SOME THEORETICAL AND APPLIED DEVELOPMENTS TO SUPPORT COGNITIVE LEARNING AND ADAPTIVE TESTING

BY

SHIYU WANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Jeff A. Douglas, Chair
Professor Hua-Hua Chang
Assistant Professor Steven A. Culpepper
Assistant Professor Georgios Fellouris
Associate Professor Jinming Zhang

Abstract

Cognitive diagnostic Modeling (CDM) and Computerized Adaptive Testing (CAT) are two useful tools to measure subjects' latent abilities from two different aspects. CDM plays a very important role in the fine-grained assessment, where the primary purpose is to accurately classify subjects according to the skills or attributes they possess, while CAT is a useful tool for coarse-grained assessment, which provides a single number to indicate the student's overall ability. This thesis discusses and solves several theoretical and applied issues related to these two areas.

The first problem we investigate related to a nonparametric classifier in Cognitive Diagnosis. Latent class models for cognitive diagnosis have been developed to classify examinees into one of the 2^K attribute profiles arising from a K-dimensional vector of binary skill indicators. These models recognize that response patterns tend to deviate from the ideal responses that would arise if skills and items generated item responses through a purely deterministic conjunctive process. An alternative to employing these latent class models is to minimize the distance between observed item response patterns and ideal response patterns, in a nonparametric fashion that utilizes no stochastic terms for these deviations. Theorems are presented that show the consistency of this approach, when the true model is one of several common latent class models for cognitive diagnosis. Consistency of classification is independent of sample size, because no model parameters need to be estimated. Simultaneous consistency for a large group of subjects can also be shown given some conditions on how sample size and test length grow with one another.

The second issue we consider is still within CDM framework, however our focus is about the model misspecification. The maximum likelihood classification rule is a standard method to classify examinee attribute profiles in cognitive diagnosis models. Its asymptotic behavior is well understood when the model is assumed to be correct, but has not been explored in the case of misspecified latent class models. We investigate the consequences of using a simple model when the true model is different. In general, when a CDM is misspecified as a conjunctive model, the MLE for attribute profiles is not necessarily consistent. A sufficient condition for the MLE to be a consistent estimator under a misspecified DINA model is found. The true model can be any conjunctive models or even a compensatory model. Two examples are provided

to show the consistency and inconsistency of the MLE under a misspecified DINA model. A Robust DINA MLE technique is proposed to overcome the inconsistency issue, and theorems are presented to show that it is a consistent estimator for attribute profile as long as the true model is a conjunctive model. Simulation results indicate that when the true model is a conjunctive model, the Robust DINA MLE and the DINA MLE based on the simulated item parameters can result in relatively good classification results even when the test length is short. These findings demonstrate that simple models can be fitted without severely affecting classification accuracy in some cases.

The last one discusses and solves a controversial issue related to CAT. In Computerized Adaptive Testing (CAT), items are selected in real time and are adjusted to the test-taker's ability. A long debated question related to CAT is that they do not allow test-takers to review and revise their responses. The last chapter of this thesis presents a CAT design that preserves the efficiency of a conventional CAT, but allows test-takers to revise their previous answers at any time during the test, and the only imposed restriction is on the number of revisions to the same item. The proposed method relies on a polytomous Item Response Theory model that is used to describe the first response to each item, as well as any subsequent revisions to it. The test-taker's ability is updated on-line with the maximizer of a partial likelihood function. I have established the strong consistency and asymptotic normality of the final ability estimator under minimal conditions on the test-taker's revision behavior. Simulation results also indicated this proposed design can reduce measurement error and is robust against several well-known test-taking strategies.

To my parents, Liping Jiang and Xuehong Wang

Acknowledgments

I would never have been able to finish my dissertation without the support of many people. First of all, I want to show my greatest gratitude to my two advisors, Professor Jeff Douglas and Professor Hua-Hua Chang. I am thankful for Professor Jeff Douglas's generous guidance and support, starting from the first semester of my Ph.D study. My very first research, first presentation, and first job hunting, all were motivated and encouraged by him. His generosity in sharing research ideas, warm personality and kind encouragements have great influence to my attitude towards research and mentoring students. Also, thanks to Professor Hua-Hua Chang, who has helped me prepare my future career by teaching me from writing research papers and dealing with hard review comments, helping me improve both oral and written communication skills, and always creating many opportunities for me to present my research work everywhere. His enthusiasm , positive attitude and determination encouraged me whenever I a hard time in my graduate study.

In addition, I wish to thank the members of my thesis committee. Thanks to Professor Georgios Fellouris, who has provided countless hours of assistance and guidance throughout my dissertation work, which contributed to the Chapter 3 of my thesis. I would like to express my sincere appreciation to Professor Steven Culpepper and Professor Jinming Zhang for their insightful suggestions and constant support. I also thank the faculty of the Statistics Department at the University of Illinois. Each of them has had an influence on my education, particularly many thanks to Professors Xiaofeng Shao and Professor Annie Qu, who have generously offered their helpful discussions on my work and guidance to my future career.

Thanks also goes to the fellow students in the Statistics Department, who have helped me a lot during my five-year study. Special thanks to Chung Eun Lee, Peibei Shi, Yeonjoo Park, Xianyang Zhang, Jin Wang, Jianjun Hu, Srijan Sengupta, Xuan Bi, Xiwei Tang, Xueying Zheng, Christopher Kinson and many others. Thank you very much for taking courses with me, sharing memories and experiences with me. I also owe special thanks to Haiyan Lin, who is my friend and also my internship mentor at ACT, and Xin Li, my friend at ACT, for their helpful discussions of research work and suggestions for my career development. Many thanks also goes to my friends, Yan Yang and Yiling Hu, for their generosity of sharing and supporting to my research work.

Lastly, I would like to show my tremendous gratitude to my beloved parents and husband. Special thanks to my father, Xuehong Wang, who has gave me greatest treasure of personalities like grit, zest and self-motivation, and without whom, I could not have the chance to go to college. I also want to thank my dearest mother, Liping Jiang, who raises me to become a person with optimism, gratitude and curiosity. Those characteristics walk me through various difficulties in my life. I can never come so close to my dream without you. I also want to express my sincere thanks to my husband, Houping Xiao, for his unconditional love and support in this long journey.

Table of Contents

List of Tables

List of Figures

Chapter 1

Consistency of Nonparametric Classification in Cognitive Diagnosis

1.1 Introduction

Interest in cognitive diagnostic models (CDMs) which allow for the profiling of subjects according to a variety of latent characteristics has been observed since the 1990s. Especially, motivated by the No Child Left Behind Act of 2001 (Law, 2002), most developments and applications of CDMs have taken place in educational contexts, where CDMs aim to provide students with information concerning whether or not they have mastered each of a group of specific skills, which are often generically referred to as attributes. At the same time, CDMs have also been widely applied outside of the field of education. For example, CDMs were used as a tool for providing diagnosis of psychological disorders (Jaeger et al., 2006; Templin and Henson, 2006; De La Torre et al., 2015). Some researchers have also referred to CDMs as diagnostic classification models (DCMs; Rupp and Templin (2008b); Rupp et al. (2010)) for greater generality of the applications of CDMs.

The first chapter provides a theoretical foundation for a nonparametric method of cognitive diagnosis studied in Chiu and Douglas (2013). We begin with a review of latent class models for cognitive diagnosis, focusing on three particular models that play a role in assumptions of consistency theorems.

1.1.1 Cognitive Diagnostic Models

Latent class models for cognitive diagnosis are generally restricted to reflect some assumptions about the underlying process by which examinees respond to items. We focus on a few such models that assume a conjunctive response process, and a more thorough review of cognitive diagnostic models can be found in Rupp and Templin (2007).

An important feature in the models we consider is a Q matrix (Tatsuoka,1985). This matrix records which attributes or skills are required to correctly respond to each item. Suppose that there are N subjects, J items and K attributes to classify. Entry q_{jk} in the $J \times K$ matrix Q indicates whether item j requires the k^{th} attribute. Let $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_N$ be random item response vectors of N subjects, where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{iJ})'$.

Let $\boldsymbol{\alpha}_i$ denote the attribute pattern for subject i, where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iK})'$ and each α_{ik} takes values of either 0 or 1 for $k = 1, 2, ..., K$. Specifically α_{ik} is an indicator of whether the i^{th} subject possesses the k^{th} attribute.

Conjunctive latent class models for cognitive diagnosis express the notion that all attributes specified in Q for an item should be required to answer the item correctly, but allow for slips and guesses in ways that distinguish the models from one another. The DINA model, an extension of the two-class model of Macready and Dayton (1977), and named in Junker and Sijtsma (2001), is one such example. Consider ideal response patterns, patterns that would arise if attribute possession entirely determined responses. Denote this ideal response pattern by $\boldsymbol{\eta}_i = (\eta_{i1}, \eta_{i2}, ..., \eta_{iJ})'$, where $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$. It denotes whether subject i has mastered all the attributes required by item j. The DINA allows for deviations from this pattern according to slipping parameters for each item, $s_j = P(Y_{ij} = 0|\eta_{ij} = 1)$, and guessing parameters for each item, $g_j = P(Y_{ij} = 1|\eta_{ij} = 0)$. The item response function of the DINA model is then

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \mathbf{s}, \mathbf{g}) = (1 - s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})},$$

and a likelihood function may be constructed from these item response functions together with an assumption of independence of $\mathbf{Y}_i$ given the attribute vector $\boldsymbol{\alpha}_i$. Though it is a simple and practical model, the DINA has some strong restrictions (Roussos et al., 2007). In particular, it assumes that the probability of a correct item response, given non-mastery on at least one skill, does not depend on the number and type of required skills that are not mastered. The next model we consider differs in this regard, but has some restrictions of its own.

The NIDA model, introduced in Maris (1999), treats the slips and guesses at the subtask level. The ideal response patterns remain the same, but a subtask response $\eta_{ijk} = 0$ indicates whether subject i correctly applied attribute k to answer item j. In this model, $s_k = P(\eta_{ijk} = 0|\alpha_{ik} = 1, q_{jk} = 1)$, $g_k = P(\eta_{ijk} = 1|\alpha_{ik} = 0, q_{jk} = 1)$, and $Y_{ij} = 1$ only if all subtasks are correctly completed. By convention, let $P(\eta_{ijk} = 1|\alpha_{ik} = a, q_{jk} = 0) = 1$, no matter the value of α_{ik}. Then the item response function of the NIDA model is

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \mathbf{s}, \mathbf{g}) = \prod_{k=1}^{K} P(\eta_{ijk} = 1|\alpha_{ik}, s_k, g_k) = \prod_{k=1}^{K} [(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}]^{q_{jk}}.$$

A restriction of the NIDA model is that it implies items requiring the same set of attributes must have precisely the same item response functions. This can be viewed as a desired property in certain situations when the theory of the cognitive attributes is indeed correct, and the Q matrix describes the cognitive processes to solve items of a certain type sufficiently well. It is also parsimonious, which is helpful for small data sets that may not afford estimation of a more general parametric model (Roussos, Templin & Henson,

2007). However, in many situations it can readily be seen to conflict with data, because it implies such strict conditions on observed proportion correct values of the items. For instance, the model implies that any two items with the same entry in Q must have the same expected proportion correct. A generalization of this that allows slipping and guessing probabilities to vary across items is a reduced version of the Reparameterized Unified Model (Hartz and Roussos, 2008). In this model, the item response function is

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \pi^*, \mathbf{r}) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*q_{jk}(1-\alpha_k)}.$$

Here π_j^* is the probability a subject who possesses all of the required attributes answers item j correctly, and r_{jk}^* can be viewed as a penalty parameter for not possessing the k^{th} attribute, and is between 0 and 1.

These three specific models will be considered in the classification consistency theory of the next section. However, more general cognitive diagnostic models have been developed including conjunctive, disjunctive, and compensatory models. For example, see the G-DINA framework (De La Torre, 2011), the log-linear cognitive diagnostic model (Henson et al., 2009), and the general diagnostic model (Davier, 2008).

1.1.2 Nonparametric Classification for Cognitive Diagnosis

The restricted latent class models described above have become popular in cognitive diagnostic research. However, there are alternatives that do not assume any particular probability model. The rule space methodology (Tatsuoka, 1983, 1985) is a widely-known and early approach to diagnostic testing, that combines parametric modeling with the notion of an ideal response pattern. The idea behind rule space is to use Boolean descriptive functions to establish the relationship between the examinee's attribute pattern and the observed response pattern through the Q-matrix, after adjusting for a fitted item response model. Building on this method, but in a more nonparametric fashion, Barnes (2010) developed hill-climbing algorithms to build the Q-matrix and examinee classifications in a purely exploratory approach. Some recent research has attempted to classify attribute patterns by utilizing cluster analysis. Willse et al. (2007), for example, apply K-means clustering to cognitive diagnosis data generated by the reduced Reparameterized Unified Model (RUM). Ayers et al. (2008), test the performance of various common clustering methods in classifying examinees. Chiu et al. (2009), conducted a theoretical and empirical evaluation of hierarchical agglomerative and K-means clustering for grouping examinees into clusters having similar attribute patterns. They established conditions for clusters to match perfectly with corresponding latent classes with probability approaching 1 as test length increases. Park and Lee (2011), also examined a method of clustering attributes required to solve mathematics problems on the TIMSS by mapping item responses to an attribute matrix, and then conducting K-means and hierarchical agglomerative cluster analysis.

A direct approach to nonparametric classification is to match observed item response patterns to the nearest ideal response pattern. Chiu and Douglas (2013), studied this method, and found that accurate classification can be achieved when the true model is DINA and NIDA with slip and guess parameters considerably greater than 0. A step of the rule space method (Tastuoka, 1983), is quite similar. However, rule space first requires calibration of the ability parameter based on an item response model, and cannot be viewed as a wholly nonparametric method. Rule space attempts to reduce the dimensionality of observed response patterns and the ideal response scores by mapping to a pair of new variables (θ, η) in a Cartesian product space,then calculates a Mahalanobis distance between the two patterns to identify the attribute pattern for each subject. By contrast, the Chiu and Douglas (2013) method requires fewer steps. The estimator of $\boldsymbol{\alpha}$ in this method would be perfect if all slip and guess parameters were 0, but still performs with good relative efficiency even when this is not the case. To formally define the estimator, first recall that $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_N$ are random item response vectors of N subjects to J items, where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{iJ})'$. Define $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ to be the j^{th} component of the ideal response pattern from the i^{th} subject, and let $\boldsymbol{\eta}$ denote this pattern. Then all possible ideal response patterns, $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, ..., \boldsymbol{\eta}_{2^K}$, can be constructed from all 2^K possible values for $\boldsymbol{\alpha}_i$. Because the $\boldsymbol{\eta}_i$ is determined by $\boldsymbol{\alpha}_i$, we define the distance between the observed item response vector for the i^{th} subject $\mathbf{y}_i$ and the ideal response pattern under attribute profile $\boldsymbol{\alpha}_m$ to be $D(\mathbf{y}_i, \boldsymbol{\alpha}_m)$, for $m = 1, 2, ..., 2^K$.

The nonparametric classification estimator $\hat{\boldsymbol{\alpha}}$ arises by minimizing some measure of distance over all possible ideal response vectors, and determining the $\boldsymbol{\alpha}$ associated with the nearest ideal response vector. It is natural to use Hamming distance for clustering with binary data, which simply counts the components of $\mathbf{y}_i$ and $\boldsymbol{\eta}_m$ that disagree,

$$D(\mathbf{y}_i, \boldsymbol{\alpha}_m) = \sum_{j=1}^{J} |y_{ij} - \eta_{mj}| = \sum_{j=1}^{J} d_j^i(\boldsymbol{\alpha}_m). \tag{1.1}$$

Minimizing this distance over all possible values of the latent attribute vector produces the estimator,

$$\hat{\boldsymbol{\alpha}}_i = \arg \min_{m \in \{1,2,...,2^K\}} D(\mathbf{y}_i, \boldsymbol{\alpha}_m). \tag{1.2}$$

This estimator can result in ties, especially in short exams. The probability of a tie converges to 0 as exam length increases, so ties play no role in the theory of the estimator. However, in practice one must decide how to break ties. This can be done by randomly choosing among the tied values, or by implementing a weighted version Hamming distance to reduce their frequency. The next section considers the asymptotic theory for $\hat{\boldsymbol{\alpha}}$, and consistency theorems are given when the true model is the DINA, NIDA, or Reparameterized Unified

model.

1.2 Consistent Classification Theory

Though the nonparametric classifier based on minimizing Hamming distance to ideal response patterns is clearly consistent when slipping and guessing parameters are 0 and data are ideal response patterns, we show it also yields consistent classification under a variety of models when the stochastic terms differ considerably from 0. Chiu and Douglas (2013) gave a heuristic justification for the theoretical underpinnings of the classifier, and a formal analysis is given below. First we provide the assumptions and conditions for consistency under different latent class models, along with their justifications.

1.2.1 Assumptions and Conditions

First we make the standard assumptions of independent subjects and conditional independence.

Assumption 1: The item response vectors $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_N$ for subjects $1, 2, ..., N$ are statistically independent.

Assumption 2: For subject i, the item response $Y_{i1}, Y_{i2}, ..., Y_{iJ}$ are statistically independent conditional on attribute vector $\boldsymbol{\alpha}_i$.

Let q_{jk} be the j, k element of the Q matrix, and define $B_j = \{k|q_{jk} = 1\}$.

For some number $\delta \in (0, .5)$ we have the following conditions on parameters of the possible true model:

Condition (a.1): When data arise from the DINA model, parameters g_j and s_j satisfy that $g_j < 0.5 - \delta$ and $s_j < 0.5 - \delta$.

Condition (a.2): When data arise from the NIDA model, $g_k < 0.5 - \delta$, for $k = 1, 2, ..., K$, and $\prod_{k \in B_j}(1 - s_k) > 0.5 + \delta$, for $j = 1, 2, ..., J$.

Condition (a.3): When data arise from the Reduced RUM model, $\pi_j^* > 0.5 + \delta$ for every j, and for some $k \in B_j, r_{jk}^* < 0.5 - \delta$.

Condition (b): Define $A_{m,m'} = \{j|\eta_{mj} \neq \eta_{m'j}\}$, where m and $m^{'}$ index different attribute patterns among the 2^K possible patterns. Card$(A_{m,m'}) \to \infty$ as $J \to \infty$.

Condition (c): The number of subjects and the test length satisfy the relationship that $\forall \varepsilon > 0$, $Ne^{-2J\varepsilon^2} \to 0$ as $J \to \infty$.

Conditions (a.1) and (a.2) bound slipping and guessing parameters in the DINA and NIDA models away from 0.5. These are reasonable assumptions for a valid model, because the probability of a subject answering an item correctly should be at least primarily determined by possession or nonpossession of the required

attributes. If such assumptions are not met, diagnostic modeling will not be as useful, either with the nonparametric classifier or with the parametric model. The condition essentially says that the most likely response for someone who has mastered the attributes is 1, and the most likely response for someone who has not mastered the required attributes is 0. Certainly masters of the attributes should have a higher probability of success, though requiring it is at least 0.5 does make the conditional somewhat restrictive. Condition (a.3) expresses the same notion for the Reduced RUM model, which can be rewritten as a NIDA model in which slipping and guessing parameters can vary with the item. Condition (b) guarantees that for each pair of attribute patterns, there is an infinite amount of information to separate them as the number of items grows to infinity. Finally Condition (c) is established in order to get the simultaneous consistent classification of a whole sample of subjects, and is unnecessary when considering the consistent classification of a single subject.

1.2.2 Asymptotic Results

In this section, three propositions will be introduced first in order to prove the consistency results for a single subject and a sample of subjects.

Proposition 1 Under Assumptions 1, 2, Conditions (a.1), (a.2), (a.3) and (b), for every $i \in \{1, 2, ..., N\}$, the true attribute pattern will minimize $E[D(\mathbf{Y}_i, \boldsymbol{\alpha}_m)]$ $(m = 1, 2, ..., 2^K)$, that is

$$\boldsymbol{\alpha}_0 = \underset{\boldsymbol{\alpha}_m}{\operatorname{argmin}} \quad E\left[D(\mathbf{Y}_i, \boldsymbol{\alpha}_m)\right].$$

Proof. Suppose the real attribute pattern for a fixed item response vector $\mathbf{Y}_i$ is $\boldsymbol{\alpha}_1$. Let $\boldsymbol{\alpha}_2$ be another attribute pattern. Because $E[D(\mathbf{Y}_i, \boldsymbol{\alpha}_m)] = \sum_{j=1}^{J} E|Y_{ij} - \eta_{mj}|$, we just need to compare $E(|Y_{ij} - \eta_{1j}|)$ and $E(|Y_{ij} - \eta_{2j}|)$ for every j. Note that if $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_2$, there must be some j such that $\eta_{1j} \neq \eta_{2j}$. Let $A_{1,2} = \{j|\eta_{1j} \neq \eta_{2j}\}$. Then for every $j \in A_{1,2}$, we have

$$\eta_{1j} = 1, \quad \eta_{2j} = 0, \quad E(|Y_{ij} - \eta_{1j}|) - E(|Y_{ij} - \eta_{2j}|) = 1 - 2P(Y_{ij} = 1);$$
$$\eta_{1j} = 0, \quad \eta_{2j} = 1, \quad E(|Y_{ij} - \eta_{1j}|) - E(|Y_{ij} - \eta_{2j}| = 2P(Y_{ij} = 1) - 1.$$

The problem then turns out to be deriving the specific $P(Y_{ij} = 1)$ under different models.

(1) When data arise from DINA model, $P(Y_{ij} = 1) = (1 - s_j)^{\eta_j} * g_j^{1-\eta_j}$.

When $\eta_{1j} = 1$ and $\eta_{2j} = 0$,

$$E\left(|Y_{ij} - \eta_{1j}|\right) - E(|Y_{ij} - \eta_{2j}|) = 2s_j - 1.$$

When $\eta_{1j} = 0$ and $\eta_{2j} = 1$,

$$E(|Y_{ij} - \eta_{1j}|) - E(|Y_{ij} - \eta_{2j}| = 2g_j - 1.$$

From Condition (a.1) we know that $g_j < 0.5 - \delta$, $s_j < 0.5 - \delta$ for some positive number δ. So we can get $E(|Y_{ij} - \eta_{1j}|) < E(|Y_{ij} - \eta_{2j}|)$ under the DINA model.

(2) When data arise from the NIDA model, $P(Y_{ij} = 1) = \prod_{k=1}^{K} \left[(1 - s_k)^{\alpha_k} g_k^{1-\alpha_k}\right]^{q_{jk}}$

When $\eta_{1j} = 1$ and $\eta_{2j} = 0$,

$$E(|Y_{ij} - \eta_{1j}|) - E(|Y_{ij} - \eta_{2j}|) = 1 - 2 \prod_{k \in B_j} (1 - s_k).$$

When $\eta_{1j} = 0$ and $\eta_{2j} = 1$,

$$E(|Y_{ij} - \eta_{1j}|) - E(|Y_{ij} - \eta_{2j}| = 2 * \prod_{k \in B_j} g_k^{1-\alpha_k} (1 - s_k)^{\alpha_k} - 1.$$

According to Condition (a.2), for some positive number δ, $\prod_{k \in B_j}(1-s_k) > 0.5+\delta$, so $1-2\prod_{k \in B_j}(1-s_k) < 0$. Furthermore, when $\eta_{1j} = 0$, there must be some $k^{'} \in B_j$ such that $\alpha_{k'} = 0$. Then $g_{k'} < 0.5 - \delta < 0.5$, we can get $2 * \prod_{k \in B_j} g_k^{1-\alpha_k}(1 - s_k)^{\alpha_k} - 1 < 0$. So $E(|y_{ij} - \eta_{1j}|) < E(|y_{ij} - \eta_{2j}|)$ is also correct under the NIDA model.

