have shown that the MMLE/EM algorithm can be used to accurately recover item and attribute parameters under varied conditions.

Assuming a sequential process in problem solving and modeling each step separately are not novel in the field of educational measurement because items of this type are common in achievement testing. Examples of IRT models for items of this type include Tutz's (1997) sequential model, Verhelst, Glas, and de Vries's (1997) step model, and Samejima's (1997) acceleration model. These models belong to a more general model framework, namely, continuation ratio model (Agresti, 2013), and their properties have been investigated by van der Ark (2001), and Hemker, van der Ark, and Sijtsma (2001). However, these IRT models have gained less attentions than other polytomous IRT models such as Samejima's (1969) graded response model, because, among other reasons, under unidimensional assessment, all these models tend to perform similarly (Verhelst et al., 1997). The sequential G-DINA model is also a special case of the continuation ratio model, but compared with other polytomous CDMs, it allows researchers to model the association between steps and attributes directly. This is a distinct feature of CDAs from traditional unidimensional assessment. Because the attributes are typically finer-grained, it is very likely that different attributes are involved at different steps. Considering this information, as in the sequential G-DINA model, might lead to more accurate classification.

Like most CDMs, the proposed sequential G-DINA model relies on a Q-matrix (Tatsuoka, 1983) to specify the association between attributes and items. Nevertheless, it goes one step further and can consider the attribute and step association specified in a category level Q-matrix or $\mathrm{Q_C}$-matrix. The $\mathrm{Q_C}$-matrix provides extra information, which could facilitate person classification; but developing the $\mathrm{Q_C}$-matrix may be more challenging than developing the Q-matrix, and thus misspecifications are more likely to occur. Chapter 3 of this dissertation developed a stepwise procedure for empirically validating the $\mathrm{Q_C}$-matrix. Because the sequential G-DINA model is equivalent to the G-DINA model for dichotomous response data, the stepwise procedure can also be used for dichotomous data without assuming the specific condensation rule, as long as the underlying model is subsumed by the G-DINA model. Compared with other Q-matrix validation procedures for dichtomous responses (e.g., Chiu, 2013; de la Torre & Chiu, 2016), the stepwise procedure takes both hypothesis test and effect size measure into consideration. Under various conditions through the simulation studies, the stepwise method performs well in terms of both false positive and true positive rates. It should be noted that the stepwise Q-matrix validation procedure assumes that a provisional $\mathrm{Q_C}$-matrix is available and largely correct. This is different from Q-matrix learning algorithms (e.g., J. Liu, Xu, & Ying, 2012) that aim to recover the Q-matrix without a provisional one.

Although no condensation rule is specified for each step in the sequential G-DINA model, knowing the condensation rule could provide additional insights into how attributes are translated into manifest item responses. Specifying models with condensation rules in line with the underlying cognitive processes makes the data analysis more defensible. Compared with the G-DINA model used in the sequential G-DINA model, models based on specific condensation rule could be simpler, and therefore yield more reliable parameter estimation. However, for domain experts, identifying the condensation rule for each step of an item could be as difficult as, if not more difficult than, determining the required attributes for this step. To address this issue, Chapter 4 of this dissertation examined whether the Wald test and likelihood ratio test can be used to empirically determine several types of condensation rules for each step of an item. Type I error and power of these hypothesis tests were examined. The performance of hypothesis tests was influenced by many factors. Under unfavorable conditions, all hypothesis tests have inflated Type I error, but under favorable conditions, the Type I error rates for all tests are close to the nominal level.

To sum up, this dissertation developed the sequential G-DINA model for a special type of polytomously scored items, as well as two procedures for validating the attribute and step association and for determining the condensation rule for each step. The sequential model and the associated procedures offer a set of psychometric tools for analyzing polytomously scored items. Despite promising results, further research along this line is needed. First of all, although the sequential G-DINA model can be used for both graded and nominal response data, the stepwise Q-matrix validation and condensation rule selection procedures in Chapters 3 and 4 are only suitable for the graded response data. It is important to generalize these procedures so that they can be used for the nominal responses. Also, in the simulation studies across the three chapters, five dichotomous items each requiring single attribute were included to ensure the Q-matrix is identifiable. Identifiability is a critical issue in CDMs. Although many studies can be found in literature (e.g., Y. Chen, Liu, Xu, & Ying, 2015; Köhn & Chiu, 2017), all of them are based on dichotomous response data. It is not clear whether the identifiability can be achieved at the step level in the Q_C-matrix. Additionally, the stepwise Q-matrix validation procedure assumes that the number of attributes measured by the test is known; in practice, however, it is possible that students use some attributes that are not identified by experts.

Apart from the Q-matrix validation and condensation rule selection, many other procedures are needed. For example, evaluating model-data fit for the sequential G-DINA model is a topic that is worth exploring. Many indices have been developed for CDMs for dichotomous responses, such as indices based on the transformed correlation and log odds ratio in J. Chen, de la Torre, and Zhang (2013), and M_2 statistic (Hansen, Cai, Monroe, & Li, 2016; Y. Liu, Tian, & Xin, 2016), and several other measures examined by Hu, Miller, Huggins-Manley, and Chen (2016). The performance of these indices in conjunction with the sequential G-DINA moel for polytomous response data needs to be examined.

Last but not least, the ultimate goal of developing the sequential G-DINA model and the associated statistical procedures is to provide valid and reliable feedback to teachers and students immediately after the exam. This could be challenging when constructed response items are involved because these items may need to be graded by human raters. It is intuitive to consider some automatic scoring algorithms to accelerate the grading process. In addition, to fulfill the potential of the diagnostic assessments, it is important to explore how the test can be administered adaptively, and how the assessments can be embeded in class instructions.

5.1 References

Agresti, A. (2013). *Categorical data analysis*. New York: John Wiley & Sons.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats - It does make a difference for diagnostic purposes. *Applied Psychological Measurement*, *11*, 385–395.

Birenbaum, M., Tatsuoka, K. K., & Gutvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement*, *16*, 353–363.

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational measurement: Issues and practice*, *33*, 2–14.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*, 123–140.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based on diagnostic classification models. *Journal of the American Statistical Association*, *110*, 850–866.

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

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

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

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology*, *69*, 225–252.

Hemker, B. T., van der Ark, L. A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika*, *66*, 487–506.

Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y.-H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing*, *16*, 119–141.

Köhn, H.-F., & Chiu, C.-Y. (2017). A procedure for assessing the completeness of the q-matrices of cognitively diagnostic tests. *Psychometrika*, *82*, 112-132.

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

Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics*, *41*, 3–26.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern Item Response Theory* (pp. 85–100). Springer.

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

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.

Tutz, G. (1997). Sequential models for ordered responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 139–152). New York: Springer-Verlag.

van der Ark, L. A. (2001). Relationships and Properties of Polytomous Item Response Theory Models. *Applied Psychological Measurement*, *25*, 273–282.

Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A step model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–138). New York: Springer-Verlag.