(3) When data arise from the Reduced RUM model, $P(Y_{ij} = 1) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*q_{jk}(1-\alpha_k)}$.

When $\eta_{1j} = 1$ and $\eta_{2j} = 0$,

$$E(|Y_{ij} - \eta_{1j}|) - E(|Y_{ij} - \eta_{2j}|) = 1 - 2\pi_j^*.$$

When $\eta_{1j} = 0$ and $\eta_{2j} = 1$,

$$E(|Y_{ij} - \eta_{1j}|) - E(|Y_{ij} - \eta_{2j}|) = 2 * \pi_j^* \prod_{k \in B_j} r_{jk}^{*(1-\alpha_k)} - 1.$$

Similar to the argument for the NIDA model, there must be some $k^{'} \in B_j$ such that $\alpha_{k'} = 0$. With Condition (a.3) that $\pi_j^* > 0.5 + \delta$ for each j, we can get $1 - 2\pi_j^* < 0$. And $r_{jk'} < 0.5 - \delta$, then $\pi_j^* \prod_{k \in B_j} r_{jk}^{*(1-\alpha_k)} = r_{jk'} \pi_j^* \prod_{k \in B_j, k \neq k'} r_{jk}^* < 0.5$

From the above argument,we can see that no matter which of the models is true, $E(|Y_{ij} - \eta_{1j}|) < E(|Y_{ij} - \eta_{2j}|)$, when $j \in A_{1,2}$. Otherwise, for every $j \in A_{1,2}^C$, $E(|Y_{ij} - \eta_{1j}|) = E(|Y_{ij} - \eta_{2j}|)$. Then we can

see that:

$$E\left[\sum_{j=1}^{J}|Y_{ij}-\eta_{1j}|\right] < E\left[\sum_{j=1}^{J}|Y_{ij}-\eta_{2j}|\right].$$

$\diamondsuit$

The next proposition states that the true attribute pattern will be well separated from one another as test length goes to infinity.

Proposition 2 Under the assumptions and conditions of Proposition 1, in addition to Condition(b), and suppose $\boldsymbol{\alpha}_1$ is the true attribute profile, $\boldsymbol{\alpha}_2$ is another different attribute profile,

$$\lim_{J\to\infty} E\left[D(\mathbf{Y}_i,\boldsymbol{\alpha}_2)\right] - E\left[D(\mathbf{Y}_i,\boldsymbol{\alpha}_1)\right] = \infty.$$

Proof. Note that the difference between $E[D(\mathbf{Y}_i,\boldsymbol{\alpha}_1)]$ and $E[D(\mathbf{Y}_i,\boldsymbol{\alpha}_2)]$ is only determined by the values when $j \in A_{1,2}$. We first prove this proposition under the DINA model.

$$\begin{aligned}\lim_{J\to\infty} E[D(\mathbf{Y}_i,\boldsymbol{\alpha}_2)] - E[D(\mathbf{Y}_i,\boldsymbol{\alpha}_1)] &= \lim_{J\to\infty}\sum_{j\in J_1} 1-2g_j + \lim_{J\to\infty}\sum_{j\in J_2} 1-2s_j \\ &> \lim_{J\to\infty}\sum_{j\in J_1}\delta + \lim_{J\to\infty}\sum_{j\in J_2}\delta = \infty\end{aligned}$$

Here $J_1 = \{j \in A_{1,2}|\eta_{1j}=0,\eta_{2j}=1\}$, $J_2 = \{j \in A_{1,2}|\eta_{1j}=1,\eta_{2j=0}\}$, and the infinite sum results from the cardinality of J_1 and J_2 going to infinity, as guaranteed by Condition (b).

The same argument can be applied to the NIDA model and the Reduced RUM.

$\diamondsuit$

The third proposition investigates the relationship between the average of $d_j^i(\boldsymbol{\alpha}_m)$ and its expectation $E[d_j^i(\boldsymbol{\alpha}_m)]$, for fixed i and every $m \in \{1,2,...,2^K\}$, when the test length goes to infinity.

Proposition 3 Under Assumptions 1 and 2, $\forall\varepsilon > 0$ and a fixed $i \in \{1,2,...,N\}$, define

$$B_\varepsilon(J) = \left\{\max_{m\in\{1,2,...,2^K\}} |\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])| \geq \varepsilon\right\}.$$

Then

$$\lim_{J\to\infty} P(B_\varepsilon(J)) = 0.$$

In order to prove Proposition 3, we need to apply the Hoeffding's Inequality as it is stated below:

Hoeffding's Inequality Let $Z_1, Z_2, ..., Z_n$ be independent random variables such that $\forall i, 0 \leq Z_i \leq 1$, Then

$$P\left(\{|\tfrac{1}{n}(\textstyle\sum_{i=1}^{n} Z_i - E[Z_i])| \geq \varepsilon\}\right) \leq 2\exp(-2n\varepsilon^2).$$

Proof of Proposition 3:

First we must show that for every $\varepsilon > 0$, $P(\{|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])| \geq \varepsilon\}) \leq 2\exp(-2J\varepsilon^2)$, and this is obtained by Hoeffding's Inequality.

Note that $\forall j \in 1, 2, ..., J, d_j^i(\boldsymbol{\alpha}_m) = |Y_{ij} - \eta_{mj}|$. For subject i, $Y_{ij}, j \in \{1, 2, ..., J\}$ are independent random variables, conditional on the true attribute pattern. This implies that $d_j^i(\boldsymbol{\alpha}_m), j \in \{1, 2, ..., J\}$ are independent random variables, and $0 \leq d_j^i(\boldsymbol{\alpha}_m) \leq 1$ is obvious. So $d_1^i(\boldsymbol{\alpha}_m), d_2^i(\boldsymbol{\alpha}_m), ..., d_J^i(\boldsymbol{\alpha}_m))$ are independent random variables which satisfy the conditions of Hoeffding's Inequality, which states that for every $\varepsilon > 0$,

$$P(\{|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])| \geq \varepsilon\}) \leq 2\exp(-2J\varepsilon^2).$$

Using this result we see that

$$\begin{aligned} P\left(\bigcup_{m=1}^{2^K}\left\{|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])| \geq \varepsilon\right\}\right) &\leq \sum_{m=1}^{2^K} P\left(\left\{|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])| \geq \varepsilon\right\}\right) \\ &\leq 2^{K+1}\exp(-2J\varepsilon^2) \end{aligned}$$

which implies that

$$\begin{aligned} &P\left(\left\{\max_{m\in\{1,2,...,2^K\}}|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])| \geq \varepsilon\right\}\right) \\ &= 1 - P\left(\left\{\max_{m\in\{1,2,...,2^K\}}|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])| < \varepsilon\right\}\right) \\ &= 1 - P\left(\bigcap_{m=1}^{2^K}\left\{|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])| < \varepsilon\right\}\right) \\ &= P\left(\bigcup_{m=1}^{2^K}\left\{|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])| \geq \varepsilon\right\}\right) \\ &\leq 2^{K+1}\exp(-2J\varepsilon^2) \end{aligned}$$

Note that 2^{K+1} and ε are constant, so this probability converges to 0 when J goes to infinity. $\diamondsuit$

The preceding propositions are now used to prove Theorem 1 , along with a corollary to show that for a single subject, the estimate of $\hat{\boldsymbol{\alpha}}$ by the nonparametric method will converge to the true attribute vector almost surely when test length goes to infinity.

Theorem 1. *For a particular subject i with true attribute pattern $\boldsymbol{\alpha}_i$, under Assumptions 1 and 2 and Conditions (a.1), (a.2), (a.3) and (b), the estimator $\hat{\boldsymbol{\alpha}}_i$ derived from the nonparametric method of equation 2 is a consistent estimator of $\boldsymbol{\alpha}_i$, provided one of the DINA model, NIDA model or Reduced RUM holds. Specifically,*

$$\lim_{J\to\infty} P(\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i) = 1.$$

Proof. For fixed subject i, $\forall\varepsilon > 0$, let event $A_\varepsilon(J) = \{|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i| > \varepsilon\}$, and $B_\varepsilon(J)$ as defined in Proposition 3. Then we show that $A_\varepsilon(J) \subset B_\varepsilon(J)$. In order to prove this, we can prove $B_\varepsilon(J)^C \subset A_\varepsilon(J)^C$.

Suppose $B_\varepsilon(J)^C$ is true, that is $\forall m \in \{1, 2, ..., 2^K\}, \forall\varepsilon > 0$

$$\left| \frac{1}{J} \sum_{j=1}^{J} \left(d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)] \right) \right| < \varepsilon$$

Then we can get that

$$\frac{1}{J} \sum_{j=1}^{J} E[d_j^i(\boldsymbol{\alpha}_m)] - \varepsilon < \frac{1}{J} \sum_{j=1}^{J} d_j^i(\boldsymbol{\alpha}_m) < \frac{1}{J} \sum_{j=1}^{J} E[d_j^i(\boldsymbol{\alpha}_m)] + \varepsilon.$$

If $\hat{\boldsymbol{\alpha}}_i \neq \boldsymbol{\alpha}_i$, then $\sum_{j=1}^{J} d_j^i(\hat{\boldsymbol{\alpha}}_i) < \sum_{j=1}^{J} d_j^i(\boldsymbol{\alpha}_i)$ and $\frac{1}{J}\sum_{j=1}^{J} d_j^i(\hat{\boldsymbol{\alpha}}_i) < \frac{1}{J}\sum_{j=1}^{J} d_j^i(\boldsymbol{\alpha}_i)$. These inequalities imply that

$$\frac{1}{J} \sum_{j=1}^{J} E[d_j^i(\hat{\boldsymbol{\alpha}}_i)] - \varepsilon < \frac{1}{J} \sum_{j=1}^{J} d_j^i(\hat{\boldsymbol{\alpha}}_i) < \frac{1}{J} \sum_{j=1}^{J} d_j^i(\boldsymbol{\alpha}_i) < \frac{1}{J} E[d_j^i(\boldsymbol{\alpha}_i)] + \varepsilon, \quad \forall\varepsilon > 0.$$

Thus for small enough ε we have

$$\sum_{j=1}^{J} E[d_j^i(\hat{\boldsymbol{\alpha}}_i)] < \sum_{j=1}^{J} E[d_j^i(\boldsymbol{\alpha}_i)].$$

This is contradictory to Proposition 1 that shows the true attribute pattern will minimize $E[D(\mathbf{Y}_i, \boldsymbol{\alpha}_m)]$, and Proposition 2 that when $J \to \infty$,the difference of the expectation of the distance defined by (1) under the wrong attribute pattern with that of under the true attribute pattern will go to infinity. So we may conclude that $B_\varepsilon(J)^C \subset A_\varepsilon(J)^C$, and equivalently $A_\varepsilon(J) \subset B_\varepsilon(J)$.

By the above claim and Proposition 3, we have $\forall \varepsilon > 0$

$$P\left(|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i| > \varepsilon\right) \leq P(B_\varepsilon(J)) \leq 2^{K+1} \exp(-2J\varepsilon^2) \to 0, \quad \text{as} \quad J \to \infty$$

Thus we have proved that if Assumptions 1 and 2 and Conditions (a.1), (a.2), (a.3) and (b) are satisfied, $\lim_{J\to\infty} P(\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i) = 1$ ♢

Corollary 1. Under Assumptions 1 and 2 and Conditions (a.1), (a.2), (a.3) and (b),

$$\lim_{J\to\infty} P(|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i| > \varepsilon, i.o) = 0.$$

Proof. We only need to prove that:

$$\sum_{J=1}^{\infty} P(B_\varepsilon(J)) < \infty$$

and the result will follow from the Borel-Cantelli Theorem. Note that

$$\begin{aligned} \sum_{J=1}^{\infty} P(B_\varepsilon(J)) &< \sum_{J=1}^{\infty} 2^{K+1} \exp(-2J\varepsilon^2) \\ &= 2^{K+1} \sum_{J=1}^{\infty} \exp(-2J\varepsilon^2). \end{aligned}$$

Define $\mathrm{f}(J) = \exp(-2J\varepsilon^2)$.

According to the convergence rule of series,

$$\begin{aligned} \lim_{J\to\infty} \frac{\mathrm{f}(J+1)}{\mathrm{f}(J)} &= \lim_{J\to\infty} \frac{\exp(-2(J+1)\varepsilon^2)}{\exp(-2J\varepsilon^2)} \\ &= \exp(-2\varepsilon^2) < 1 \end{aligned}$$

Then we have

$$\sum_{J=1}^{\infty} P(A_\varepsilon(J)) < \sum_{J=1}^{\infty} P(B_\varepsilon(J)) < \infty$$

which completes the proof of the corollary.

♢

Finally we investigate the joint consistency of a sample of N subjects. Essentially the same results hold, but there must be some control on the relative sizes of N and J as both go to infinity.

Theorem 2. *Under Assumptions 1 and 2 and Condition (a.1), (a.2), (a.3),(b)and (c)in section 2.1.1, provided one of the DINA model, NIDA model or Reduced RUM holds,*

$$\lim_{J\to\infty} P\left(\bigcap_{i=1}^{N}\{\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i\}\right) = 1.$$

Proof. Note that the only difference between Theorem 1 and Theorem 2 is that Theorem 2 has sample size N but Theorem 1 has only one subject. So Proposition 1 and Proposition 2 still hold for every subject in Theorem 2. For Proposition 3, $\forall\varepsilon > 0$, we can define $B_\varepsilon(N,J) = \{\max_{i\in\{1,2,\ldots,N\}}\{\max_{m\in\{1,2,\ldots,2^K\}} |\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - Ed_j^i(\boldsymbol{\alpha}_m))| \geq \varepsilon\}\}$. Then under Condition (c) we can show that

$$\lim_{J\to\infty} P\left(B_\varepsilon(N,J)\right) = 0.$$

For every $i \in \{1,2,...,N\}$, $P\left(\left\{\left|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])\right| \geq \varepsilon\right\}\right) \leq 2\exp(-2J\varepsilon^2)$ holds, so we see that,

$$\begin{aligned}
&P\left(\bigcup_{i=1}^{N}\bigcup_{m=1}^{2^K}\left\{|\frac{1}{J}\sum_{j=1}^{J}(d_j^J(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])| \geq \varepsilon\right\}\right)\\
&\leq \sum_{i=1}^{N}\sum_{m=1}^{2^K} P\left(\left\{\left|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])\right| \geq \varepsilon\right\}\right)\\
&\leq 2^{K+1}N\exp(-2J\varepsilon^2).
\end{aligned}$$

$\Longrightarrow$

$$
\begin{aligned}
& P\left(\left\{\max_{i\in\{1,2,\ldots,N\}}\left\{\max_{m\in\{1,2,\ldots,2^K\}}\left|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m)-E[d_j^i(\boldsymbol{\alpha}_m)])\right|\geq\varepsilon\right\}\right\}\right)\\
&=1-P\left(\left\{\max_{i\in\{1,2,\ldots,N\}}\left\{\max_{m\in\{1,2,\ldots,2^K\}}\left|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m)-E[d_j^i(\boldsymbol{\alpha}_m)])\right|<\varepsilon\right\}\right\}\right)\\
&=1-P\left(\bigcap_{i=1}^{N}\bigcap_{m=1}^{2^K}\left\{|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m)-E[d_j^i(\boldsymbol{\alpha}_m)])|<\varepsilon\right\}\right)\\
&=P\left(\bigcup_{i=1}^{N}\bigcup_{m=1}^{2^K}\left\{|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m)-E[d_j^i(\boldsymbol{\alpha}_m)])|\geq\varepsilon\right\}\right)\\
&\leq 2^{K+1}N\exp(-2J\varepsilon^2).
\end{aligned}
$$

Note that 2^{K+1} is constant, so this probability converges to 0 provided the sample size and test length have the relationship $Ne^{-2J\varepsilon^2}\to 0$ as $J\to\infty$ (Condition (c)).

Now define $A_\varepsilon(N,J)=\left\{\max_{i\in\{1,2,\ldots,N\}}|\hat{\boldsymbol{\alpha}}_i-\boldsymbol{\alpha}_i|>\varepsilon\right\}$, with the same argument as that in the proof of Theorem 1, we can prove that $A_\varepsilon(N,J)\subset B_\varepsilon(N,J)$. Then we can get that:

$$
\begin{aligned}
P\left(\bigcup_{i=1}^{N}\{|\hat{\boldsymbol{\alpha}}_i-\boldsymbol{\alpha}_i|>\varepsilon\}\right) &\leq P\left\{\max_{i\in\{1,2,\ldots,N\}}\left\{\max_{m\in\{1,2,\ldots,2^K\}}\left|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m)-E[d_j^i(\boldsymbol{\alpha}_m)])\right|\geq\varepsilon\right\}\right\}\\
&\leq 2^{K+1}N\exp(-2J\varepsilon^2)\longrightarrow 0,
\end{aligned}
$$

provided that $N\exp(-2J\varepsilon^2)\to 0$ as $J\to\infty$. This completes the proof of Theorem 2. $\diamondsuit$

Corollary 2. Under Assumptions 1 and 2 and Conditions (a.1), (a.2), (a.3), (b), and (c), if the test length J and sample size N satisfy the relationship that:

$$J^{n_1}\leq N<J^{n_2},\quad 1<n_1<n_2<\infty.$$

Then

$$\lim_{J\to\infty}P\left(\bigcup_{i=1}^{N}\left\{\hat{\boldsymbol{\alpha}}_i\neq\boldsymbol{\alpha}_i\right\},i.o\right)=0.$$

Proof. We only need to prove that:

$$\textstyle\sum_{N=1}^{J^{n_2}}\sum_{J=1}^{\infty}P(B_\varepsilon(N,J))<\infty,$$

and the result will follow from the Borel-Cantelli Theorem like in Corollary 1. Note that

$$\begin{aligned}\sum_{N=1}^{J^{n_2}}\sum_{J=1}^{\infty}P(B_\varepsilon(N,J)) &< \sum_{N=1}^{J^{n_2}}\sum_{J=1}^{\infty}2^K N\exp(-2J\varepsilon^2)\\ &= 2^K\sum_{J=1}^{\infty}(\sum_{N=1}^{J^{n_2}}N)\exp(-2J\varepsilon^2)\\ &= 2^K\sum_{J=1}^{\infty}\frac{1+J^{n_2}}{2}\exp(-2J\varepsilon^2)\end{aligned}$$

Define $\mathrm{f}(J)=\frac{1+J^{n_2}}{2}\exp(-2J\varepsilon^2)$. According to the convergence rule of series,

$$\begin{aligned}\lim_{J\to\infty}\frac{\mathrm{f}(J+1)}{\mathrm{f}(J)} &= \lim_{J\to\infty}\frac{\frac{1+(1+J)^{n_2}}{2}\exp(-2(J+1))\varepsilon^2}{\frac{1+J^{n_2}}{2}\exp(-2J\varepsilon^2)}\\ &= \lim_{J\to\infty}\frac{1+(J+1)^{n_2}}{1+J^{n_2}}\exp(-2\varepsilon^2)\\ &= \exp(-2\varepsilon^2)<1.\end{aligned}$$

$\Rightarrow\sum_{J=1}^{\infty}\frac{1+J^{n_2}}{2}\exp(-2J\varepsilon^2)<\infty$.

Then we have

$$\textstyle\sum_{N=1}^{J^{n_2}}\sum_{J=1}^{\infty}P(A_\varepsilon(N,J))<\sum_{N=1}^{J^{n_2}}\sum_{J=1}^{\infty}P(B_\varepsilon(N,J))<\infty$$

and the proof is complete. ♢

These theorems for consistency of the nonparametric method, assume that one of several possible true models hold, and we have focused on some of the common latent class models for cognitive diagnosis. The purpose was to show that the simple nonparametric method may be used without calibrating a model, no matter which of those models hold. However, essentially the same general condition was used in the proof of each particular model, and here we focus on the most general condition that any cognitive diagnosis model must satisfy for the nonparametric technique to yield consistent classification. The key steps for the proofs of consistency require that Proposition 1, Proposition 2 and Proposition 3 hold, under proper regularity conditions. If we replace the conditions for the model parameters (Conditions (a.1), (a.2) and (a.3)) with the more general condition that involves no model parameters,

Condition ($a^{'}$): $P(Y_j=1|\eta_j=1)>0.5+\delta$, $P(Y_j=1|\eta_j=0)<0.5-\delta$, for some positive number $\delta>0$,

the three propositions still hold. This means that the consistency results (Theorem 1, Corollary 1, Theorem 2 and Corollary 2) still hold for any models which satisfy the Condition ($a^{'}$). The theory of the previous

results essentially utilized this condition, but phrased it in terms of what it required of model parameters. More generally, these conditions on model parameters can be replaced by Condition ($a^{'}$).

1.3 Numerical Studies

1.3.1 Study Design

In this section, we report simulated examples to illustrate finite test length behavior. The simulation conditions are similar to those in Chiu & Douglas (2013) and were formed by crossing test length, the data generation model, and the expected departure from ideal response patterns. For each condition, 1000 subjects were simulated using either DINA or NIDA model. For each data set, $K = 3$ attributes were required and response profiles consisting of $J = 20$ or 40 items were generated. For a much more thorough simulation study that covers more conditions and compares with competing parametric approaches, see Chiu & Douglas (2013).

Two methods were used to generate the attribute profiles. The first sampled attribute patterns, $\boldsymbol{\alpha}$, from a uniform distribution on the 2^K possible values. The second approach utilized a multivariate normal threshold method. Discrete $\boldsymbol{\alpha}$ were linked to an underlying multivariate normal distribution, $MVN(\mathbf{0}_K, \Sigma)$, with covariance matrix, Σ, structured as

$$\Sigma = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix},$$

with $\rho = 0.5$. Let $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iK})^{'}$ denote the K-dimensional vector of latent continuous scores for subject i. The attribute pattern $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iK})^{'}$ was determined by

$$\alpha_{ik} = \begin{cases} 1, & if \quad \theta_{ik} \geq \mathbf{\Phi}^{-1}(\frac{k}{K+1}); \\ 0, & otherwise. \end{cases}$$

The item parameters for the DINA and NIDA models were generated from uniform distributions with left endpoints of 0 and right endpoints, denoted as $\max(s, g)$, either 0.1, 0.3 or 0.5.

The Q-matrices for tests of 20 items with $K = 3$ were designed as in Table 1, and those for tests of 40 items were obtained by doubling the length of the Q matrix in Table 1. For the simulation with misspecified Q-matrices, 10% or 20% of misspecified Q entries were randomly arranged in the Q-matrix for

each replication.

Table 1.1: Q-matrices for test of 20 items

Attribute	Item																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	0	0	1	1	0	1	0	0	1	1	0	1	0	0	1	1	0	1	1
2	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	1
3	0	0	1	0	1	1	0	0	1	0	1	1	0	0	1	0	1	1	1	1

1.3.2 Results

Results are summarized by an index called pattern-wise agreement rate (PAR), denoting the proportion of attribute patterns accurately estimated according to $PAR = \sum_{i=1}^{N} \frac{I[\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i]}{N}$. The nonparametric estimator based on minimizing Hamming distance can result in some ties, and these ties were randomly broken, though there might be room for developing a more sophisticated technique.

In the first example, we investigate the impact of the model parameters and test length on the PAR of the nonparametric method based on Hamming distance, when actual data were generated from the DINA model and the NIDA model.

Table 1.2: Classification rates for the nonparametric method with DINA data

$\max(s, g)$	$J = 20$	$J = 40$
	Uniform Attribute Patterns	$K = 3$
0.1	0.9925	1.000
0.3	0.9200	0.9825
0.5	0.8100	0.8850
	Multivariate Normal Attribute Patterns	$K = 3$
0.1	0.9900	0.9975
0.3	0.9500	0.9950
0.5	0.7050	0.8675

Table 1.2 documents the effectiveness of the nonparametric method when applied to responses generated from the DINA model. In support of the theoretical results, this approach produces nearly perfect classifications when the slipping and guessing parameter are less than 0.1. As the item parameters become close to 0.5, the classification become worse, but much better than random assignment, which would have an expected classification rate of 0.125 when $K = 3$. Consistent with the asymptotic theory, classification

rates clearly improve as test length increases. Table 1.3 below presents the results when the responses were generated from the NIDA model. Note that the conditions for consistency are somewhat different for the NIDA (Condition (a.2)), and could be violated for several items, in the case where s and g parameters are allowed to be as large as 0.5.

Table 1.3: Classification rates for the nonparametric method with NIDA data

$\max(s, g)$	$J = 20$	$J = 40$
	Uniform Attribute Patterns	$K = 3$
0.1	0.9950	1.000
0.3	0.8225	0.8725
0.5	0.6775	0.8023
	Multivariate Normal Attribute Patterns	$K = 3$
0.1	0.9925	0.9990
0.3	0.8900	0.9125
0.5	0.4650	0.4800

Next we consider the robustness of the nonparametric method when several entries of Q-matrix are misspecified. In each replication, 10% or 20% of the entries in a given Q-matrix were randomly changed. The misspecified Q-matrix was used for classifying examinees with the nonparametric method. Table 1.4 reports the results for the DINA data with attribute patterns generated from a uniform distribution.(The results when the attribute patterns generated from multivariate normal distribution have similar patterns thus omit here.) Table 1.4 shows that classification agreement decreases with the rate of misspecification. However, as the test length increases, the correct classification rate still increases. In all cases, classification rates are well above random assignment. Table 1.5 show the similar results for NIDA data. Though it requires theoretical proof, we speculate that for a certain range of the misspecified percent of entries in Q-matrix, the consistency theories may still hold.

Table 1.4: Classification results for the nonparametric methods with DINA data and a uniform distribution on $\boldsymbol{\alpha}$ when Q is misspecified

$\max(s, g)$	$J = 20$	$J = 40$
	10% misspecified q entries	$K = 3$
0.1	0.9125	0.9650
0.3	0.7825	0.8775
0.5	0.4625	0.8021
	20% misspecified q entries	$K = 3$
0.1	0.7231	0.8024
0.3	0.6621	0.7642
0.5	0.4321	0.5861

Table 1.5: Classification results for the nonparametric method with NIDA data and a uniform distribution on $\boldsymbol{\alpha}$ when Q is misspecified

$\max(s, g)$	$J = 20$	$J = 40$
	10% misspecified q entries	$K = 3$
0.1	0.9125	0.9550
0.3	0.7123	0.8275
0.5	0.4232	0.4532
	20% misspecified q entries	$K = 3$
0.1	0.7135	0.7824
0.3	0.6213	0.6120
0.5	0.3346	0.4032

1.3.3 Discussion

The consistency results of the previous section demonstrate that nonparametric classification can be effective under a variety of underlying conjunctive models. This can greatly expand potential applications, by allowing for conducting cognitive diagnosis when calibration of a parametric model is not feasible. Nonparametric classification based on minimizing Hamming distance to ideal response patterns is simple and fast and can be used with a large number of attributes. One advantage over parametric modeling is that no model calibration is needed, and it can be performed with a sample size as small as 1. Requiring no large samples

or calibration allows for small scale implementation, such as in the classroom setting, where diagnosis can be most important.

The appealing property of the nonparametric method, is that it is consistent under a variety of possible true parametric models, and can be viewed as robust in that sense. Here we studied the properties of the classifier under the DINA, NIDA, and RED-RUM models, but consistency is not be restricted to them, as a conjunctive response process and knowledge of the Q matrix are the critical assumptions. As discussed following the theoretical results, the only general condition required of the underlying item response functions, is that the probability of a correct response for masters of the attributes is bounded above 0.5 for each item, and the probability for non-masters is bounded below 0.5. If the true model satisfies these simple conditions, nonparametric classification will be consistent as the test length increases.

Though consistency results were demonstrated, Chiu and Douglas (2013) show that maximum likelihood estimation with the correct parametric model is more efficient, which can be expected. Other advantages of using parametric statistical models is that one can use general statistical techniques for goodness-of-fit, model selection, and gain a sense of variability and the chance of errors. For instance, classification using parametric latent class models for cognitive diagnosis allows one to compute posterior probabilities for any attribute pattern, which cannot be done with the nonparametric classifier.

Nevertheless, the impressive relative efficiency (Chiu and Douglas, 2013) of the nonparametric classifier and its consistency properties suggest that the approach may be a useful alternative when calibration of a parametric model is not feasible. This approach can be implemented as soon as an item bank with a corresponding Q matrix has been developed, and the computational simplicity allows one to construct reliable computer programs for classification that amount to exhaustively searching through all possible patterns, which is guaranteed to identify an optimal solution. Some promising directions for future research in nonparametric classification include development of fit indices and algorithms for computerized adaptive testing. Another ongoing issue for future research is identifying or validating the correct specification of the Q matrix. In the numerical study we see the effect of misspecification rate on performance. Incorrect Q matrix entries affect both parametric and nonparametric techniques, and suggest that robust methods could be a fruitful area of research.

Chapter 2

Model Misspecification in Cognitive Diagnosis

2.1 Introduction

In cognitive diagnosis, many models have been developed to provide a profile defining mastery or non-mastery of a set of predefined skills or attributes. Based on the assumptions about how attributes influence test performance, CDMs can be categorized as noncompensatory models or compensatory models. Chapter 1.1.1 reviews several common conjunctive models, which express the notion that all attributes specified in the attribute-by-item Q matrix for an item should be required to answer the item correctly, but allow for slips and guesses in ways that distinguish the models from one another. Disjunctive models define the probability of a correct response such that mastering a subset of the attributes is sufficient to have high probability of a correct response. An example is the Deterministic Input, Noisy "Or" gate model (DINO; Templin and Henson (2006)). Unlike noncompensatory models, compensatory models allow an individual to compensate for what is lacked in some measured skills by having mastered other skills. Some representatives are the General Diagnostic Model (GDM; Davier (2005)) and a special case of the GDM called the compensatory RUM (Hartz (2002)). Encompassing the traditional categories of the reduced CDMs, several general models based on different link functions (e.g., G-DINA; De La Torre (2011) and log-linear CDM; Henson et al. (2009)) have been developed to include many of the common compensatory and noncompensatory models.

Starting the analysis with a general model obviates the need to identify the specific form of the CDM, and a general model also has better model fit compared with other reduced models. However, specific and simple models may be preferred under some circumstances for several reasons. First, simple models have more straightforward interpretations (De La Torre et al., 2015). Second, they require smaller sample sizes to be estimated accurately (Rupp and Templin, 2008b). Finally, appropriate reduced models can sometimes provide better classification rates than general models, particularly when the sample size is small (Rojas et al. (2012)). This phenomenon is essentially the bias-variance tradeoff, and is familiar in regression when we can sometimes used a small but biased model for more efficient prediction, for example. When selecting a model, which is surely never the true model, a natural question to consider is how accurately the misspecified model

can classify examinees, which is the ultimate objective. In particular, we are concerned with the behavior of the maximum likelihood estimator, when the likelihood is computed assuming an incorrect model.

The objective of this chapter is to investigate the consequences of using a simple but incorrect model. We are interested in the impact of model misspecification to the classification accuracy, which can be investigated by analyzing the asymptotic behavior of the MLE for attribute profiles under misspecified CDMs. The asymptotic behavior of the MLE under misspecified models associated with parameters in a continuous space has been studied for nearly 50 years (Huber, 1967; Berk et al., 1966; White, 1982; Wainer and Wright, 1980). However, CDMs are in a discrete classification space, and the theory for estimation of discrete attribute profiles is different from that of the estimation of a parameter in a continuous parameter space. In the CDM framework, several empirical studies have been conducted to investigate the effect of model misspecification or Q matrix misspecification on parameter estimation and classification accuracy (Rupp and Templin, 2008a; Chen et al., 2013). However, no theory of the asymptotic behavior of the MLE for attribute profile under misspecified CDMs has been developed.

Specifically, our research questions include: 1) What is the behavior of MLE for attribute profile when the true model and misspecified model both are conjunctive models? 2) Under what conditions is the MLE for attribute profiles a consistent estimator, when the true model satisfying certain conditions, including both conjunctive models and compensatory models, is misspecified as a DINA model? 3) If the MLE is not consistent, can a robust estimation procedure be developed to solve the inconsistency issue under a misspecified DINA model? Investigating these questions can provide practitioners some guidelines about in what situations they can choose a simple model without severely affecting the classification accuracy.

2.2 Behavior of MLE in Misspecified Models

In this section, we investigate whether the MLE of attribute profiles can provide consistent classification when the model is misspecified as a conjunctive model. Two examples about the behavior of MLE under a misspecified DINA model are given.

2.2.1 Problem Formulation

Throughout this section, we focus on the case of a single examinee with true attribute profile denoted as $\boldsymbol{\alpha}^*$. Suppose this examinee finishes a test with J items which measure K attributes. Let $\mathbf{Y} = (Y_1, Y_2, ..., Y_J)^T$ be the examinee's item response vector. The true attribute pattern in this case can be written as $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_K^*)^T$, and each α_k^* takes a value of either 0 or 1 for $k = 1, 2, ..., K$. Specifically, α_k^* is an indicator

of whether this examinee possesses the k^{th} attribute.

We assume that the Q matrix is the same for the true model and the misspecified model, which is a $J \times K$ matrix with entry q_{jk} to indicate whether item j requires the k^{th} attribute. Define $\eta_j(\boldsymbol{\alpha}^*) = \prod_{k=1}^{K}(\alpha_k^*)^{q_{jk}}$ to be the j^{th} component of the ideal response pattern for this examinee, and let $\boldsymbol{\eta}(\boldsymbol{\alpha}^*) = (\eta_1(\boldsymbol{\alpha}^*), ..., \eta_J(\boldsymbol{\alpha}^*))^T$ denote this pattern. For an attribute pattern $\boldsymbol{\alpha}$, we denote the true item response function as $\pi_j^*(\boldsymbol{\alpha})$, $j = 1, 2, ..., J$. Based on the true model, the correct response probability for item j given any attribute profile $\boldsymbol{\alpha}$ can be written as:

$$\pi_j^*(\boldsymbol{\alpha}) = P^*(Y_j = 1|\boldsymbol{\alpha}) = \begin{cases} \pi_{j1}^*(\boldsymbol{\alpha}), & \text{if} \quad \eta_j(\boldsymbol{\alpha}) = 1; \\ \pi_{j2}^*(\boldsymbol{\alpha}), & \text{if} \quad \eta_j(\boldsymbol{\alpha}) = 0. \end{cases}$$

In reality, the true model is not known, and after we collect data and we believe the responses come from a conjunctive process, then an appropriate conjunctive model needs to be selected. Denote $\pi_j(\boldsymbol{\alpha})$ as the item response function for the model we choose, and different conjunctive models are defined through different forms of $\pi_j(\boldsymbol{\alpha})$. Three types of conjunctive models are introduced here.

Example 1 (The DINA model)**.** *The DINA model is the simplest conjunctive model, and the item response function is totally determined by $\eta_j(\boldsymbol{\alpha})$. Thus there are only two types of correct response probabilities for each item under the DINA model. $\forall \boldsymbol{\alpha}$,*

$$\pi_j(\boldsymbol{\alpha}) = P(Y_j = 1|\boldsymbol{\alpha}) = \begin{cases} \pi_{j1}(\boldsymbol{\alpha}) = \pi_{j1} = 1 - s_j \quad , & \textit{if} \quad \eta_j(\boldsymbol{\alpha}) = 1; \\ \pi_{j2}(\boldsymbol{\alpha}) = \pi_{j2} = g_j \quad , & \textit{if} \quad \eta_j(\boldsymbol{\alpha}) = 0. \end{cases}$$

Example 2 (The NIDA model)**.** *Different from the DINA model, the slipping and guessing parameters for the NIDA model are defined through the attribute levels. Define $H_j = \{k|q_{jk} = 1\}$, for the NIDA model,*

$$\pi_j(\boldsymbol{\alpha}) = P(Y_j = 1|\boldsymbol{\alpha}) = \begin{cases} \pi_{j1}(\boldsymbol{\alpha}) = \pi_{j1} = \prod_{k \in H_j}(1 - s_k) \quad , & \textit{if} \quad \eta_j(\boldsymbol{\alpha}) = 1; \\ \pi_{j2}(\boldsymbol{\alpha}) = \prod_{k \in H_j}(1 - s_k)^{\alpha_k} g_k^{1-\alpha_k} \quad , & \textit{if} \quad \eta_j(\boldsymbol{\alpha}) = 0. \end{cases}$$

Example 3 (The Reduced RUM model)**.** *The Reduced RUM model can be generalized from the NIDA model with the item response function defined as:*

$$\pi_j(\boldsymbol{\alpha}) = P(Y_j = 1|\boldsymbol{\alpha}) = \begin{cases} \pi_{j1}(\boldsymbol{\alpha}) = \pi_{j1} = \pi_j \quad , & \textit{if} \quad \eta_j(\boldsymbol{\alpha}) = 1; \\ \pi_{j2}(\boldsymbol{\alpha}) = \pi_j \prod_{k \in H_j}(r_{jk}^*)^{1-\alpha_k} \quad , & \textit{if} \quad \eta_j(\boldsymbol{\alpha}) = 0. \end{cases}$$

Note that for the NIDA model and the Reduced RUM model, π_{j1} is the same for different $\boldsymbol{\alpha}$ such that

$\eta_j(\boldsymbol{\alpha}) = 1$, however $\pi_{j2}(\boldsymbol{\alpha})$ can be different for different $\boldsymbol{\alpha}$. Unlike the DINA model, for each item there can be more than two types of correct response probabilities if $K \geq 1$.

With the assumption of conditional independence, the likelihood function of $\boldsymbol{\alpha}$ constructed from the chosen model based on J questions is:

$$L^{(J)}(\boldsymbol{\alpha}) = \prod_{j=1}^{J} (\pi_j(\boldsymbol{\alpha}))^{Y_j} (1 - \pi_j(\boldsymbol{\alpha}))^{(1-Y_j)}.$$

Note that the superscript (J) indicates the corresponding function depends on J, and we will use such notation from now on. The corresponding log-likelihood function can be written as:

$$l^{(J)}(\boldsymbol{\alpha}) = \log(L^J(\boldsymbol{\alpha})) = \sum_{j=1}^{J} \left[Y_j \log(\pi_j(\boldsymbol{\alpha})) + (1 - Y_j) \log(1 - \pi_j(\boldsymbol{\alpha})) \right].$$

When our selected model is different from the true model, we wish to investigate whether the MLE for $\boldsymbol{\alpha}^*$ arising from the wrong model is consistent. Let $A = \{\boldsymbol{\alpha}^h \neq \boldsymbol{\alpha}^*, h = 1, 2, ..., 2^K - 1\}$ denote the set of the $2^K - 1$ alternative attribute patterns. Our problem is to consider if

$$\lim_{J \to \infty} P\Big(\max_{\boldsymbol{\alpha}^h \in A} l^{(J)}(\boldsymbol{\alpha}^h) < l^{(J)}(\boldsymbol{\alpha}^*) \Big) = 1. \tag{2.1}$$

To investigate whether (2.1) can hold in the model misspecification case, we use a condition on the identifiability of attribute patterns afforded by the Q-matrix, to eliminate a trivial cause for inconsistency.

Condition (I): Define $A_{m,m'} = \Big\{ j | \eta_j(\boldsymbol{\alpha}^m) \neq \eta_j(\boldsymbol{\alpha}^{m'}) \Big\}$, where m and m' index different attribute patterns among the 2^K possible patterns. Then $\liminf_{J \to \infty} \frac{\text{Card}(A_{m,m'})}{J} > 0$.

This condition guarantees that as the test length grows, more and more information becomes available to distinguish attribute patterns from one another. Though the actual rate need not be as the same order of J, a condition of this nature would be needed for consistency, even when using the true model.

To simplify the argument in the next section, we use the notation $a_n \asymp b_n$ to denote that sequences $\{a_n\}$ and $\{b_n\}$ have the same order. This means that there exists constants $0 < m < M < \infty$ and an integer n_0 such that for all $n > n_0$, $m < |\frac{a_n}{b_n}| < M$.

2.2.2 The asymptotic behavior of MLE under misspecified conjunctive CDMs

In this section, we first discuss the behavior of MLE when the true model and misspecified model both belong to conjunctive categories by using a simple case when the number of attributes $K = 1$. Then we extend the analysis to the general case where $K \geq 1$ and when the true model can be any CDM satisfying

some necessary conditions and is misspecified as a DINA model.

A simple case when $K = 1$

Now let's assume the true model and the misspecified model both are conjunctive models, and there is only one attribute ($K = 1$) associated with J items. The corresponding Q matrix is simply a vector with all elements equal to 1. Suppose the true attribute is $\boldsymbol{\alpha}^*$ for this examinee, and the set A only has one element $\boldsymbol{\alpha}^1$. Because only 1 attribute is required for each item, there are only two types of correct response probabilities for each item no matter what the true conjunctive model is. The possible values for $\boldsymbol{\alpha}$ and the corresponding correct response probabilities for each item under the true model and the misspecified model are summarized in Table 2.1 below:

Table 2.1: Attributes and Item Parameters

	Correct Response Probability	
Attribute Profile	True Model	Misspecified Model
$\boldsymbol{\alpha}$	$\pi_j^*(\boldsymbol{\alpha})$	$\pi_j(\boldsymbol{\alpha})$
0	π_{2j}^*	π_{2j}
1	π_{1j}^*	π_{1j}

The MLE for the attribute pattern under the misspecified model is

$$\hat{\boldsymbol{\alpha}}_{MLE} = \underset{\boldsymbol{\alpha}\in\{0,1\}}{\arg\max} \quad l^{(J)}(\boldsymbol{\alpha}).$$

Furthermore, let E^* denote the expectation under the true model, and define

$$\Delta^{(J)} = E^*[l^{(J)}(\boldsymbol{\alpha}^*) - l^{(J)}(\boldsymbol{\alpha}^1)].$$

$\Delta^{(J)}$ quantifies the expected distance between the log-likelihood value of the true attribute profile and that of the alternative attribute profile under the probability measure from the true model.

For conjunctive models, it is natural to assume that the correct response probability for each item when the examinee masters all the required attributes will be greater than when the examinee does not. So we assume that there exists two positive numbers $\delta^* > 0$ and $\delta^1 > 0$, and for any attribute profile $\boldsymbol{\alpha}^h$ and $\boldsymbol{\alpha}^{h'}$ such that $\eta(\boldsymbol{\alpha}^h) = 1$ and $\eta(\boldsymbol{\alpha}^{h'}) = 0$, the true conjunctive model satisfies,

Condition (T) : $\pi_{j1}^*(\boldsymbol{\alpha}^h) - \pi_{j2}^*(\boldsymbol{\alpha}^{h'}) > \delta^*$,

and the misspecified conjunctive model satisfies,

Condition (M) : $\pi_{j1}(\boldsymbol{\alpha}^h) - \pi_{j2}(\boldsymbol{\alpha}^{h'}) > \delta^1$.

Besides these two conditions, we assume that for the misspecified model, the two types of correct response probabilities are bounded away from 0 and 1. The following theorem describes the behavior $\hat{\boldsymbol{\alpha}}_{MLE}$ in this simplest case.

Theorem 3. *When $K = 1$, for one examinee with true attribute profile $\boldsymbol{\alpha}^*$, suppose both the true model and misspecified model are conjunctive models and satisfy Condition (T) and Condition (M) respectively, and the correctly specified Q-matrix satisfies Condition (I) as the exam length increases. If $0 < \Delta^{(J)} \asymp J$, $\lim_{J\to\infty} P(l^{(J)}(\boldsymbol{\alpha}^*) > l^{(J)}(\boldsymbol{\alpha}^1)) = 1$. If $0 > \Delta^{(J)} \asymp J$, $\lim_{J\to\infty} P(l^{(J)}(\boldsymbol{\alpha}^1) > l^{(J)}(\boldsymbol{\alpha}^*)) = 1$.*

Proof. First rewrite the probability in another format,

$$\begin{aligned}
&P\Big(l^{(J)}(\boldsymbol{\alpha}^1) - l^{(J)}(\boldsymbol{\alpha}^*) > 0\Big) \\
&= P\Big(l^{(J)}(\boldsymbol{\alpha}^1) - l^{(J)}(\boldsymbol{\alpha}^*) - E^*[l^{(J)}(\boldsymbol{\alpha}^1) - l^{(J)}(\boldsymbol{\alpha}^*)] > -E^*[l^{(J)}(\boldsymbol{\alpha}^1) - l^{(J)}(\boldsymbol{\alpha}^*)]\Big) \\
&= P\Big(l^{(J)}(\boldsymbol{\alpha}^1) - l^{(J)}(\boldsymbol{\alpha}^*) - (-\Delta^{(J)}) > \Delta^{(J)}\Big).
\end{aligned}$$

Note that when $K = 1$, there are only two possible situations for the values of $\boldsymbol{\alpha}^*$ and $\boldsymbol{\alpha}^1$: $\boldsymbol{\alpha}^* = 0, \boldsymbol{\alpha}^1 = 1$ and $\boldsymbol{\alpha}^* = 1, \boldsymbol{\alpha}^1 = 0$. We prove the theorem under the two cases respectively.

When $\boldsymbol{\alpha}^* = 0$, $\boldsymbol{\alpha}^1 = 1$, and recall that $\Delta^{(J)} = E^*[l^{(J)}(\boldsymbol{\alpha}^*) - l^{(J)}(\boldsymbol{\alpha}^1)]$, then it's easy to verify that

$$l^{(J)}(\boldsymbol{\alpha}^1) - l^{(J)}(\boldsymbol{\alpha}^*) - (-\Delta^{(J)}) = \sum_{j=1}^{J}(Y_j - \pi_{j2}^*)\left(\log\left(\frac{\pi_{j1}}{\pi_{j2}}\frac{1-\pi_{j2}}{1-\pi_{j1}}\right)\right).$$

Now define $Z_j = (Y_j - \pi_{j2}^*)$, then $0 < |Z_j| < 1$. Based on the conditional independence assumption, we have $Z_1, Z_2, ..., Z_j$ are independent and $E^*[Z_j] = 0, j = 1, 2, ..., J$. Let $C_j := \log(\frac{\pi_{j1}}{\pi_{j2}}\frac{1-\pi_{j2}}{1-\pi_{j1}})$. Based on Condition (M), $\forall j$, C_j is positive and bounded away from 0. Suppose the upper bound is a positive number C.

Then we have

$$P\Big(l^{(J)}(\boldsymbol{\alpha}^1) - l^{(J)}(\boldsymbol{\alpha}^*) > 0\Big) = P\Big(\sum_{j=1}^{J} Z_j C_j > \Delta^{(J)}\Big).$$

To investigate the value of the above equation when $J \to \infty$, we need to discuss the behavior of $\Delta^{(J)}$ in the following two situations.

(a) If $0 < \Delta^J \asymp J$, then for large enough J, there exists a constant $m_0 > 0$ such that $\Delta^{(J)} = m_0 J$. Then

by Hoeffding's inequality we have

$$P\Big(l^{(J)}(\boldsymbol{\alpha}^1) > l^{(J)}(\boldsymbol{\alpha}^0)\Big) < \exp(-\frac{2(\Delta^{(J)})^2}{JC^2}) \sim \exp(-2J\epsilon),$$

where $\epsilon = \frac{m_0^2}{C^2} > 0$. Then (2.1) holds in this situation, and the MLE using the misspecified model is consistent.

(b) If $\Delta^{(J)} \asymp J$ but $\Delta^{(J)} < 0$, then

$$P\Big(l^{(J)}(\boldsymbol{\alpha}^1) > l(\boldsymbol{\alpha}^0)\Big) = P\Big(\sum_{j=1}^{J}(-Z_jC_j) < -\Delta^{(J)}\Big) = 1 - P(\sum_{j=1}^{J}(-Z_jC_j) > -\Delta^{(J)}).$$

By using Hoeffiding's inequality again,

$$P(\sum_{j=1}^{J}(-Z_jC_j) > -\Delta^{(J)}) \leq \exp(-J\epsilon) \to 0, \quad \text{as} \quad J \to \infty$$

Then

$$\lim_{J\to\infty} P(l^{(J)}(\boldsymbol{\alpha}^1) > l^{(J)}(\boldsymbol{\alpha}^0)) = 1.$$

Similarly, when $\boldsymbol{\alpha}^* = 1$ and $\boldsymbol{\alpha}^1 = 0$, repeat the above analysis, and we can get the exact same conclusion. Thus when $\Delta^{(J)} < 0$ and has the same order of J, $\hat{\boldsymbol{\alpha}}_{MLE} = \boldsymbol{\alpha}^1$ with probability approaching 1, and is not a consistent estimator of $\boldsymbol{\alpha}^*$. ♢

Theorem 3 indicates that whether the consistency of $\hat{\boldsymbol{\alpha}}_{MLE}$ holds under model misspefication depends heavily on the quantity $\Delta^{(J)} = E^*[l^{(J)}(\boldsymbol{\alpha}^*) - l^{(J)}(\boldsymbol{\alpha}^1)]$. If $0 < \Delta^{(J)} \asymp J$, then the consistency of MLE under model misspecification can hold. If $\Delta^{(J)} < 0$ and has the same order of J, $\hat{\boldsymbol{\alpha}}_{MLE} = \boldsymbol{\alpha}^1$ with probability approaching 1, and is not a consistent estimator of $\boldsymbol{\alpha}^*$. To see more directly how $\Delta^{(J)}$ is related to the parameters of the true and the misspecified model, and in what situations $\Delta^{(J)}$ can guarantee the consistent estimator under a misspecified model, consider the case when $\boldsymbol{\alpha}^* = 0$ and $\boldsymbol{\alpha}^1 = 1$. Now with this specific example, it's easy to verify that

$$\Delta^{(J)} = \sum_{j=1}^{J}(1-\pi_{j2}^*)\log\left(\frac{1-\pi_{j2}}{1-\pi_{j1}}\right) + \pi_{j2}^*\log\left(\frac{\pi_{j2}}{\pi_{j1}}\right) = \sum_{j=1}^{J}\Delta_j.$$

Let's consider two cases,

A. When $\pi_{j2} \geq \pi_{j2}^*$, because of Condition (M) for the misspecified model we have $\log(\frac{\pi_{j2}}{\pi_{j1}}) < 0$ and

$\log(\frac{1-\pi_{j2}}{1-\pi_{j1}}) > 0$, and they imply that

$$\Delta^{(J)} \geq \sum_{j=1}^{J} \left[(1-\pi_{j2}) \log \left(\frac{1-\pi_{j2}}{1-\pi_{j1}} \right) + \pi_{j2} \log \left(\frac{\pi_{j2}}{\pi_{j1}} \right) \right] = \sum_{j=1}^{J} KL\left(\pi_{j2} || \pi_{j1}\right) > 0,$$

where $KL(\pi_{j2}||\pi_{j1})$ is the K-L divergence between π_{j2} and π_{j1}, and will be strictly positive as long as $\pi_{j2} \neq \pi_{j1}$. In this case, $0 < \Delta^{(J)} \asymp J$ and consistency holds.

B. When $\pi_{j2} < \pi^*_{j2}$, then each component of $\Delta^{(J)}$ can be either negative or positive. For example, when $\pi^*_{j1} = 0.6$, $\pi^*_{j2} = 0.4$, $\pi_{j1} = 0.7$, if $\pi_{j2} = 0.1$, then $\Delta_j = -0.119 < 0$, if $\pi_{j2} = 0.3$, then $\Delta_j = 0.169 > 0$. So, in this case $\Delta^{(J)} > 0$ and whther it has the same order of J can not always be guaranteed. This means that the MLE under the misspecified model may not be a consistent estimator for $\boldsymbol{\alpha}^*$.

Similarly, if $\boldsymbol{\alpha}^* = 1$ and $\boldsymbol{\alpha}^1 = 0$, we can conclude that if $\pi^*_1 \geq \pi_1$, then the consistency for MLE holds, otherwise, the MLE may not be consistent. By using this simple example we can see that when the true model and the misspecified model are both conjunctive models, the MLE of attribute profile under such model misspecification is not necessarily a consistent estimator.

General Situation

In the general situation where there are $K \geq 1$ attributes associated with test, the analysis of the MLE under misspecified conjunctive models is more complicated. However, if the true model is a model which satisfies condition (T) and is misspecified as a DINA model which satisfies condition (M), a sufficient condition for the consistency of the MLE can be found in this situation.

Theorem 4. *When $K \geq 1$, for an examinee with true attribute profile $\boldsymbol{\alpha}^*$, suppose the true model satisfies Condition (T), and we misspecify it as a DINA model which satisfies condition (M). Let the Q-matrix be correctly specified and satisfy Condition (I) as the exam length increases. Furthermore, if the correct response probability under the true model and the misspecified model satisfy: Condition (S): $\forall j$, $0 < \pi^*_{j2}(\boldsymbol{\alpha}^*) \leq \pi_{j2} < \pi_{j1} \leq \pi^*_{j1}(\boldsymbol{\alpha}^*)$, then the MLE for $\boldsymbol{\alpha}^*$ is a consistent estimator.*

Proof. Suppose the response vector for a particular examinee is $\mathbf{Y} = (Y_1, Y_2, ..., Y_J)$. Let the true attribute pattern for this examinee be $\boldsymbol{\alpha}^*$, and denote the ideal response pattern under $\boldsymbol{\alpha}^*$ as $\boldsymbol{\eta}(\boldsymbol{\alpha}^*)$. Let $A = \{\boldsymbol{\alpha}^h, h = 1, 2, ..., 2^K - 1\}$ denote the set of the $2^K - 1$ alternative attribute patterns.

Define $B_m(\boldsymbol{\alpha}) = \{j | \eta_j(\boldsymbol{\alpha}) = m\}, m \in \{0, 1\}$. The DINA log-likelihood can be written as:

$$l^{(J)}(\boldsymbol{\alpha}) \quad = \quad \sum_{j \in B_1(\boldsymbol{\alpha})} l_j(\boldsymbol{\alpha} | \eta_j(\boldsymbol{\alpha}) = 1) + \sum_{j \in B_0(\boldsymbol{\alpha}^h)} l_j(\boldsymbol{\alpha} | \eta_j(\boldsymbol{\alpha}) = 0),$$

we need to prove

$$\lim_{J\to\infty} P\left(\max_{\boldsymbol{\alpha}^h\in A} l^{(J)}(\boldsymbol{\alpha}^h) < l^{(J)}(\boldsymbol{\alpha}^*)\right) = 1. \tag{2.2}$$

Note that

$$P\left(\max_{\boldsymbol{\alpha}^h\in A} l^{(J)}(\boldsymbol{\alpha}^h) < l^{(J)}(\boldsymbol{\alpha}^*)\right) \geq 1 - \sum_{\boldsymbol{\alpha}^h\in A} P\Big(\{l^{(J)}(\boldsymbol{\alpha}^h) > l^{(J)}(\boldsymbol{\alpha}^*)\}\Big).$$

In order to prove equation (2.2), it is equivalent to prove:

$$\lim_{J\to\infty} \sum_{\boldsymbol{\alpha}^h\in A} P\left(\{l^{(J)}(\boldsymbol{\alpha}^h) > l^{(J)}(\boldsymbol{\alpha}^*)\}\right) = 0.$$

Because $|A| < \infty$, it can be further simplified to prove that for any $\boldsymbol{\alpha}^h \in A$,

$$\lim_{J\to\infty} P\Big(\{l^{(J)}(\boldsymbol{\alpha}^h) > l^{(J)}(\boldsymbol{\alpha}^*)\}\Big) = 0.$$

Define set $B_{md}(\boldsymbol{\alpha}^h) = \{j|\eta_j(\boldsymbol{\alpha}^*) = m, \eta_j(\boldsymbol{\alpha}^h) = d\}, m \in \{0,1\}, d \in \{0,1\}$. The relationship of $B_m(\boldsymbol{\alpha}^h)$ and $B_{md}(\boldsymbol{\alpha}^h)$ is:

$$B_m(\boldsymbol{\alpha}^*) = \bigcup_{d\in\{0,1\}} B_{md}(\boldsymbol{\alpha}^h), m \in \{0,1\};, \quad B_d(\boldsymbol{\alpha}^h) = \bigcup_{m\in\{0,1\}} B_{md}(\boldsymbol{\alpha}^h), d \in \{0,1\}.$$

Then

$$\begin{aligned} l^{(J)}(\boldsymbol{\alpha}^h) - l^{(J)}(\boldsymbol{\alpha}^*) &= \sum_{j\in B_{01}(\boldsymbol{\alpha}^h)} \left[Y_j \log\left(\frac{\pi_{j1}(1-\pi_{j2})}{(1-\pi_{j1})\pi_{j2}}\right) - \log\left(\frac{1-\pi_{j2}}{1-\pi_{j1}}\right)\right] \\ &+ \sum_{j\in B_{10}(\boldsymbol{\alpha}^h)} \left[Y_j \log\left(\frac{\pi_{j2}(1-\pi_{j2})}{(1-\pi_{j2})\pi_{j1}}\right) - \log\left(\frac{1-\pi_{j1}}{1-\pi_{j2}}\right)\right] \\ &= I_1^{(J)} + I_2^{(J)} \end{aligned}$$

We need to prove that

$$\lim_{J\to\infty} P\left(I_1^{(J)} + I_2^{(J)} > 0\right) = 0. \tag{2.3}$$

Define $J_1(\boldsymbol{\alpha}^h) = |B_{01}(\boldsymbol{\alpha}^h)|$ and $J_2(\boldsymbol{\alpha}^h) = |B_{10}(\boldsymbol{\alpha}^h)|$. The next step is to prove the above equation in three different situations.

Based on Condition (I) that $\liminf \frac{J_1(\boldsymbol{\alpha}^h)+J_2(\boldsymbol{\alpha}^h)}{J} > 0$, there are three possible cases $J_1(\boldsymbol{\alpha}^h)$ and $J_2(\boldsymbol{\alpha}^h)$.

(1.1) $\lim_{J\to\infty} J_1(\boldsymbol{\alpha}^h) = \infty$ and $\lim_{J\to\infty} J_2(\boldsymbol{\alpha}^h) = \infty$.

(1.2) $\lim_{J\to\infty} J_1(\boldsymbol{\alpha}^h) = \infty$ and $\lim_{J\to\infty} J_2(\boldsymbol{\alpha}^h) < \infty$.

(1.3)$\lim_{J\to\infty} J_1(\boldsymbol{\alpha}^h) < \infty$ and $\lim_{J\to\infty} J_2(\boldsymbol{\alpha}^h) = \infty$.

Proof of (1.1)

$$P\left(I_1^{(J)} + I_2^{(J)} > 0\right) \leq P\left(I_1^{(J)} > 0) + P(I_2^{(J)} > 0\right)$$

Then we only need to prove

$$\lim_{J\to\infty} P\left(I_1^{(J)} > 0\right) = 0, \quad \text{and} \quad \lim_{J\to\infty} P\left(I_2^{(J)} > 0\right) = 0.$$

Using similar techniques as in the proof of Theorem 3, we have

$$P\left(I_1^{(J)} > 0\right) = P\left(I_1^{(J)} - E^*[I_1^{(J)}] > -E^*[I_1^{(J)}]\right).$$

Note that, for $j \in B_{01}(\boldsymbol{\alpha}^h)$, $E^*(Y_j) = P^*(Y_j = 1|\eta_j(\boldsymbol{\alpha}^*) = 0) = \pi^*_{j2}(\boldsymbol{\alpha}^*)$. Then we can get that

$$P\left(I_1^{(J)} > 0\right) = P\left(\sum_{j\in B_{01}} (Y_j - \pi^*_{j2}(\boldsymbol{\alpha}^*)) \left(\log\left(\frac{\pi_{j1}}{\pi_{j2}}\frac{1-\pi_{j2}}{1-\pi_{j1}}\right)\right) > -E^*[I_1^{(J)}]\right),$$

where $-E^*[I_1^{(J)}] = \sum_{j\in B_{01}(\boldsymbol{\alpha}^h)}(1-\pi^*_{j2}(\boldsymbol{\alpha}^*))\log(\frac{1-\pi_{j2}}{1-\pi_{j1}}) + \pi^*_{j2}(\boldsymbol{\alpha}^*)\log(\frac{\pi_{j2}}{\pi_{j1}})$.

Using a similar argument as in the prove of Theorem3, let $C_j = \log(\frac{\pi_{j1}}{\pi_{j2}}\frac{1-\pi_{j2}}{1-\pi_{j1}}))$ and because π_{j1}, π_{j2} are bounded away from 0 and 1, then C_j is a finite number and assume the upper bound is C. In order to use Hoeffding's inequality, we only need to guarantee that $-E^*[I_1^{(J)}]$ is positive and has the same order as $J_1(\boldsymbol{\alpha}^h)$. Note that because $\pi^*_{j2}(\boldsymbol{\alpha}^*) \leq \pi_{j2}$ and $\pi_{j2} < \pi_{j1}$,

$$\begin{aligned}
-E^*[I_1^{(J)}] &= \sum_{j\in B_{01}(\boldsymbol{\alpha}^h)} (1-\pi^*_{j2}(\boldsymbol{\alpha}))\log\left(\frac{1-\pi_{j2}}{1-\pi_{j1}}\right) + \pi^*_{j2}(\boldsymbol{\alpha})\log\left(\frac{\pi_{j2}}{\pi_{j1}}\right) \\
&\geq \sum_{j\in B_{01}(\boldsymbol{\alpha}^h)} (1-\pi_{j2})\log\left(\frac{1-\pi_{j2}}{1-\pi_{j1}}\right) + \pi_{j2}\log\left(\frac{\pi_{j2}}{\pi_{j1}}\right) = \sum_{j\in B_{01}(\boldsymbol{\alpha}^h)} KL\left(\pi_{j2}||\pi_{j1}\right).
\end{aligned}$$

Based on the condition that $\lim_{J\to\infty} J_1(\boldsymbol{\alpha}^h) = \infty$, and by Hoeffiding's inequality we have that

$$P\left(I_1^{(J)} > 0\right) \leq P\left(\sum_{j\in B_{01}} (Y_j - \pi^*_{j2}(\boldsymbol{\alpha}))C > \sum_{j\in B_{01}(\boldsymbol{\alpha}^h)} KL(\pi_{j2}||\pi_{j1})\right)$$
$$\leq \exp\left(\frac{-2\left(\sum_{j\in B_{01}(\boldsymbol{\alpha}^h)} KL(\pi_{j2}||\pi_{j1})\right)^2}{J_1(\boldsymbol{\alpha}^h)C^2}\right) \sim \exp\left(-2J_1(\boldsymbol{\alpha}^h)\epsilon\right),$$

where $\epsilon > 0$. Thus $\lim_{J\to\infty} P\left(I_1^{(J)} > 0\right) = 0$. Similarly, we can get that

$$P\left(I_2^{(J)} > 0\right) \leq \exp\left(\frac{-2\left(\sum_{j\in B_{10}(\boldsymbol{\alpha}^h)} KL(\pi_{j1}||\pi_{j2})\right)^2}{J_2(\boldsymbol{\alpha}^h)C^2}\right) \sim \exp\left(-2J_2(\boldsymbol{\alpha}^h)\epsilon\right),$$

and this completes the proof of (1.1)

Proofs of (1.2) and (1.3)

Here we only discuss (1.2) in detail and the conclusion for (1.3) is found by precisely the same argument. In situation (1.2), $J_2(\boldsymbol{\alpha}^h) < \infty$, so

$$\lim_{J\to\infty} I_2^{(J)} = \lim_{J\to\infty} \sum_{j\in B_{10}(\boldsymbol{\alpha}^h)} \left[Y_j \log\left(\frac{\pi_{j2}(1-\pi_{j2})}{(1-\pi_{j2})\pi_{j1}}\right) - \log\left(\frac{1-\pi_{j1}}{1-\pi_{j2}}\right)\right]$$
$$\leq \lim_{J\to\infty} \sum_{j\in B_{10}(\boldsymbol{\alpha}^h)} \left[\log\left(\frac{\pi_{j2}(1-\pi_{j2})}{(1-\pi_{j2})\pi_{j1}}\right) - \log\left(\frac{1-\pi_{j1}}{1-\pi_{j2}}\right)\right] < \infty.$$

So in order to prove (2.3), we only need to prove $\lim_{J\to\infty} P\left(I_1^{(J)} > 0\right) = 0$. The proof is this case is the same as that in (1.1).

Similarly in situation (1.3), when $J \to \infty$, $I_1^{(J)}$ will be a finite number and in order to prove (2.3), we only need to prove $\lim_{J\to\infty} P\left(I_2^{(J)} > 0\right) = 0$. To summarize, for some positive constant ϵ[1], an upper bound for $P\left(l^{(J)}(\boldsymbol{\alpha}^h) > l^{(J)}(\boldsymbol{\alpha}^*)\right)$ can be simplified according to Table 3.8 below:

Table 2.2: Asymptotic form of the upper bound

Asymptotic Behavior	Form of upper bound
$\lim_{J\to\infty} J_1(\boldsymbol{\alpha}^h) = \infty, \lim_{J\to\infty} J_2(\boldsymbol{\alpha}^h)(\boldsymbol{\alpha}^h) = \infty$	$\exp(-2J_1(\boldsymbol{\alpha}^h)\epsilon) + \exp(-2J_2(\boldsymbol{\alpha}^h)\epsilon)$
$\lim_{J\to\infty} J_1(\boldsymbol{\alpha}^h) = \infty, \lim_{J\to\infty} J_2(\boldsymbol{\alpha}^h) < \infty$	$\exp(-2J_1(\boldsymbol{\alpha}^h)\epsilon)$
$\lim_{J\to\infty} J_1(\boldsymbol{\alpha}^h) < \infty, \lim_{J\to\infty} J_2(\boldsymbol{\alpha}^h) = \infty$	$\exp(-2J_2(\boldsymbol{\alpha}^h)\epsilon)$

[1]The constant in the different exponential bound are different, for simplicity, we just use ϵ to stand for that positive constant in each bound.

Note that at least one of $J_1(\boldsymbol{\alpha}^h)$ and $J_2(\boldsymbol{\alpha}^h)$ will have the same order as J. Thus, we can summarize the upper bound as $\exp\left(-2J^*(\boldsymbol{\alpha}^h)\epsilon\right)$, where $J^*(\boldsymbol{\alpha}^h) \asymp J$ and $\epsilon > 0$. $\diamondsuit$

Condition (T) is a natural assumption for conjunctive models, and Condition (S) is less intuitive, and expresses that consistency will hold if the assumed discrimination of the item is no greater than the true discrimination.

Though Theorem 4 is expressed for conjuctive models, the results can apply more generally. Note that it's also reasonable to assume Condition (T) applies to some compensatory models. For example, when the Compensatory RUM model is the true model, the item response function is:

$$P^*\left(Y_j = 1|\boldsymbol{\alpha}\right) = \frac{\exp\left(\sum_{k=1}^K r_{jk}\alpha_{jk}q_{jk} - p_j\right)}{1 + \exp\left(\sum_{k=1}^K r_{jk}\alpha_{jk}q_{jk} - p_j\right)},$$

where $r_{jk} > 0$.

Define $f_j(\boldsymbol{\alpha}) = \sum_{k=1}^K r_{jk}\alpha_k q_{jk} - p_j$. For $\boldsymbol{\alpha}^1$ and $\boldsymbol{\alpha}^2$ such that $\eta_j(\boldsymbol{\alpha}^1) = 1, \eta_j(\boldsymbol{\alpha}^2) = 0$, the positive coefficients r_{jk} result in the relationship that : $f_j(\boldsymbol{\alpha}^1) > f_j(\boldsymbol{\alpha}^2)$. Because $G(x) = \frac{\exp(x)}{1+\exp(x)}$ is an increasing function, we have $\pi^*_{j1}(\boldsymbol{\alpha}^1) > \pi^*_{j2}(\boldsymbol{\alpha}^2)$.

However, if the true model is a disjunctive model, and is misspecified as a DINA model, we can not get consistent classification. For example, if the DINO model is used, the item response function is determined by

$$\omega_j = 1 - \prod_{k=1}^K (1 - \alpha_k)^{q_{jk}},$$

where $\omega_j = 1$ denotes that the examinee has mastered at least one of the required attributes for the jth item, while $\omega_j = 0$ denotes that one has not mastered any of the required attributes. Given ω_j, the probability of a correct response is defined as

$$P^*\left(Y_j = 1|\omega_j\right) = (1 - s_j)^{\omega_j}\, g_j^{1-\omega_j}.$$

Suppose $K = 3$, for item j, $q_j = c(0, 1, 1)$, $\boldsymbol{\alpha}^1 = c(0, 1, 1)$, $\boldsymbol{\alpha}^2 = c(0, 0, 1)$. Then $\eta_j(\boldsymbol{\alpha}^1) = 1$, $\eta_j(\boldsymbol{\alpha}^2) = 0$. However, $\omega_j(\boldsymbol{\alpha}^1) = \omega_j(\boldsymbol{\alpha}^2) = 1$, which means $\pi^*_{j1}(\boldsymbol{\alpha}^1) = \pi^*_{j2}(\boldsymbol{\alpha}^2)$. Because of such equality, when data arise from a disjunctive model, Conditions (T) cannot be satisfied. Thus, the consistency theorem cannot hold.

2.2.3 Examples of the MLE under a misspecified DINA model

In this section, we provide two examples for the MLE of attribute profile under a misspecified DINA model. Example 1 shows the inconsistency of the MLE when the true model is a DINA model and is misspecified as a DINA model with different item parameters. Example 2 shows the consistency of the MLE when the true model is a compensatory RUM model and is misspecified as a DINA model. In both examples, the consistency is measured by the pattern-wise agreement rate (PAR), which is defined as PAR $= \sum_{i=1}^{N} \frac{I[\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i]}{N}$ to denote the proportion of accurately estimated attribute patterns.

Example 1. Inconsistency of MLE under a misspecified DINA model. In this example, the number of attributes $K = 3$. The true model is a DINA model with fixed slipping and guessing parameters for each item: $s_t = 0.4$, $g_t = 0.4$. In particular, for $\forall j$, $\pi^*_{j1} = 1 - 0.4 = 0.6$, $\pi^*_{j2} = 0.4$. For the misspecified DINA model, the guessing and slipping parameters are fixed at $s_m = 0.3$, $g_m = 0.1$ for each item, violating Condition (S). That is, $\forall j$, $\pi_{j1} = 1 - 0.3 = 0.7$, $\pi_{j2} = 0.1$. To inspect the consistency, a sample of 1000 examinees with true attribute profiles simulated from a uniform distribution for each of 9 levels of test length: $J = 20, 40, 100, 200, 300, 400, 500, 600, 700$. When test length equaled 20 items, a Q matrix which contains 12 items each measuring a single attribute, 6 items each measuring 2 attributes and 2 items each measuring 3 attributes was used. The Q matrices for longer tests were obtained by replicating this Q matrix. For each condition, the examinees' responses were simulated based on the true model, and their attribute profiles were estimated based on the true DINA model and the misspecified DINA model. Results are summarized by the Figure 2.1 below:

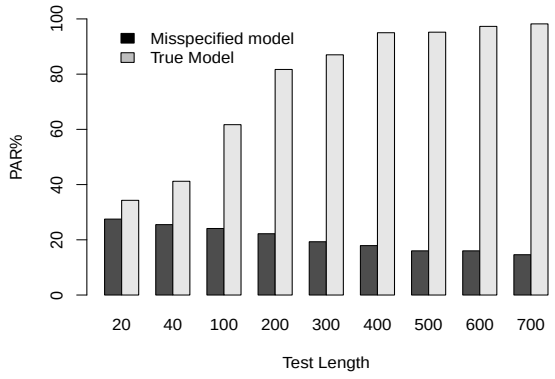

Figure 2.1: Inconsistency of MLE under model misspecification

From Figure 2.1 we can see that, under the misspecified DINA model, the pattern-wise agreement rate decreases as the test length increases, which indicates the MLE is not a consistent estimator in this case.

Example 2: Consistency of the MLE under a misspecified DINA model. In this example, $K = 3$ and the item parameters for the true model, a compensatory RUM model, were generated from uniform distributions on appropriate intervals. In order to make the item parameters satisfy condition (T), the lower bound for p_j is controlled by 0.3 and the upper bound for r_{jk} is controlled by 0.8. A sample of 1000 examinees was generated with true attribute profiles simulated from a uniform distribution for each of the 5 levels of test length: $J = 20, 40, 100, 200, 300$. For each condition, the Q matrix is constructed in the same way as that in Example 1, and the examinees' responses were simulated based on the true model. A DINA model is calibrated by the "CDM" package (Robitzsch et al. (2014)) in R based on the examinees responses in each condition. The results in terms of PAR were summarized in Figure 2.2 below. We can see that, with the increase of the test length, the PAR grows slowly to 100%, which indicates the MLE is a consistent estimator.

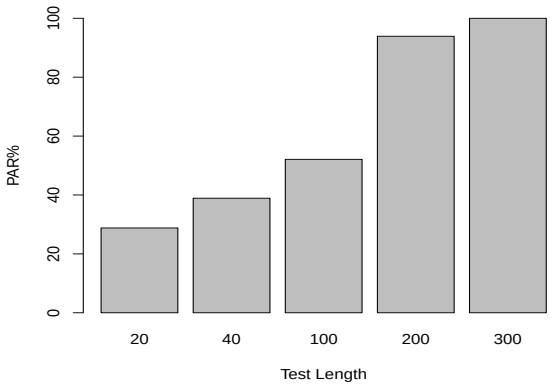

Figure 2.2: Consistency of MLE under model misspecification

2.3 Robust Estimation

In the previous section, we discuss the inconsistency of the MLE for the attribute vector under misspecified conjunctive models. Although a sufficient condition for consistency under a misspecified DINA model has been found, it is difficult to verify whether the calibrated item parameters from a DINA model satisfy Condition (S) or not. This is because one cannot know $\pi_j^*(\boldsymbol{\alpha}^*)$, thus suggesting a need to develop a robust estimator that guarantees consistency of attribute vector classification under model misidentification.

The discussion in Section 2 indicates that the problematic part under model misspecification is that we can not guarantee that $E^*[l^{(J)}(\boldsymbol{\alpha}^*) - l^{(J)}(\boldsymbol{\alpha}^h)]$, $\forall \boldsymbol{\alpha}^h \in A$ is positive and of order J. If some concave function can be incorporated into the log-likelihood function to compensate for this problematic part, then the modified log-likelihood function may guarantee the unique maximum is $\boldsymbol{\alpha}^*$. One solution is to utilize

the loss function of the nonparametric estimation method of Chiu and Douglas (2013), that was shown by Wang and Douglas (2015) to be consistent for the attribute profile under very general conditions. Following some regularity conditions, the loss function for nonparametric estimation method is strictly convex with an expected value that has a unique minimum at $\boldsymbol{\alpha}^*$. Thus, by incorporating a multiple of this loss function with the negative log-likelihood, a robust classifier is defined that has desirable consistency properties, and is more efficient than simply using the nonparametric technique. The nonparametric estimation method can be found in section 1.

2.3.1 Robust DINA MLE

One can regard the MLE as the minimizer of a loss function, which is the negative log-likelihood function. The idea of the Robust DINA MLE is to shrink this loss function towards the distance function of the consistent nonparametric estimation method to ensure consistency while obtaining greater efficiency than using the nonparametric method alone. Suppose now the model is misspecified as a DINA model with the item parameters:

$$\pi_{j1} = 1 - s_j, \pi_{j2} = g_j.$$

For any attribute pattern $\boldsymbol{\alpha}$, the robust likelihood under $\boldsymbol{\alpha}$ is

$$L_r^{(J)}(\boldsymbol{\alpha}) = \prod_{j=1}^{J} (\pi_j(\boldsymbol{\alpha}))^{Y_j} (1 - \pi_j(\boldsymbol{\alpha}))^{1-Y_j} \exp\left(-\lambda_j(\alpha)|Y_j - \eta_j(\boldsymbol{\alpha})|\right),$$

where

$$\lambda_j(\boldsymbol{\alpha}) = \begin{cases} \log(\frac{\pi_{j1}}{\pi_{j2}}) & \text{if} \quad , \quad \eta_j(\boldsymbol{\alpha}) = 1; \\ \log(\frac{1-\pi_{j2}}{1-\pi_{j1}}) & \text{if} \quad , \quad \eta_j(\boldsymbol{\alpha}) = 0. \end{cases}$$

Then the Robust DINA MLE estimator is :

$$\hat{\boldsymbol{\alpha}}_{RMLE} = \underset{\boldsymbol{\alpha}^h \in A \cup \{\boldsymbol{\alpha}^*\}}{\operatorname{argmax}} \quad L_r^{(J)}(\boldsymbol{\alpha}) \tag{2.4}$$

Note that any conjunctive model could be used for the parametric term, and the DINA is merely chosen for its simplicity. The item response function is totally determined by the ideal response pattern and the slipping and guessing parameters for each item. The item response functions are more complicated for other conjunctive models, with item parameters that are generally more difficult to calibrate. As demonstrated

below, the Robust DINA MLE has good statistical properties, and applying the simple DINA model, can always yield relatively efficient classifications of attribute patterns under a wide class of underlying true models. There are some choices for calibrating the slipping and guessing parameters in DINA portion of the model. Based on the consistency theorem discussed in the next section, those parameters can either be randomly simulated from a reasonable interval or calibrated from the original DINA log-likelihood function by using existing software. This greatly simplifies the item parameter calibration procedure. Next, we provide theory to show the consistency of the Robust DINA MLE estimator under model misspecification.

2.3.2 Asymptotic behavior of the Robust DINA MLE

In this section, we first provide the regularity conditions required for consistency, then state and prove the main consistency result for a single examinee, using the Robust DINA MLE. Then we extend the consistency result to a finite sample of subjects.

The true model must satisfy the following conditions: for some known positive small number δ_0,

Condition (T.1) $0.5 + \delta_0 < \pi^*_{j1}(\boldsymbol{\alpha}) \leq 1$,

Condition (T.2) $0 \leq \pi^*_{j2}(\boldsymbol{\alpha}) < 0.5 - \delta_0$.

The corresponding two conditions are required for the misspecified DINA model: there exists positive small numbers δ_1, $\delta_1^{'}$ such that

Condition (M.1) $\pi_{j1} - \pi_{j2} > \delta_1 > 0$

Condition (M.2) $\delta_1^{'} < \pi_{j1}, \pi_{j2} < 1 - \delta_1^{'}$.

We also make the usual assumption that independence of the item response vector given the attribute vector holds for the true model, and item response vectors of different examinees are independent.

Theorem 5. *For an examinee with true attribute vector $\boldsymbol{\alpha}^*$, suppose the true CDM satisfies Conditions (T.1) and (T.2). If a misspecified DINA model is used which satisfies Conditions (M.1) and (M.2), and we correctly specify a Q-matrix that satisfies Condition (I) as the exam length increases, then the Robust DINA MLE estimator is a consistent estimator of $\boldsymbol{\alpha}^*$. Specifically,*

$$\lim_{J\to\infty} P\left(\max_{\boldsymbol{\alpha}^h \in A} L_r^{(J)}(\boldsymbol{\alpha}^h) < L_r^{(J)}(\boldsymbol{\alpha}^*)\right) = 1. \tag{2.5}$$

Proof. The proof of Theorem 2.3.2 using exactly same techniques as those in the proof of Theorem 4, but the structure of the robust log-likelihood is different. Note that the Robust DINA log-likelihood can be

written as:

$$\begin{aligned} l_r^{(J)}(\boldsymbol{\alpha}) &= \log(L_r^{(J)}(\boldsymbol{\alpha})) = \sum_{j=1}^{J} l_{r_j}(\boldsymbol{\alpha}) \\ &= \sum_{j\in B_1(\boldsymbol{\alpha})} l_{r_j}\left(\boldsymbol{\alpha}|\eta_j(\boldsymbol{\alpha})=1\right) + \sum_{j\in B_0(\boldsymbol{\alpha}^h)} l_{r_j}\left(\boldsymbol{\alpha}|\eta_j(\boldsymbol{\alpha})=0\right), \end{aligned}$$

where

$$\begin{aligned} l_{r_j}(\boldsymbol{\alpha}|\eta_j(\boldsymbol{\alpha})=1) &= Y_j\log(\pi_{j1}) + (1-Y_j)\log(1-\pi_{j1}) - \log\left(\frac{\pi_{j1}}{\pi_{j2}}\right)(1-Y_j); \\ l_{r_j}(\boldsymbol{\alpha}|\eta_j(\boldsymbol{\alpha})=0) &= Y_j\log(\pi_{j2}) + (1-Y_j)\log(1-\pi_{j2}) - \log\left(\frac{1-\pi_{j2}}{1-\pi_{j1}}\right)Y_j. \end{aligned}$$

♢

Following the same argument as those in the proof of Theorem 4, we can have that

$$\begin{aligned} l_r^{(J)}(\boldsymbol{\alpha}^h) - l_r^{(J)}(\boldsymbol{\alpha}^*) &= \sum_{j\in B_{01}} l_{r_j}(\boldsymbol{\alpha}^h|\eta(\boldsymbol{\alpha}^h)=1) - \sum_{j\in B_{01}} l_{r_j}(\boldsymbol{\alpha}^*|\eta(\boldsymbol{\alpha}^*)=0) \\ &+ \sum_{j\in B_{10}} l_{r_j}(\boldsymbol{\alpha}^h|\eta(\boldsymbol{\alpha}^h)=0) - \sum_{j\in B_{10}} l_{r_j}(\boldsymbol{\alpha}^*|\eta(\boldsymbol{\alpha}^*)=1) \\ &= I_1^{(J)} + I_2^{(J)}. \end{aligned}$$

In order to prove Theorem , we only need to prove that

$$\lim_{J\to\infty} P\left(I_1^{(J)} + I_2^{(J)} > 0\right) = 0. \tag{2.6}$$

Again, using the same notations as the cardinalities defined in Theorem 4, we need to discuss (2.6) in three situations.

Proof of (1.1)

$$P(I_1^{(J)} + I_2^{(J)} > 0) \le P(I_1^{(J)} > 0) + P(I_2^{(J)} > 0).$$

Then we only need to prove

$$\lim_{J\to\infty} P(I_1^{(J)} > 0) = 0, \quad \text{and,} \quad \lim_{J\to\infty} P(I_2^{(J)} > 0) = 0.$$

Using similar techniques as in the proof of Theorem 3, we have

$$P\left(I_1^{(J)} > 0\right) = P\left(I_1^{(J)} - E^*[I_1^J] > -E^*[I_1^J]\right).$$

Note that for $j \in B_{01}(\boldsymbol{\alpha}^h)$, $E^*(Y_j) = P^*(Y_j = 1|\eta_j(\boldsymbol{\alpha}^*) = 0) = \pi^*_{j2}(\boldsymbol{\alpha}^*)$. Then we can get that

$$P\Big(I_1^{(J)} > 0\Big) = P\Big(\sum_{j\in B_{01}}(Y_j - \pi^*_{j2}(\boldsymbol{\alpha}^*))2\log\left(\frac{\pi_{j1}}{\pi_{j2}}\frac{1-\pi_{j2}}{1-\pi_{j1}}\right) > -E^*[I^{(J)_1}]\Big),$$

where $-E^*[I_1^{(J)}] = \sum_{j\in B_{01}(\boldsymbol{\alpha}^h)}\left[(1-2\pi^*_{j2}(\boldsymbol{\alpha}^*))\log\left(\frac{\pi_{j1}}{\pi_{j2}}\frac{1-\pi_{j2}}{1-\pi_{j1}}\right)\right]$.

If we treat $Z_j = 2(Y_j - \pi^*_{j2}(\boldsymbol{\alpha}^*))\log(\frac{\pi_{j1}}{\pi_{j2}}\frac{1-\pi_{j2}}{1-\pi_{j1}})$ as the new random variable, then

$$0 \le |Z_j| \le 2\log\left(\frac{\pi_{j1}}{\pi_{j2}}\frac{1-\pi_{j2}}{1-\pi_{j1}}\right) \le \log\left((\frac{1}{\delta_1'})^2\right) \le \infty.$$

Furthermore,

$$(1-2\pi^*_{j2}(\boldsymbol{\alpha}^*))\log\left(\frac{\pi_{j1}}{\pi_{j2}}\frac{1-\pi_{j2}}{1-\pi_{j1}}\right) > \quad 2\delta_0\log\left((1+\tfrac{\delta_1}{1+\delta_1'})^2\right) = \epsilon_1 > 0.$$

Define $\epsilon_2 = \log((\frac{1}{\delta_1'})^2)$, then we can apply Hoeffding's inequality to obtain an upper bound for $P(I_1^{(J)} > 0)$.

$$P(I_1^{(J)} > 0) < P\left(\sum_{j\in B_{01}(\boldsymbol{\alpha}^h)}(Y_j - \pi^*_{j2}(\boldsymbol{\alpha}^*))\epsilon_2 > J_1(\boldsymbol{\alpha}^h)\epsilon_1\right) \le \exp\left(-\frac{2(J_1(\boldsymbol{\alpha}^h)\epsilon_1)^2}{J_1(\boldsymbol{\alpha})\epsilon_2^2}\right) = \exp(-2J_1(\boldsymbol{\alpha}^h)\epsilon),$$

where $\epsilon = \frac{\epsilon_1^2}{\epsilon_2^2} > 0$. Thus $P\left(I_1^{(J)} > 0\right) \to 0$, as $J \to \infty$.

Similarly,

$$\begin{aligned} P\left(I_2^{(J)} > 0\right) &= P\left(I_2^{(J)} - E^*[I_2^{(J)}] > -E^*[I_2^{(J)}]\right) \\ &= P\left(\sum_{j\in B_{10}(\boldsymbol{\alpha}^h)}(Y_j - \pi^*_{j1}(\boldsymbol{\alpha}^*))2\log\left(\frac{\pi_{j2}}{\pi_{j1}}\frac{1-\pi_{j1}}{1-\pi_{j2}}\right) > -E^*[I_2^{(J)}]\right), \end{aligned}$$

where $E^*[I_2^{(J)}] = \sum_{j\in B_{10}(\boldsymbol{\alpha}^h)}(2\pi^*_{j1}(\boldsymbol{\alpha}^*) - 1)\log\left(\frac{\pi_{j1}}{\pi_{j2}}\frac{1-\pi_{j2}}{1-\pi_{j1}}\right)$. Note that $\forall j \in B_{10}(\boldsymbol{\alpha}^h)$,

$$\begin{aligned} &\big(2\pi^*_{j1}(\boldsymbol{\alpha}^*) - 1\big)\log\left(\frac{\pi_{j1}}{\pi_{j2}}\frac{1-\pi_{j2}}{1-\pi_{j1}}\right) > 2\delta_0\log\left((1+\frac{\delta_1}{1+\delta_1'})^2\right) = \epsilon_1 > 0, \\ &\left|\log\left(\frac{\pi_{j2}}{\pi_{j1}}\frac{1-\pi_{j1}}{1-\pi_{j2}}\right)\right| < \epsilon_2. \end{aligned}$$

Applying Hoeffdoing's inequality again we see that

$$\begin{aligned} P\left(I_2^{(J)} > 0\right) &< P\left(\sum_{j\in B_{10}(\boldsymbol{\alpha}^h)} (Y_j - \pi^*_{j1}(\boldsymbol{\alpha}^*)) \log\left(\frac{\pi_{j2}}{\pi_{j1}}\frac{1-\pi_{j1}}{1-\pi_{j2}}\right) > J_2(\boldsymbol{\alpha}^h)\epsilon_1\right) \\ &\leq \exp\left(-\frac{2(J_2(\boldsymbol{\alpha}^h)\epsilon_1)^2}{J_2(\boldsymbol{\alpha})\epsilon_2^2}\right) = \exp\left(-2J_2(\boldsymbol{\alpha}^h)\epsilon\right). \end{aligned}$$

Thus $P\left(I_2^{(J)} > 0\right) \to 0$, as $J \to \infty$.

Proofs of (1.2) and (1.3)

Here we only discuss (1.2) in detail and the conclusion for (1.3) is found by precisely the same argument.

In situation (1.2), $J_2(\boldsymbol{\alpha}^h) < \infty$, $\log\left(\frac{\pi_{j2}(1-\pi_{j1})}{(1-\pi_{j2})\pi_{j1}}\right) < 0$, $\log\left(\frac{1-\pi_{j2}}{1-\pi_{j1}}\frac{\pi_{j1}}{\pi_{j2}}\right) > 0$,

$$\begin{aligned} I_2^{(J)} &= \sum_{j\in B_{10}(\boldsymbol{\alpha}^h)} \left[Y_j * 2\log\left(\frac{\pi_{j2}(1-\pi_{j1})}{(1-\pi_{j2})\pi_{j1}}\right) + \log\left(\frac{1-\pi_{j2}}{1-\pi_{j1}}\right)\right] \\ &\leq \sum_{j\in B_{10}(\boldsymbol{\alpha}^h)} \log\left(\frac{1-\pi_{j2}}{1-\pi_{j1}}\frac{\pi_{j1}}{\pi_{j2}}\right) < \sum_{j\in B_{10}(\boldsymbol{\alpha}^h)} 2\log\left(\frac{1}{\delta_1'}\right) \\ &\leq \lim_{J\to\infty} J_2(\boldsymbol{\alpha}^h) 2\log\left(\frac{1}{\delta_1'}\right) = C_1 < \infty \end{aligned}$$

Then we only need to discuss $P\left(I_1^{(J)} > 0\right)$ when $J \to \infty$. Using the same technique as in the proof of Theorem 4 we have,

$$P(I_1^{(J)} > 0) \sim \exp(-2J_1(\boldsymbol{\alpha}^h)\epsilon),$$

where $\epsilon > 0$ is the same form as that in (1.1). So we conclude that $P\left(I_1^{(J)} + I_2^{(J)} > 0\right)$ will converge to 0 as $J \to \infty$.

In situation (1.3), define $C_2 = \lim_{J\to\infty} 2J_1(\boldsymbol{\alpha}^h)2\log\left(\frac{1}{\delta_1'}\right) < \infty$, we have

$$P\left(I_1^{(J)} + I_2^{(J)} > 0\right) \sim \exp\left(-2J_2(\boldsymbol{\alpha}^h)\epsilon\right).$$

To summarize, we can get the exact same upper bounds for $P\left(l_r^{(J)}(\boldsymbol{\alpha}^h) > l_r^{(J)}(\boldsymbol{\alpha}^*)\right)$ as those in Table 3.8.

Theorem 6. *For a sample of N subjects, suppose the true CDM satisfies Conditions (T.1) and (T.2). If we use a possibly misspecified model which satisfies Conditions (M.1) and (M.2), and a correctly specified*

Q-matrix that satisfies Condition (I) as the exam length increases, then

$$\lim_{J\to\infty} P\left(\bigcap_{i=1}^{N}\left\{\max_{\boldsymbol{\alpha}^h\in A} L_r^{(J)}(\boldsymbol{\alpha}^h) < L_r^{(J)}(\boldsymbol{\alpha}_i^*)\right\}\right) = 1$$

Proof. With the same techniques as those used for Theorem 2.3.2, it is equivalent to prove that

$$\lim_{J\to\infty} P\left(\bigcup_{i=1}^{N}\bigcup_{\boldsymbol{\alpha}^h\in A}\left\{l_r^{(J)}(\boldsymbol{\alpha}^h) > l_r^{(J)}(\boldsymbol{\alpha}_i^*)\right\}\right) \leq \lim_{J\to\infty}\sum_{i=1}^{N}\sum_{\boldsymbol{\alpha}^h\in A} P\left(\left\{l_r^{(J)}(\boldsymbol{\alpha}^h) > l_r^{(J)}(\boldsymbol{\alpha}_i^*)\right\}\right) = 0$$

Because N is a finite sample size, not increasing with J, the above equation is always true based on a direct application of Theorem 2.3.2. ♢

Remark 1. *Note that the upper bound for* $P\left(L_r^{(J)}(\boldsymbol{\alpha}^h) > L_r^{(J)}(\boldsymbol{\alpha}^*)\right)$ *converges to* 0 *exponentially fast, thus a strong consistency result can easily be achieved by applying the Borel-Cantelli Lemma.*

The proof of Theorem 2.3.2 indicates that conditions (T.1) and (T.2) are the key factors to guarantee the consistency of the Robust DINA MLE. For conjunctive models, the probability of a correct response when subjects have mastered all the required attributes will always be larger than that when they have not. If the true model satisfies conditions (T.1) and (T.2), and the misspecified DINA model satisfies conditions (M.1) and (M.2), the standard independence assumption and the condition for the identifiability of Q matrix guarantee that we can always get a consistent classification result by using the Robust DINA MLE. With similar arguments as those in Section 2.2.2, the consistency may sill hold if the true model is a compensatory model, but the 0.5 boundary may be too restrictive for most compensatory models. If the true model is a disjunctive model, this theorem will not hold.

2.4 Simulation

2.4.1 Simulation 1

To illustrate how the proposed Robust DINA MLE can solve the inconsistency problem for the MLE under a misspecified DINA model, a simulation based on the Example 1 in Section 2.2.3 was conducted. The Robust DINA log-likelihood was constructed based on the same wrong parameters for the misspecified model as were given in the inconsistency example. We summarize the PAR of those two methods under each test length.

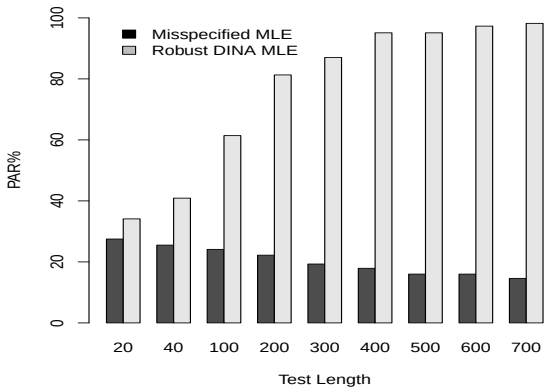

Figure 2.3: Consistency of Robust DINA MLE

From Figure 2.3 we can see that the PAR from the Robust DINA MLE converges to 100% as the test length increases. This indicates the Robust DINA MLE is a consistent estimator in this case.

2.4.2 Simulation 2

Because the nonparametric classifier is also a consistent estimator for the attribute profile, we can compare the robust DINA MLE with it in terms of the estimation efficiency. Furthermore, although there is an example which shows the MLE estimator under a misspecified DINA model is not necessarily consistent, it is still worthwhile to investigate the behavior of the MLE in some model misspecification cases.

Simulation Design

The simulation conditions were formed by crossing test length, the data generation model, and the deviation of $\pi_{j1}(\boldsymbol{\alpha}^*)$ and $\pi_{j2}(\boldsymbol{\alpha}^*)$ from 0.5 (denoted as δ_0). The DINA, the NIDA, and the Reduced-RUM models were chosen as the true models to generate examinees' responses. Responses were simulated using the DINA, the NIDA, and the Reduced-RUM. For each dataset, $K = 3$ or $K = 5$ attributes were required and response profiles consisting of $J = 20, 40, 100$ or 200 items were sampled for N = 1000 subjects. The first two test lengths were chosen to study the behavior under common conditions, and the larger two test lengths were included to provide an empirical examination of whether consistency appears to hold. The attribute pattern $\boldsymbol{\alpha}_i^*$ for $i = 1, 2, ..., 1000$ were determined based on a multivariate normal threshold model (Chiu and Douglas, 2013) to mimic a realistic situation where attributes are correlated and of unequal prevalence. Specifically, the discrete $\boldsymbol{\alpha}$ were linked to an underlying multivariate normal distribution, $MVN(\mathbf{0}_K, \Sigma)$, with covariance

matrix,

$$\Sigma = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}$$

, and $\rho = 0.5$. Let $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iK})^{'}$ denote the K- dimensional vector of latent continuous scores for examinee i. The attribute pattern $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iK})$ was determined by

$$\alpha_{ik} = \begin{cases} 1, & \text{if} \quad \theta_{ik} \geq \Phi^{-1}(\frac{k}{K+1}); \\ 0, & \text{otherwise.} \end{cases}$$

The item parameters for the three models were generated from uniform distributions on appropriate intervals. The 0.5 boundary for $\pi_{j1}(\boldsymbol{\alpha}^*)$ and $\pi_{j2}(\boldsymbol{\alpha}^*)$ is the key condition to get the consistency results for robust DINA MLE estimation and the non-parametric estimation method under model misspecification. As two types of correct response probabilities grow in distance from 0.5, quantified by δ_0, we expect greater efficiency in classification. Here we regard relative efficiency as how correct classification rates compare across two methods or two conditions. To investigate the performance with different δ_0 values, we simulated the item parameters from 3 different true models and control δ_0 as "small", "medium" and "large". For the DINA model, the three types of values for δ_0 can be achieved by setting the upper bound of s_j and g_j as 0.5, 0.3, and 0.1. For the Reduced RUM model, we control the upper bound of r^*_{jk} as 0.3, and let the lower bound of π_j vary from 0.5, 0.7 and 0.9. For the NIDA model, the lower bound for g_k is controlled by 0.5. To guarantee $\pi^*_{j1}(\boldsymbol{\alpha}) > 0.5$, for any $\boldsymbol{\alpha}$, $(1 - s_k)^K$ should be at least greater than 0.5. Because of this product effect, the lower bound of s_k varies from 0.2, 0.15, and 0.1 when $K = 3$, and 0.12, 0.07 and 0.09 when $K = 5$ to represent δ_0 as "small","medium" and "large".

No matter what the true model is, for parametric estimation methods, the DINA model is always used to estimate the attribute pattern. Two methods were used to obtain the DINA item parameters. One was to use "CDM" package in R to calibrate DINA item parameters, the other was to randomly simulate the guessing and slipping parameters for each item from a uniform distribution with lower bound as 0.01 and upper bound as 0.3.

The Q-matrices for tests of 20 items with $K = 3$ and 5 were designed as in Table 2.3 and Table 2.4, and for those for tests of 40, 100 or 200 items were obtained by replicating the Q matrix in Table 2.3 and Table 2.4.

Table 2.3: Q matrix, $K = 3$

	Item																			
Attribute	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	0	0	1	1	0	1	0	1	1	1	0	1	1	0	1	1	0	1	1
2	0	1	0	1	0	1	1	1	0	1	0	1	1	1	1	1	0	1	1	1
3	0	0	1	0	1	1	0	1	1	0	1	1	1	0	1	0	1	1	1	1

Table 2.4: Q matrix, $K = 5$

	Item																			
Attribute	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	0	1	0	0	1	1	0	1	1	0	1	0	1	0	0	0	1	1
2	0	0	0	0	1	1	0	1	0	0	0	1	0	1	0	1	1	1	1	1
3	0	1	0	0	0	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1
4	0	0	1	0	0	1	1	0	1	1	0	0	0	0	1	0	1	0	0	1
5	1	0	0	0	0	0	0	1	1	1	1	1	0	1	1	0	1	0	1	0

Results

Results are summarized by pattern-wise agreement rate (PAR) and attribute-wise agreement rate (AAR) to reflect the agreement between estimated attribute profiles and the known true attribute profiles. PAR is defined as PAR $= \sum_{i=1}^{N} \frac{I[\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i]}{N}$ to denote the proportion of accurately estimated attribute patterns. AAR refers the proportion of individual attributes that were classified correctly, and is defined as AAR $= \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{I[\hat{\alpha}_{ik} = \alpha_{ik}]}{NK}$. The nonparametric estimator based on minimizing Hamming distance can result in some ties, and these ties were randomly broken, though there might be room for developing a more sophisticated technique. In order to see a clear trend, we first present the results for PAR when $K = 5$. The barplot in Figure 2.4 documents PARs from the MLE, the robust DINA MLE and the non-parametric estimation method when the data were generated from the DINA model. Conditioning on the same deviation δ_0, the PARs from the three methods both increase with test length. The larger the two correct response probabilities deviate from 0.5, the higher the classification rates for the three methods under the same test length. Besides those general trends, several interesting results can be concluded from Figure 2.4: 1) When test length equaled 20 or 40, the robust DINA MLE method performed very well in most situations, with classification rates nearly indistinguishable from the calibrated DINA model. 2) For the DINA MLE method, the results based on the simulated item parameters were quite comparable to those based on the calibrated

item parameters. For the Robust DINA MLE method, results based on the simulated item parameters were slighter worse than those based on the calibrated item parameters. 3) When the test length was long enough (100 or 200), all three methods had similar performance and the PARs were all close to 100%. Especially, for the two parametric methods, the results based on the calibrated item parameters and simulated parameters had similar performances.

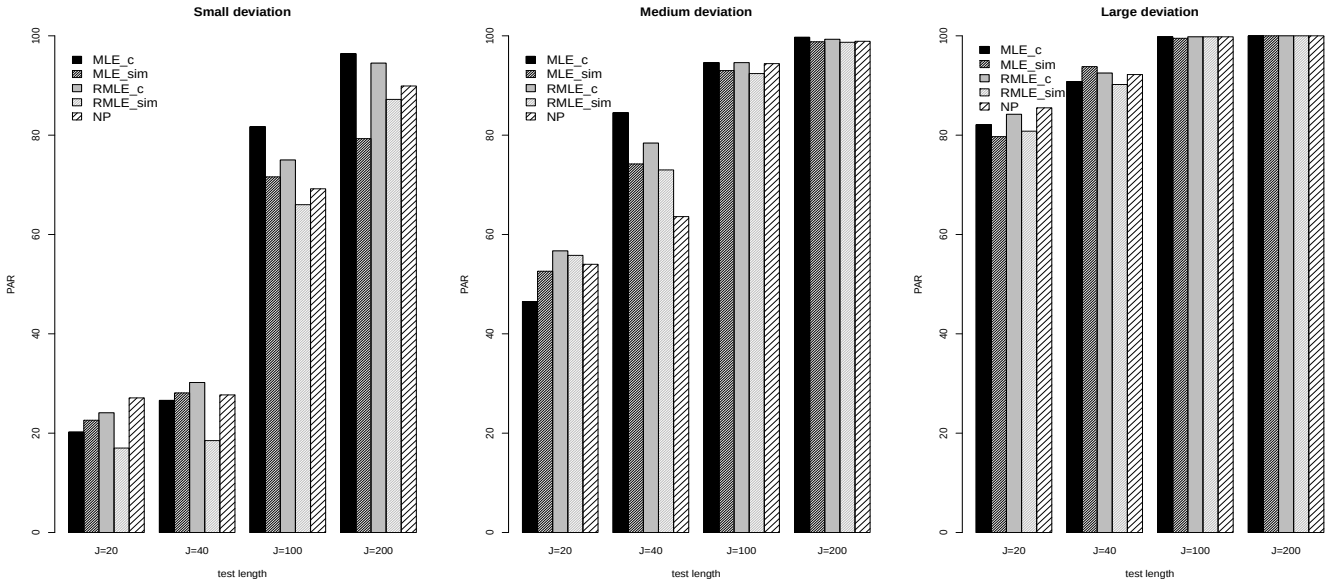

Figure 2.4: True Model: DINA

MLE_c: DINA MLE using calibrated item parameters; MLE_{sim}: DINA MLE using simulated item parameter; RMLE_c: Robust DINA MLE using calibrated item parameters; RMLE_{sim}:Robust DINA MLE using simulated item parameters; NP: the nonparametric estimation method

Figure 2.5 and Figure 2.6 document the classification results in terms of PAR of the three methods when the data were generated from the Reduced-RUM model and the NIDA model. Note that, instead of fitting the corresponding right model, we used the true model parameters to obtain the MLE for the attribute profile as a baseline (the results were denoted as "Real" in those figures). Similar conclusions as those from Figure 2.4 can be seen. However, the Robust DINA MLE displays a significant advantage over the DINA MLE in many cases. An additional conclusion is that when test length is long enough (100 or 200), the parametric methods based on a DINA model, no matter whether using calibrated item parameters or simulated item parameters, had a similar but slightly worse performance as the MLE obtained by using the true model parameters.

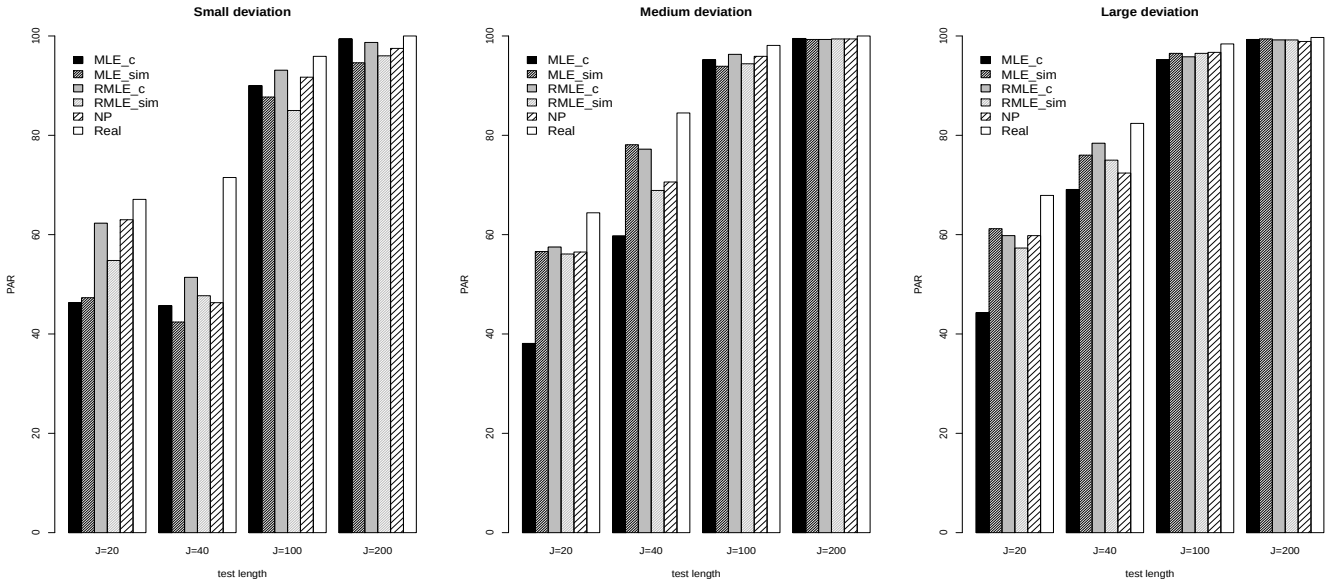

Figure 2.5: True Model: Reduced RUM

MLE_c: DINA MLE using calibrated item parameters; MLE_{sim}: DINA MLE using simulated item parameter; RMLE_c: Robust DINA MLE using calibrated item parameters; RMLE_{sim}:Robust DINA MLE using simulated item parameters; NP: the nonparametric estimation method; Real: MLE obtained using true reduced RUM model parameters

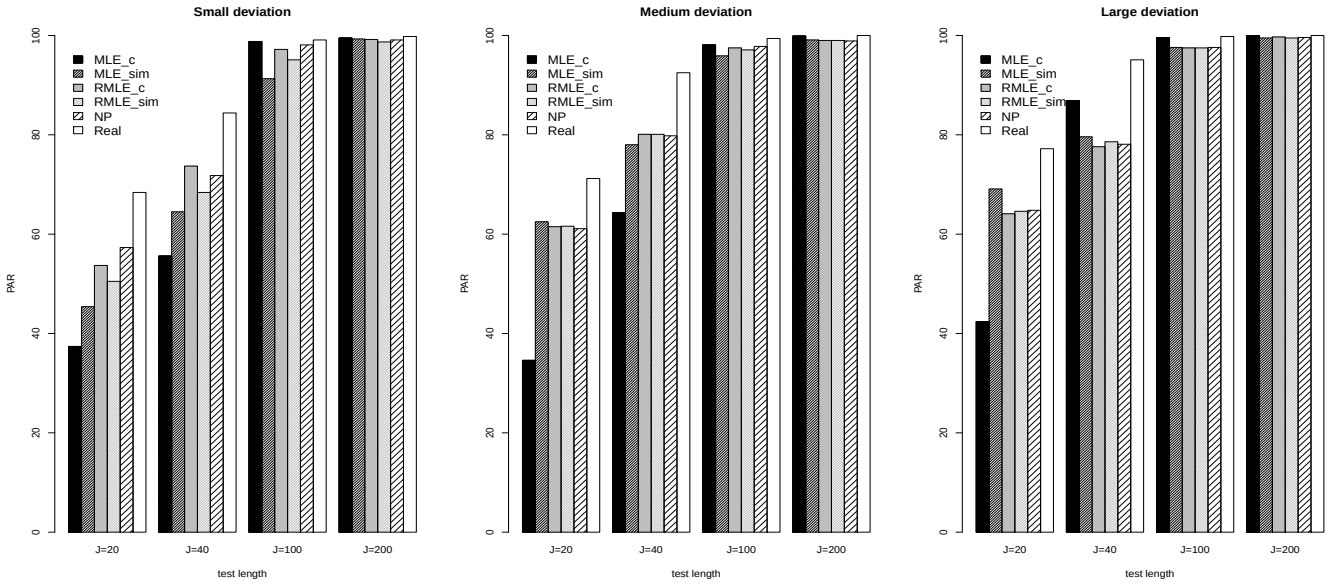

Figure 2.6: True Model: NIDA

MLE_c: DINA MLE using calibrated item parameters; MLE_{sim}: DINA MLE using simulated item parameter; RMLE_c: Robust DINA MLE using calibrated item parameters; RMLE_{sim}:Robust DINA MLE using simulated item parameters; NP: the nonparametric estimation method; Real: MLE obtained using true NIDA model parameters

Finally, the classification results when $K = 3$ are summarized in Table 2.5 to Table 2.7. The general

trends were the same as those when $K = 5$. The only difference is that now the Robust DINA MLE had a similar performance as the DINA MLE when the test length equaled 20 or 40.

Table 2.5: Classification rates for three methods with DINA data

$K = 3$			PAR				AAR			
			Test Length				Test Length			
δ_0	Estimation	Item Calibration	20	40	100	200	20	40	100	200
small	MLE	Calibration	72.7	78.2	98.5	98	87.4	90.3	99.5	99.3
		Simulation	73.2	65.6	90.2	91.5	88.7	82.9	95.9	97.0
	RMLE	Calibration	73.4	78.9	97.3	97.7	87.9	90.9	99.1	98.8
		Simulation	73.1	70.1	92.1	92.5	88.8	86.3	97.1	97.4
	NP	-	71.1	73.8	96.7	96.5	87.4	88.2	98.8	98.5
medium	MLE	Calibration	79.0	90.5	98.6	99.9	90.8	96.4	99.5	100
		Simulation	73.7	82.2	98.0	99.6	88.5	92.9	99.3	99.9
	RMLE	Calibration	73.8	90.7	98.2	99.8	87.6	96.5	99.3	100
		Simulation	76.4	83.4	97.8	99.6	90.1	93.6	99.2	99.9
	NP	-	74.5	85.2	98.3	99.7	96.7	97.7	99.9	100
large	MLE	Calibration	91.3	94.6	99.9	100	96.0	98.1	100	100
		Simulation	93.1	94.4	99.9	100	97.2	97.9	100	100
	RMLE	Calibration	90.8	93.6	99.9	100	95.7	97.7	100	100
		Simulation	93.1	94.1	99.9	100	97.2	97.8	99.9	100
	NP	-	92.4	93.5	99.9	100	96.7	97.6	99.9	100

Table 2.6: Classification rates for three methods with Reduced RUM data

$K = 3$			PAR				AAR			
			Test Length				Test Length			
δ_0	Estimation	Item Calibration	20	40	100	200	20	40	100	200
small	MLE	Calibration	88.8	93.9	97.2	99.2	95.7	97.6	99.1	99.7
		Simulation	83.9	89.8	95.2	95.1	93.2	96.2	98.3	98.3
	RMLE	Calibration	88.9	93.0	98.6	99.1	95.7	97.5	99.5	99.7
		Simulation	83.0	88.8	96.7	97.9	92.5	95.8	98.9	99.0
	NP	-	86.7	90.8	98.4	97.9	94.5	96.5	99.4	99.3
	R-RUM	-	88.0	96.8	99.4	100	95.2	98.6	99.8	100
medium	MLE	Calibration	87.9	90.9	98.2	100	95.4	96.5	99.3	100
		Simulation	85.1	90.0	98.0	99.6	93.8	96.3	99.2	99.9
	RMLE	Calibration	87.8	89.9	97.9	99.9	95.4	96.2	99.2	100
		Simulation	85.1	90.4	98.1	99.8	93.8	96.5	99.3	99.9
	NP	-	86.2	88.9	98.1	99.8	94.4	94.7	99.6	100
	R-RUM	-	87.9	96.2	99.3	100	95.1	98.6	99.8	100
large	MLE	Calibration	82.0	90.0	99.1	100	92.1	96.0	99.7	100
		Simulation	83.5	85.7	99.1	100	93.2	94.5	99.7	100
	RMLE	Calibration	82.7	86.0	99.1	100	92.6	94.2	99.7	100
		Simulation	85.3	87.5	99.2	100	94.3	95.0	99.7	100
	NP	-	84.2	86.3	99.0	100	93.6	94.7	99.6	100
	R-RUM	-	87.1	94.4	99.7	100	94.8	97.8	99.9	100

Table 2.7: Classification rates for three methods with NIDA data

$K = 3$			PAR				AAR			
			Test Length				Test Length			
δ_0	Estimation	Item Calibration	20	40	100	200	20	40	100	200
small	MLE	Calibration	78.8	95.3	96.6	99.6	90.3	98.3	98.9	99.9
		Simulation	74.6	92.2	92.1	98.7	89.1	96.7	97.1	99.6
	RMLE	Calibration	79.7	89.4	96.1	99.5	90.9	96.2	98.6	99.8
		Simulation	78.7	88.3	94.0	98.9	92.0	95.3	97.8	99.6
	NP	-	80.1	90.3	96.1	99.4	91.2	96.5	98.6	99.8
	NIDA	-	86.2	95.8	98.6	100	94.7	98.5	99.5	100
medium	MLE	Calibration	82.1	96.0	99.7	99.6	91.4	98.5	99.9	99.8
		Simulation	84.9	90.6	98.6	99.2	93.5	96.3	99.5	99.7
	RMLE	Calibration	82.3	92.5	98.5	99.2	91.5	97.3	99.5	99.7
		Simulation	82.6	90.2	98.0	99.0	91.8	96.1	99.3	99.7
	NP	-	84.6	91.8	98.5	99.3	93.2	96.9	99.5	99.7
	NIDA	-	85.2	96.9	99.9	100	93.7	98.8	100	100
large	MLE	Calibration	88.6	97.5	99.8	100	95.5	98.8	99.9	100
		Simulation	80.2	92.0	98.3	99.2	90.9	97.1	99.4	99.7
	RMLE	Calibration	86.1	91.2	98.3	99.5	94.9	96.8	99.4	100
		Simulation	83.2	91.2	98.2	99.6	92.7	96.7	99.4	100
	NP	-	82.6	92.3	98.3	99.4	92.5	97.2	99.4	100
	NIDA	-	89.3	98.5	99.9	100	95.6	99.4	100	100

To summarize, the consistency pattern for the DINA MLE, the Robust DINA MLE and the nonparametric classifier can be observed no matter whether the data were generated from the DINA model, the Reduced RUM model or the NIDA model. Although we have generated an example to show the inconsistency of the MLE under a misspecified conjunctive model, the simulation results indicate that the MLE estimator can still be consistent. This can be explained mathematically. Although there might be negative and positive terms in $\Delta^{(J)}$, this sum is still positive and of order J when the true item parameters are uniformly distributed. A very interesting finding is that if the true conjunctive model satisfies $\delta_0 > \delta$, when using the parametric methods to estimate attribute profiles, there is no need to calibrate the DINA item parameters. Simulating the DINA item parameters from a reasonable bounded uniform distribution may guarantee good classification results, especially when test length is long.

2.5 Real data analysis

In this section, we compare the classification results from the MLE, the Robust DINA MLE and the Non-parametric classifier based on a real data set. This data set contains responses to 34 items involving the square root operation from 5348 second year students in a primary school in Beijing. The Q matrix for this data set is in Table 6 below, assuming 6 attributes are required to do the square root operation. They are: (1) know the concept of square root, (2) square the number, (3) simplify the equation, (4) multiplication and division skills, (5) addition and subtraction skills, and (6) arithmetic ability.

Table 2.8: Q matrix for square root operation

	Item																
Attribute	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	1	1	1	1	0	0	1	1	1	1	1	0	0
4	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	Item																
Attribute	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	1
3	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	1	0
4	0	0	0	0	0	1	1	1	1	1	0	0	0	1	1	1	1
5	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1

In this analysis, a DINA model was used for the parametric estimation methods and the item parameters were either calibrated by the "CDM" package or simulated from a uniform distribution on (0.01, 0.3). All the calibrated parameters satisfy $1 - s_j - g_j > 0$. The total possible attribute profiles is $2^K = 64$. However, they were classified into 36 equivalence classes because the Q matrix is not complete. The $\boldsymbol{\alpha}$ which has the same ideal response pattern belongs to one equivalence class. The five estimation methods: DINA MLE based on calibrated item parameters, DINA MLE based on simulated item parameters, Robust DINA MLE

based on calibrated item parameters, Robust DINA MLE based on simulated item parameters and the nonparametric estimation method were used, and pairwise equivalent class agreement rates between the 5 estimation methods are summarized by Table 2.9 below:

Table 2.9: The Equivalence Classification Agreement Among Five Methods

	1	2	3	4	5
1	100%	49.4%	47.1%	47.0%	46.4%
2	-	100%	77.7%	91.0%	79.6%
3	-	-	100%	78.9%	80.8%
4	-	-	-	100%	80.4%
5	-	-	-	-	100%

1: DINA MLE based on calibrated item parameters

2: DINA MLE based on simulated item parameters

3: Robust DINA MLE based on calibrated item parameters

4: Robust DINA MLE based on simulated item parameters

5: Nonparametric estimation method

From the results we can see that the DINA MLE with calibrated item parameters had relatively low agreement with other methods, and the other four methods had relatively high pairwise agreement, especially for the DINA MLE with simulated item parameters and the Robust DINA MLE with simulated item parameters.

2.6 Discussion

We have examined the consequences of model misspecification in the cognitive diagnosis framework when using maximum likelihood classification of attribute profiles. Several interesting and even surprsing findings were observed. First, in general, when a CDM is misspecified as a conjunctive model, the MLE for attribute profiles is not necessarily consistent. An example was used to show that the MLE is not consistent when the true model is a DINA model and is misspecified as a DINA model with different item parameters. Second, we found a sufficient condition for the MLE to be a consistent estimator under a misspecified DINA model. The true model can be any conjunctive models or even a compensatory model that satisfies these conditions. We have provided another example to show that the MLE of attribute profile under a misspecified DINA model can be a consistent estimator even when the true model is a compensatory RUM model. Third, a Robust

DINA MLE technique was proposed to overcome the inconsistency issue, and theorems were presented to show that it is a consistent estimator for attribute profiles as long as the true model is a conjunctive model. Finally, simulation results indicated that when the true model is a conjunctive model, the robust DINA MLE and the DINA MLE based on the simulated item parameters can result in relatively good classification result even when the test length is short.

Those findings give some insights when using cognitive diagnosis models to estimate attribute profiles: simple and interpretable models can be fitted without severely affecting the classification accuracy. This is especially valuable when practitioners believe the response data come from a conjunctive process. In such cases, a simple DINA model can be used to get a relative good classification result. Surprisingly, according to the simulation results, one can use simulated item parameters instead of calibrated item parameters to get a good classification result.

Using a simple model can save much effort and expense on model calibration and offers a simple interpretation. However, the worry is that this may result in inflated missclassification. For the DINA model, the item response function is totally determined by the ideal response pattern, and if the Q matrix is not complete and the test length is short, many attribute patterns may result in the exactly the same ideal response pattern. This results in attribute patterns forming equivalence classes within which patterns may not be distinguished from one another. If at this time we have further information about the attribute patterns, not assuming they are independent, like the partially ordered set model (POSET) proposed by Tatsuoka (2002), all the attribute patterns can be identified. Particularly, by using Bayesian analysis techniques in Tatsuoka and Ferguson (2003), such POSET models can recognize confounding and equivalence classes in classifications. However, to use such methods one needs to have prior information on the distributions of the attribute profiles.

There is always a tradeoff between using a simple model and a complicated model. A complicated model may result in more accurate classification results, but one needs to devote more effort on model calibrations and result interpretations. A simple model can be easily implemented to analyze the data, but sometimes may cause some unnecessary loss of accuracy. This paper discusses the consequences of using a simple model, and gives some guidance for practitioners in what situations they can choose a simple model.

Chapter 3

Computerized Adaptive Testing that Allows for Response Revision

3.1 Introduction

A main goal in educational assessment is the accurate estimation of the test-taker's ability. In a conventional paper-pencil test, this estimation is based on the examinee's responses to a preassembled set of items. On the other hand, in Computerized Adaptive Testing (CAT), as it was originally conceived by Lord (1971), items are selected in real time and can be tailored to the examinee's ability, which is learned as the test progresses. This feature is especially important for examinees at the two extreme ends of the ability distribution, who may otherwise receive items that are either too difficult or too easy.

The design of CAT is based on Item Response Theory (IRT) models for the probability of a correct answer given the examinee's ability and the item itself. The simplest IRT model is the Rasch model (Rasch (1993)), in which the probability of a correct answer is equal to $H(\theta - b)$, where H is the cdf of the logistic distribution, θ is a scalar parameter that represents the ability of the examinee and b is the difficulty parameter of the item. A generalization of the Rasch model is the three parameter logistic model (3PL) model, in which the probability of a correct answer is equal to $c + (1 - c)H(a(\theta - b))$, where $c \in (0, 1)$ is a parameter that captures the probability of guessing the right answer and a is the discrimination parameter of the item. An intermediate model, the two parameter logistic model (2PL), arises when we set $c = 0$.

A standard approach for item selection in CAT, proposed by Lord (Lord, 1980), is to select the item with the maximum Fisher information at each step. In the case of the Rasch model, this means that item i should be selected so that its difficulty parameter b_i is equal to θ. Since θ is unknown, this suggests setting b_i equal to an estimate of θ based on the first $i - 1$ observations. For the adaptive estimation of θ, Lord (1971) proposed the Stochastic Approximation algorithm of Robbins and Monro (1951). However, this non-parametric approach can be very inefficient with binary data, as it was shown by Lai and Robbins (1979). For this reason, Wu (1985, 1986) suggested using the Maximum Likelihood Estimator (MLE) of θ based on the first $i - 1$ observations. Following this approach, coupled with the information maximizing item selection strategy, Ying and Wu (1997) established the strong consistency and asymptotic normality of

the resulting estimator under the Rasch model, whereas Chang and Ying (2009) extended these results to the case of the 2PL and 3PL model.

Thanks to the above statistical advances, as well as the rapid development of modern technology, CAT has become popular for many kinds of measurement tasks, such as educational testing, patient reported outcomes, and quality of life measurement. Examples of large-scale CATs include the Graduate Management Admission Test (GMAT), the National Council Licensure Examination (NCLEX) for nurses, and the Armed Services Vocational Aptitude Battery (ASVAB) (Chang and Ying, 2007). Moreover, beyond the problem of ability estimation, CAT has been applied to mastery testing (Sie et al., 2015; Bartroff et al., 2008) and cognitive diagnosis (Liu et al. (2015)). However, in none of the currently operational CAT programs are the examinees allowed to revise their responses to previously administered items during the test (Vispoel et al. (1999)) and this is one of the reasons that some testing programs have decided to switch to other modes of testing (Luecht and Nungester, 1998).

The main argument against response revision among practitioners and researchers who oppose this feature is that it violates the adaptive nature of CAT. Specifically, it has been argued that allowing for response revision decreases estimation efficiency and increases bias (Stocking, 1997; Vispoel et al., 1999), as well as that it gives the opportunity to disingenuous examinees to artificially inflate their test scores by adopting deceptive test-taking strategies (Wainer, 1993; Kingsbury, 1996; Wise et al., 1999). On the other hand, it has been argued that response revision in CAT may lead to more accurate inference, by minimizing measurement error, and to a friendlier testing environment, by lowering the anxiety of the examinees (Wise, 1996; Vispoel et al., 2000). Indeed, it has been reported that examinees would favor the response revision feature in CAT (Vispoel and Coffman, 1992), whereas the desire for review opportunities has also been verified in other studies of computerized tests, e.g., (Schmidt et al., 1978).

Overall, the absence of the opportunity to revise in CAT has been a main concern for both examinees and testing companies and a number of modified CAT designs have been proposed in order to incorporate this feature (Stocking, 1997; Vispoel et al., 2000; Han, 2013). In order to prevent the potential dangers of revision, these designs have postulated quite limited revision rules. For example, they may impose an upper bound on the number of items that can be revised, or they may allow the test-taker to revise only at specific times during the test. Under such rules and restrictions, it has been reported that response revision does not impact the estimation accuracy and efficiency of CAT. However, these conclusions were based only on simulation experiments and were not supported theoretically.

In this chapter, we propose and analyze a novel CAT design whose goal is to preserve the advantages of conventional CAT with respect to estimation efficiency, but at the same time to allow examinees to

revise their answers at any time during the test. The only restriction is on the number of revisions *to the same item*. Thus, we impose significantly more relaxed revision constraints on the examinees in comparison to alternative CAT designs that have been proposed to allow for response revision. In order to achieve this, we use a different modeling framework than that of a typical CAT design. Indeed, although most operational CAT programs employ multiple-choice items, they model them in a dichotomous way, specifying the probability of each response being either right or wrong. On the contrary, we use a polytomous IRT model, the nominal response model proposed by Bock (1972), and specify the probability that the examinee selects each category of a given item. Based on this model, we postulate a joint probability model for the first answer to each item and any subsequent revisions to it and we update the ability parameter after each response with the maximizer of the likelihood of all responses, first answers *and* revisions. However, we do not make any modeling assumptions for the decision of the examinee to revise or not at each step. Finally, whenever the examinee asks for a new item, we select the one with the maximum Fisher information at the current estimate of the ability level.

We provide an asymptotic analysis of the proposed method, which to our knowledge is the first rigorous theoretical analysis of a CAT design that allows for response revision. First, we show that the resulting estimator is strongly consistent as the number of administered items goes to infinity (Theorem 3.4.1). Then, we show that it is asymptotically normal under a stability assumption on the cumulative Fisher information that is satisfied when the number of revisions is small relative to the number of items (Theorem 3.4.2).

Moreover, we consider separately the case of a conventional CAT (that does not allow for response revision) that is based on the nominal response model. To our knowledge, there has not been any theoretical analysis of a standard CAT that is based on a polytomous model, therefore the corresponding consistency (Theorem 3.3.1) and asymptotic normality (Theorem 3.3.2) results are new and of independent interest. Most importantly, they help us illustrate the conceptual and technical differences between the traditional CAT setup, where the number of observed responses coincides with the number administered items at any time during the test, and the proposed setup in which the number of responses and items are in general different.

The theoretical analysis of our design does not rely on any assumptions about the examinee's revision strategies, i.e., nothing was assumed regarding when and which items the examinee chooses to revise. However, before this design is implemented in practice, it is important to understand how it behaves under various test-taking strategies that may even violate some of its underlying modeling assumptions. Thus we also investigate the proposed design under different revision strategies. First, in order to illustrate that this design can reduce measurement error, we consider examinees that only correct careless mistakes caused by

misreading the items or temporary lapses in memory. It is reasonable to expect that correcting such mistakes can yield scores that represent more accurately the examinee's skill level Vispoel (1998). Second, in order to address the issue of cheating, which is a main concern for researchers and practitioners regarding response revision in CAT Wise (1996), we consider examinees that adopt test-taking strategies which take advantage of the response revision feature, such as the Wainer strategy and the generalized Kingsbury strategy (Wainer, 1993; Green et al., 1984; Wise et al., 1999).

Throughout the paper, we focus on a single examinee whose ability is quantified by an unknown, scalar parameter $\theta \in \mathbb{R}$ and we denote by $\mathsf{P}_\theta/\mathsf{E}_\theta/\mathsf{Var}_\theta$ the corresponding probability measure/expectation/variance.

3.2 Nominal Response Model

Let X be the response to a generic multiple-choice item with $m \geq 2$ categories. We write $X = k$ when the examinee chooses category k and we assume that

$$\mathsf{P}_\theta(X = k) = \frac{\exp(a_k\theta + c_k)}{\sum_{h=1}^m \exp(a_h\theta + c_h)}, \quad 1 \leq k \leq m, \tag{3.1}$$

where $\{a_k, c_k\}_{1\leq k\leq m}$ are known, item-specific real numbers that satisfy

$$\sum_{k=1}^m |a_k| \neq 0 \quad \text{and} \quad \sum_{k=1}^m |c_k| \neq 0 \tag{3.2}$$

and the identifiability conditions that $\sum_{k=1}^m a_k = \sum_{k=1}^m c_k = 0$.

Thus, the distribution of X is specified by the ability of the examinee, θ, and the item-specific vector $\mathbf{b} = (a_2, \ldots, a_m, c_2, \ldots, c_m)$ and we write

$$p_k(\theta; \mathbf{b}) := \mathsf{P}_\theta(X = k), \quad 1 \leq k \leq m, \tag{3.3}$$

whereas we use the following notation for the corresponding log-likelihood, score function and Fisher information

$$\begin{aligned} \ell(\theta; \mathbf{b}, X) &:= \log \mathsf{P}_\theta(X) = \sum_{k=1}^m \log\big(p_k(\theta; \mathbf{b})\big)\, \mathbb{1}_{\{X=k\}}, \\ s(\theta; \mathbf{b}, X) &:= \frac{d}{d\theta}\ell(\theta; \mathbf{b}, X), \quad J(\theta; \mathbf{b}) := \mathsf{Var}_\theta[s(\theta; \mathbf{b}, X)]. \end{aligned} \tag{3.4}$$

Remark: In the special case of binary data ($m = 2$), $a_1 = -a_2$, $c_1 = -c_2$,

$$p_2(\theta; \mathbf{b}) = 1 - p_1(\theta; \mathbf{b}) = \frac{\exp(2a_2\theta + 2c_2)}{1 + \exp(2a_2\theta + 2c_2)}, \tag{3.5}$$

thus, we recover the 2PL model with discrimination parameter $2|a_1|$ and difficulty parameter $-c_2/a_2$.

For a given item with parameter vector $\mathbf{b}$, we denote by $a^*(\mathbf{b})$ and $a_*(\mathbf{b})$ the maximum and minimum of the a_k's respectively, i.e.,

$$a^*(\mathbf{b}) := \max_{1 \le k \le m} a_k \quad \text{and} \quad a_*(\mathbf{b}) := \min_{1 \le k \le m} a_k.$$

Furthermore, we denote by $k^*(\mathbf{b})$ and $k_*(\mathbf{b})$ the category with the largest and smallest a-value, respectively, i.e.,

$$k^*(\mathbf{b}) := \operatorname*{argmax}_{1 \le k \le m} a_k \quad \text{and} \quad k_*(\mathbf{b}) := \operatorname*{argmin}_{1 \le k \le m} a_k. \tag{3.6}$$

For simplicity and without loss of generality, we assume that the categories are ordered, so that $k^*(\mathbf{b}) = m$ and $k_*(\mathbf{b}) = 1$. We assume that $\mathbf{b}$ takes values in some *compact* set $\mathbb{B} \subset \mathbb{R}^{2m-2}$ that represents the underlying item bank/pool. Then, we denote by $p^*(\theta)$ (resp. $p_*(\theta)$) the maximum probability of selecting the category with the largest (resp. smallest) a-value, i.e.,

$$p^*(\theta) := \sup_{\mathbf{b} \in \mathbb{B}} p_{k^*(\mathbf{b})}(\theta; \mathbf{b}) \quad \text{and} \quad p_*(\theta) := \sup_{\mathbf{b} \in \mathbb{B}} p_{k_*(\mathbf{b})}(\theta; \mathbf{b}), \tag{3.7}$$

and by $J^*(\theta)$ (resp. $J_*(\theta)$) the maximal (resp. minimal) Fisher information in the pool, i.e.,

$$J_*(\theta) := \inf_{\mathbf{b} \in \mathbb{B}} J(\theta; \mathbf{b}) \quad \text{and} \quad J^*(\theta) := \sup_{\mathbf{b} \in \mathbb{B}} J(\theta; \mathbf{b}). \tag{3.8}$$

The next two lemmas will be used throughout the paper.

Lemma 1. *If $g : \mathbb{R} \times \mathbb{B} \to \mathbb{R}$ is a jointly continuous function, then*

(i) $\sup_{\mathbf{b} \in \mathbb{B}} g(\cdot, \mathbf{b})$ *and* $\inf_{\mathbf{b} \in \mathbb{B}} g(\cdot, \mathbf{b})$ *are also continuous functions.*

(ii) If $x_n \to x_0$, then

$$\limsup_n \sup_{\mathbf{b}\in\mathbb{B}} g(x_n;\mathbf{b}) \leq \sup_{\mathbf{b}\in\mathbb{B}} g(x_0;\mathbf{b})$$

$$\liminf_n \inf_{\mathbf{b}\in\mathbb{B}} g(x_n;\mathbf{b}) \geq \inf_{\mathbf{b}\in\mathbb{B}} g(x_0;\mathbf{b})$$

and $\sup_{\mathbf{b}\in\mathbb{B}} |g(x_n,\mathbf{b}) - g(x_0,\mathbf{b})| \to 0$.

Proof. Part (i) follows from the so-called Maximum Theorem (see, e.g., Sundaram, R. K. (1996), p. 239), so we will only prove (ii). In order to do so, we will first prove that for every $x_0 \in \mathbb{R}$ we have

$$\limsup_n \sup_{\mathbf{b}\in\mathbb{B}} g(x_n;b) \leq \sup_{\mathbf{b}\in\mathbb{B}} g(x_0;b). \tag{3.9}$$

For any given $\mathbf{b} \in \mathbb{B}$ and $\rho > 0$ define the function:

$$\phi(x,\mathbf{b},\rho) := \sup\{g(x;\mathbf{b}');\ \mathbf{b}' \in \overline{S(\mathbf{b},\rho)}\}, \quad x \in \mathbb{R},$$

where $S(\mathbf{b},\rho) := \{\mathbf{b}' \in \mathbb{B} : ||\mathbf{b}' - \mathbf{b}|| < \rho\}$ is an open ball with center $\mathbf{b}$ and radius ρ and $\overline{S(\mathbf{b},\rho)}$ its closure. From (i) it follows that $\phi(\cdot,\mathbf{b},\rho)$ is a continuous function.

For every x, $g(x;\cdot)$ is continuous, thus, upper semi-continuous, therefore we have

$$\limsup_{\rho\to 0} \phi(x;\mathbf{b},\rho) \leq g(x;\mathbf{b}).$$

Thus, if we fix some $\epsilon > 0$, for any $b \in \mathbb{B}$ we can find some $\rho_b > 0$ small enough so that

$$\phi(x;\mathbf{b},\rho_{\mathbf{b}}) \leq g(x;\mathbf{b}) + \epsilon. \tag{3.10}$$

Since $\{S(\mathbf{b},\rho_{\mathbf{b}})\}_{\mathbf{b}\in\mathbb{B}}$ is an open cover of $\mathbb{B}$ and $\mathbb{B}$ is compact, there is a finite set $\{\mathbf{b}_1,\ldots,\mathbf{b}_J\} \subset \mathbb{B}$ so that $\{S(\mathbf{b}_j,\rho_{\mathbf{b}_j}\}_{1\leq j\leq J}$ is also a cover of $\mathbb{B}$. This means that for some arbitrary $\mathbf{b} \in \mathbb{B}$ there exists a $j \in \{1,\ldots,J\}$ so that $\mathbf{b} \in S(\mathbf{b}_j,\rho_{\mathbf{b}_j})$ and, consequently,

$$g(x_n;\mathbf{b}) \leq \phi(x_n;\mathbf{b}_j,\rho_{\mathbf{b}_j}) \leq \max_{1\leq j\leq J} \phi(x_n;\mathbf{b}_j,\rho_{b_j})$$

Since the right-hand side is free of $\mathbf{b}$, we have

$$\sup_{\mathbf{b}\in\mathbb{B}} g(x_n;\mathbf{b}) \leq \max_{1\leq j\leq J} \phi(x_n;\mathbf{b}_j,\rho_{\mathbf{b}_j}). \tag{3.11}$$

Since $x_n \to x_0$ and $\phi(\cdot,\mathbf{b},\rho)$ is a continuous function, we have $\phi(x_n;\mathbf{b}_j,\rho_{\mathbf{b}_j}) \to \phi(x_0,\mathbf{b}_j,\rho_{\mathbf{b}_j})$ for every $1\leq j\leq J$ and, consequently,

$$\max_{1\leq j\leq J} \phi(x_n;\mathbf{b}_j,\rho_{\mathbf{b}_j}) \to \max_{1\leq j\leq J} \phi(x_0,\mathbf{b}_j,\rho_{\mathbf{b}_j}). \tag{3.12}$$

From (3.11) and (3.12) we conclude that

$$\limsup_n \sup_{\mathbf{b}\in\mathbb{B}} g(x_n;\mathbf{b}) \leq \max_{1\leq j\leq J} \phi(x_0;\mathbf{b}_j,\rho_{\mathbf{b}_j}). \tag{3.13}$$

However, from the definition of the $\rho_{\mathbf{b}}$'s in (3.10) we have

$$\max_{1\leq j\leq J} \phi(x;\mathbf{b}_j,\rho_{\mathbf{b}_j}) \leq \max_{1\leq j\leq J} g(x_0;\mathbf{b}_j) + \epsilon \leq \sup_{\mathbf{b}\in\mathbb{B}} g(x_0;\mathbf{b}) + \epsilon \tag{3.14}$$

From (3.13) and (3.14) we conclude that

$$\limsup_n \sup_{\mathbf{b}\in\mathbb{B}} g(x_n;\mathbf{b}) \leq \sup_{\mathbf{b}\in\mathbb{B}} g(x_0;\mathbf{b}) + \epsilon \tag{3.15}$$

Since ϵ is arbitrary, this proves (3.9). Now, in order to prove (ii), let us assume that $g(x_0;\mathbf{b}) = 0$ for every $\mathbf{b}\in\mathbb{B}$. This can be done without any loss of generality, since we may otherwise work with $g(x_n,\mathbf{b}) - g(x_0,\mathbf{b})$. Then, we simply need to show that

$$\begin{aligned}\limsup_n \sup_{\mathbf{b}\in\mathbb{B}} |g(x_n;\mathbf{b})| &= \limsup_n \ \max\{\sup_{\mathbf{b}\in\mathbb{B}} g(x_n;\mathbf{b}) \ , \ \sup_{\mathbf{b}\in\mathbb{B}} -g(x_n;\mathbf{b})\} \\ &= 0.\end{aligned} \tag{3.16}$$

Applying (3.9) with $-g$ implies that

$$\limsup_n \sup_{\mathbf{b}\in\mathbb{B}} -g(x_n;\mathbf{b}) \leq 0. \tag{3.17}$$

Combining (3.9) and (3.17) proves (3.16).

$\diamondsuit$

Lemma 2. *(i) Fix $\mathbf{b} \in \mathbb{B}$. Then,*

$$s(\theta; \mathbf{b}, X) := \sum_{k=1}^{m} \big[a_k - \bar{a}(\theta; \mathbf{b})\big]\, \mathbb{1}_{\{X=k\}}, \tag{3.18}$$

$$J(\theta; \mathbf{b}) := \sum_{k=1}^{m} \Big(a_k - \bar{a}(\theta; \mathbf{b})\Big)^2 p_k(\theta; \mathbf{b}); \tag{3.19}$$

where $\bar{a}(\theta; \mathbf{b})$ is the following weighted average of the a_k's:

$$\bar{a}(\theta; \mathbf{b}) := \sum_{h=1}^{m} a_h\, p_h(\theta; \mathbf{b}), \tag{3.20}$$

Moreover, the derivative of $s(\theta; \mathbf{b}, X)$ with respect to θ does not depend on X. Specifically,

$$s'(\tilde{\theta}; \mathbf{b}) := \frac{d}{d\theta} s(\theta; \mathbf{b}, X)\Big|_{\theta=\tilde{\theta}} = -J(\theta; \mathbf{b}). \tag{3.21}$$

(ii) Fix $\mathbf{b} \in \mathbb{B}$. Then, $\bar{a}(\theta; \mathbf{b}) \to a_*(\mathbf{b})$, $p_{k_*(\mathbf{b})}(\theta; \mathbf{b}) \to 1$ as $\theta \to -\infty$, $\bar{a}(\theta; \mathbf{b}) \to a^*(\mathbf{b})$ and $p_{k^*(\mathbf{b})}(\theta; \mathbf{b}) \to 1$ as $\theta \to +\infty$. Moreover,

$$\lim_{|\theta| \to \infty} J(\theta; \mathbf{b}) = 0. \tag{3.22}$$

(iii) The functions $\theta \to J_*(\theta)$ and $\theta \to J^*(\theta)$ are continuous.

(iv) For any ability level θ, $p^*(\theta) < 1$, $p_*(\theta) < 1$, and for any X we have $|s(\theta; b, X)| \le K$, $0 < J_*(\theta) \le J^*(\theta) \le K$, where K is a constant that does not depend on θ or $\mathbf{b}$.

Proof. Parts (i) and (ii) follow from direct computation. Part (iii) follows from Lemma 1, since $J(\theta; \mathbf{b})$ is jointly continuous and $\mathbb{B}$ compact. Finally, for part (iv), for any θ and $\mathbf{b}$ we have

$$|s(\theta; b, X)| \le \max_{1 \le k \le m} |a_k - \bar{a}(\theta; \mathbf{b})| \le 2a^*(\mathbf{b}) \le 2 \sup_{\mathbf{b} \in \mathbb{B}} a^*(\mathbf{b}).$$

Moreover,

$$0 < J(\theta; \mathbf{b}) \le \sum_{k=1}^{m} a_k^2\, p_k(\theta; \mathbf{b}) \le m\, \left(a^*(\mathbf{b})\right)^2 \le m \sup_{\mathbf{b} \in \mathbb{B}} \left(a^*(\mathbf{b})\right)^2,$$

where the first inequality holds because the a_k's cannot be identical, due to (2.2) and the identification condition, and the second follows from (2.10). The upper bound does not depend on $\mathbf{b}$ or θ, therefore it is a bound for $J^*(\theta)$. On the other hand, since J^* is continuous and $\mathbb{B}$ compact, we have $J_*(\theta) > 0$.

♢

3.3 Standard CAT with Nominal Response Model

In this section, we consider the design of a CAT that is based on the nominal response model, but is conventional in that it does not allow for response revision.

3.3.1 Problem formulation

Let n be the total number of items that will be administered to the examinee and let X_i denote the response to item i, where $1 \leq i \leq n$. In order to lighten the notation, we assume that each item has $m \geq 2$ categories and we write $X_i = k$ if the examinee chooses category k in item i, where $1 \leq k \leq m$ and $1 \leq i \leq n$. The responses are assumed to be governed by the nominal response model (3.1)-(3.3), so that

$$\mathsf{P}_\theta(X_i = k) := p_k(\theta; \mathbf{b}_i), \quad 1 \leq k \leq m,\ 1 \leq i \leq n, \tag{3.23}$$

where θ is the scalar parameter of interest that represents the ability of the examinee and $\mathbf{b}_i := (a_{i2}, \ldots, a_{im}, c_{i2}, \ldots, c_{im})$ is vector that characterizes item i, satisfies (3.2) and takes values in the compact set $\mathbb{B} \subset \mathbb{R}^{2m-2}$. Note that in practice each $\mathbf{b}_i$ can not take any value in $\mathbb{B}$, because there is only a finite number of items in a given item bank and there are further restrictions on the exposure rate of the items (Chang and Ying (1999)). Nevertheless, this assumption will allows us to obtain a benchmark for the large-sample performance that this method can attain in practice.

We assume that the responses are conditionally independent given the selected items, in the sense that

$$\mathsf{P}_\theta(X_{1:i} |\, \mathbf{b}_{1:i}) = \prod_{j=1}^{i} \mathsf{P}_\theta(X_j | \mathbf{b}_j), \quad 1 \leq i \leq n, \tag{3.24}$$

where $X_{1:i} \equiv (X_1, \ldots, X_i)$ and $\mathbf{b}_{1:i} \equiv (\mathbf{b}_1, \ldots, \mathbf{b}_i)$. In a paper-pencil test, the selected items are fixed in advance, thus, the corresponding item parameters, $\mathbf{b}_{1:n}$, are deterministic. However, this is not the case in CAT, where items are determined in real time based on the already observed responses. Specifically, if we denote by $\mathcal{F}_i^X$ the information contained in the first i responses, i.e., $\mathcal{F}_i^X := \sigma(X_1, ..., X_i)$, then $\mathbf{b}_i$ must be a $\mathcal{F}_{i-1}^X$-measurable, $\mathbb{B}$-valued random vector for every $2 \leq i \leq n$, whereas $\mathbf{b}_1$ is arbitrary. As a result, despite assumption (3.24), the responses are far from independent and, in fact, they may have a complex dependence structure.

The problem in a standard CAT design is to find an *ability estimator*, $\hat{\theta}_n$, at the end of the test, i.e., an $\mathcal{F}_n^X$-measurable estimator of θ, and *an item selection strategy*, $(\mathbf{b}_i)_{2 \leq i \leq n}$, so that the accuracy of $\hat{\theta}_n$ can be optimized. We suggest selecting the items in order to maximize the Fisher information at the current

estimate of the ability level, i.e., item $i+1$ should be selected so that

$$\hat{\mathbf{b}}_{i+1} = \underset{\mathbf{b}\in\mathbb{B}}{\operatorname{argmax}}\, J(\hat{\theta}_i; \mathbf{b}), \quad 1 \le i \le n-1, \tag{3.25}$$

where J is the Fisher information function of the nominal response model given by (3.19) and $\hat{\theta}_i$ is an estimate of θ based on the first i responses, i.e, an $\mathcal{F}_i$-measurable statistic. Thus, we need to estimate θ not only at the end of the test, but after each response. Therefore, in order to implement the item selection strategy (3.25), we need to estimate the test-taker's ability in an on-line fashion.

We will define the proposed adaptive ability estimator for an arbitrary item selection strategy , not necessarily (3.25). To be more specific, due to (3.23) and (3.24), the conditional log-likelihood and score function of the first i responses given arbitrary selected items $\mathbf{b}_{1:i}$ take the form

$$\begin{aligned} L_i(\theta) &:= \log \mathsf{P}_\theta(X_{1:i} \mid \mathbf{b}_{1:i}) = \sum_{j=1}^{i} \ell(\theta; \mathbf{b}_j, X_j), \\ S_i(\theta) &:= \frac{d}{d\theta} L_i(\theta) = \sum_{j=1}^{i} s(\theta; \mathbf{b}_j, X_j), \end{aligned} \tag{3.26}$$

where $\ell(\theta; \mathbf{b}_j, X_j)$ and $s(\theta; \mathbf{b}_j, X_j)$ are the log-likelihood and score, respectively, of the j^{th} response, defined in (3.4). Then, our estimate of θ based on the first i observations is the root of $S_i(\theta)$, which exists and is unique as long as at least one of the first i responses does not correspond to the category with the largest a-value or to the category with the smallest a-value, that is as long as $i > n_0$, where

$$n_0 := \max\Big\{ i \in \{1, \ldots, n\} : X_j = k^*(\mathbf{b}_j)\ \forall j \le i \quad \text{or} \quad X_j = k_*(\mathbf{b}_j)\ \forall j \le i \Big\}$$

(recall the definition of $k^*(\mathbf{b})$ and $k_*(\mathbf{b})$ in (3.6)). For example, if we have items with $m = 4$ categories in which the largest (resp. smallest) a-value is associated with category 4 (resp. 1), we have $n_0 = 3$ in a sequence of responses $1, 1, 1, 3, \ldots$.

For $i \le n_0$, we need an alternative ability estimator, such as the Bayesian estimator in Bock and Aitkin (1981). Alternatively, for $i \le n_0$ we may set $\hat{\theta}_0 = 0$ and $\hat{\theta}_i = \hat{\theta}_{i-1} + d$ (resp. $\hat{\theta}_i = \hat{\theta}_{i-1} - d$) if the initial responses have the largest (resp. smallest) a-value, where d is some predetermined constant. In any case, the following lemma shows that, for large n, the final ability estimator, $\hat{\theta}_n$, is the root of the score function with probability 1.

Lemma 3. $\mathsf{P}_\theta(S_n(\hat{\theta}_n) = 0$ *for all large* $n) = 1$.

Proof. The final ability estimator, $\hat{\theta}_n$, is not a root of $S_n(\theta)$ on the event $A_n \cup B_n$, where

$$A_n = \{X_1 = k^*(\mathbf{b}_1), \ldots, X_n = k^*(\mathbf{b}_n)\},$$
$$B_n = \{X_1 = k_*(\mathbf{b}_1), \ldots, X_n = k_*(\mathbf{b}_n)\}.$$

Thus, it suffices to show that $\mathsf{P}_\theta(\limsup_n A_n) = 0$ and $\mathsf{P}_\theta(\limsup_n B_n) = 0$. We will prove only the first identity, since the second can be shown in a similar way. Indeed, $\mathsf{P}_\theta(A_n) = \mathsf{E}_\theta\left[\mathsf{P}_\theta\left(A_n \,|\, \mathbf{b}_{1:n}\right)\right]$ and

$$\mathsf{P}_\theta\left(A_n \,|\, \mathbf{b}_{1:n}\right) = \prod_{i=1}^{n} p_{k^*(\mathbf{b}_i)}(\theta; \mathbf{b}_i) \leq \left(p^*(\theta)\right)^n,$$

where the equality follows the assumption of conditional independence (3.2) and the inequality from the definition of $p^*(\theta)$ in (2.7). Since $p^*(\theta) < 1$ (recall Lemma 2(iv)), we have $\sum_{n=1}^{\infty} \mathsf{P}_\theta(A_n) < \infty$ and from the Borel-Cantelli lemma we obtain $\mathsf{P}_\theta(\limsup_n A_n) = 0$, which completes the proof.

♢

3.3.2 Asymptotic analysis

In this section, we assume that (3.23) and (3.24) hold and we establish the asymptotic properties of the final ability estimator, $\hat{\theta}_n$. Specifically, we establish its strong consistency for an arbitrary item selection strategy and its asymptotic normality when the information-maximizing item selection strategy (3.25) is adopted.

First of all, in Lemma 4 we show that for an arbitrary item selection strategy, $(\mathbf{b}_i)_{1 \leq i \leq n}$, the corresponding score function $S_n(\theta)$ is a martingale with mean 0 and predictable variation equal to the conditional Fisher information

$$I_n(\theta) := \sum_{i=1}^{n} J(\theta; \mathbf{b}_i), \tag{3.27}$$

where $J(\theta; \mathbf{b}_i)$ is the Fisher information of the i^{th} item given by (3.19). Note also that from (3.21) it follows that

$$S_n^{'}(\tilde{\theta}) := \frac{d}{d\theta} S_n(\theta)\Big|_{\theta=\tilde{\theta}} = \sum_{i=1}^{n} -J(\tilde{\theta}; \mathbf{b}_i) = -I_n(\tilde{\theta}). \tag{3.28}$$

Lemma 4. *For any item selection strategy, the score function $\{S_n(\theta)\}_{n\in\mathbb{N}}$ is a $\{\mathcal{F}_n\}$-martingale under P_θ,*

with bounded increments, mean 0 *and predictable variation*

$$\langle S(\theta)\rangle_n := \sum_{i=1}^{n} \mathsf{E}_\theta \left[\left(S_i(\theta) - S_{i-1}(\theta) \right)^2 \mid \mathcal{F}_{i-1} \right] = I_n(\theta).$$

Proof. Fix $n \in \mathbb{N}$. Then,

$$S_n(\theta) - S_{n-1}(\theta) = s(\theta; \mathbf{b}_n, X_n)$$

and from Lemma 2 (iv) it follows that $|S_n(\theta) - S_{n-1}(\theta)| \leq K$. Moreover, since $\mathbf{b}_n$ is $\mathcal{F}_{n-1}$-measurable, from representation (2.9) it follows that

$$\mathsf{E}_\theta[S_n(\theta) - S_{n-1}(\theta)|\mathcal{F}_{n-1}] = \mathsf{E}_\theta[s(\theta; \mathbf{b}_n, X_n)|\mathcal{F}_{n-1}] = 0,$$

which proves the martingale property of $S_n(\theta)$. Next, from (2.10) it follows that

$$\mathsf{E}_\theta[(S_n(\theta) - S_{n-1}(\theta))^2|\mathcal{F}_{n-1}] = \mathsf{E}_\theta[s^2(\theta; \mathbf{b}_n, X_n)|\mathcal{F}_{n-1}] = J(\theta; \mathbf{b}_n),$$

which proves that $\langle S(\theta)\rangle_n = \sum_{i=1}^{n} J(\theta; \mathbf{b}_i)$.

♢

The following theorem establishes the strong consistency of $\hat{\theta}_n$.

Theorem 3.3.1. *For any item selection strategy, as* $n \to \infty$ *we have* $\mathsf{P}_\theta(\hat{\theta}_n \to \theta) = 1$ *and*

$$\frac{I_n(\hat{\theta}_n)}{I_n(\theta)} \to 1 \quad \mathsf{P}_\theta - a.s. \tag{3.29}$$

Proof. Let $(\mathbf{b}_n)_{n\in\mathbb{N}}$ be an arbitrary item selection strategy. From Lemma 4 it follows that $S_n(\theta)$ is a P_θ-martingale with mean 0 and predictable variation $I_n(\theta) \geq nJ_*(\theta) \to \infty$, since $J_*(\theta) > 0$. Then, from the Martingale Strong Law of Large Numbers (see, e.g., Williams, D. (1991), p. 124), it follows that as $n \to \infty$

$$\frac{S_n(\theta)}{I_n(\theta)} \to 0 \quad \mathsf{P}_\theta - \text{a.s.} \tag{3.30}$$

From a Taylor expansion of $S_n(\theta)$ around $\hat{\theta}_n$ it follows that there exists some $\tilde{\theta}_n$ that lies between $\hat{\theta}_n$ and θ so that

$$\begin{aligned} 0 = S_n(\hat{\theta}_n) &= S_n(\theta) + S_n'(\tilde{\theta}_n)(\hat{\theta}_n - \theta) \\ &= S_n(\theta) - I_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta) \quad \mathsf{P}_\theta - \text{a.s.} \end{aligned} \tag{3.31}$$

where the second equality follows from (3.6). From (3.30) and (3.40) we then obtain

$$\frac{I_n(\tilde{\theta}_n)}{I_n(\theta)}\,(\hat{\theta}_n - \theta) \to 0 \quad \mathsf{P}_\theta - \text{a.s.}$$

The strong consistency of $\hat{\theta}_n$ will then follow as long as we can guarantee that the fraction in the last relationship remains bounded away from 0 as $n \to \infty$. However, for every n we have

$$\frac{I_n(\tilde{\theta}_n)}{I_n(\theta)} = \frac{\sum_{i=1}^n J(\tilde{\theta}_n; \mathbf{b}_i)}{\sum_{i=1}^n J(\theta; \mathbf{b}_i)} \geq \frac{nJ_*(\tilde{\theta}_n)}{nJ^*(\theta)} = \frac{J_*(\tilde{\theta}_n)}{J^*(\theta)}.$$

Since $J^*(\theta) > 0$, it suffices to show that $\mathsf{P}_\theta(\liminf_n J_*(\tilde{\theta}_n) > 0) = 1$. Since $J_*(\theta)$ is continuous, positive and bounded away from 0 when $|\theta|$ is bounded away from infinity (recall (2.13)) and $\tilde{\theta}_n$ lies between $\hat{\theta}_n$ and θ, it suffices to show that

$$\mathsf{P}_\theta(\limsup_n |\hat{\theta}_n| = \infty) = 0. \tag{3.32}$$

In order to prove (3.32), we observe first of all that since $S_n(\hat{\theta}_n) = 0$ for large n, (3.30) can be rewritten as follows:

$$\frac{S_n(\theta) - S_n(\hat{\theta}_n)}{I_n(\theta)} \to 0 \quad \mathsf{P}_\theta - \text{a.s.} \tag{3.33}$$

But for every n we have $I_n(\theta) \leq nJ^*(\theta)$ and

$$\begin{aligned} S_n(\theta) - S_n(\hat{\theta}_n) &= \sum_{i=1}^n \Big[s(\theta; \mathbf{b}_i, X_i) - s(\hat{\theta}_n; \mathbf{b}_i, X_i) \Big] \\ &= \sum_{i=1}^n \Big[\bar{a}(\hat{\theta}_n; \mathbf{b}_i) - \bar{a}(\theta; \mathbf{b}_i) \Big] \geq n \inf_{\mathbf{b} \in \mathbb{B}} \Big[\bar{a}(\hat{\theta}_n; \mathbf{b}) - \bar{a}(\theta; \mathbf{b}) \Big], \end{aligned}$$

therefore we obtain

$$\frac{S_n(\theta) - S_n(\hat{\theta}_n)}{I_n(\theta)} \geq \frac{\inf_{\mathbf{b} \in \mathbb{B}} \Big[\bar{a}(\hat{\theta}_n; \mathbf{b}) - \bar{a}(\theta; \mathbf{b}) \Big]}{J^*(\theta)}. \tag{3.34}$$

On the event $\{\limsup_n \hat{\theta}_n = \infty\}$ there exists a subsequence $(\hat{\theta}_{n_j})$ of $(\hat{\theta}_n)$ such that $\hat{\theta}_{n_j} \to \infty$. Consequently, for any $\mathbf{b} \in \mathbb{B}$ we have

$$\lim_{n_j \to \infty} \Big[\bar{a}(\hat{\theta}_{n_j}; \mathbf{b}) - \bar{a}(\theta; \mathbf{b}) \Big] = a^*(\mathbf{b}) - \bar{a}(\theta; \mathbf{b}) > 0 \tag{3.35}$$

and from Lemma 2 we obtain

$$\liminf_{n_j\to\infty} \inf_{\mathbf{b}\in\mathbb{B}} \left[\bar{a}(\hat{\theta}_{n_j};\mathbf{b}) - \bar{a}(\theta;\mathbf{b})\right] \geq \inf_{\mathbf{b}\in\mathbb{B}} [a^*(\mathbf{b}) - \bar{a}(\theta;\mathbf{b})] > 0. \tag{3.36}$$

From (3.34) and (3.36) it follows that

$$\liminf_{n_j\to\infty} \frac{S_{n_j}(\theta) - S_{n_j}(\hat{\theta}_{n_j})}{I_{n_j}(\theta)} > 0$$

and comparing with (3.33) we conclude that $\mathsf{P}_\theta(\limsup_n \hat{\theta}_n = \infty) = 0$. In an identical way we can show that $\mathsf{P}_\theta(\liminf_n \hat{\theta}_n = -\infty) = 0$, which establishes (3.32) and completes the proof of the strong consistency of $\hat{\theta}_n$. In order to prove (3.7), we observe that

$$\begin{aligned}\frac{|I_n(\hat{\theta}_n) - I_n(\theta)|}{I_n(\theta)} &\leq \frac{1}{nJ_*(\theta)} \sum_{i=1}^{n} |J(\hat{\theta}_n;\mathbf{b}_i) - J(\theta;\mathbf{b}_i)| \\ &\leq \frac{1}{J_*(\theta)} \sup_{\mathbf{b}\in\mathbb{B}} |J(\hat{\theta}_n;\mathbf{b}) - J(\theta;\mathbf{b})|.\end{aligned}$$

But since $J(\theta;\mathbf{b})$ is jointly continuous and $\hat{\theta}_n$ strongly consistent, from Lemma 1 it follows that

$$\sup_{\mathbf{b}\in\mathbb{B}} |J(\hat{\theta}_n;\mathbf{b}) - J(\theta;\mathbf{b})| \to 0 \quad \mathsf{P}_\theta - \text{a.s.} \tag{3.37}$$

which completes the proof.

♢

It is interesting to note that (3.29) remains valid for any strongly consistent estimator of θ and that the strong consistency of $\hat{\theta}_n$ is established for any item selection strategy. This is due to the compactness of the item parameter space, $\mathbb{B}$. If this is not the case, the resulting estimator may fail to be consistent (see Chang and Ying (2009) for a counterexample). On the other hand, as we show in the following theorem, the information-maximizing item selection strategy (3.25) guarantees the asymptotic normality and efficiency of $\hat{\theta}_n$.

Theorem 3.3.2. *If $I_n(\theta)/n$ converges in probability to some positive constant under P_θ, then as $n \to \infty$ we have*

$$\sqrt{I_n(\hat{\theta}_n)}\,(\hat{\theta}_n - \theta) \longrightarrow \mathcal{N}(0,1). \tag{3.38}$$

This is in particular the case when the information-maximizing item selection strategy (3.25) *is adopted, in*

which case

$$\sqrt{n}(\hat{\theta}_n - \theta) \to \mathcal{N}\left(0, [J^*(\theta)]^{-1}\right). \tag{3.39}$$

Proof. From Lemma 4 we know that $\{S_n(\theta)\}_{n\in\mathbb{N}}$ is a martingale with bounded increments, mean 0 and predictable variation $I_n(\theta)$. Then, if $I_n(\theta)/n$ converges in probability to some positive constant under P_θ, we can apply the Martingale Central Limit Theorem (see, e.g., Billingsley (2008), Ex. 35.19, p. 481) and obtain

$$\frac{S_n(\theta)}{\sqrt{I_n(\theta)}} \longrightarrow \mathcal{N}(0,1).$$

From Lemma 3 and a Taylor expansion of $S_n(\theta)$ around $\hat{\theta}_n$ it follows that there exists some $\tilde{\theta}_n$ that lies between $\hat{\theta}_n$ and θ so that

$$\begin{aligned} 0 = S_n(\hat{\theta}_n) &= S_n(\theta) + S_n^{'}(\tilde{\theta}_n)(\hat{\theta}_n - \theta) \\ &= S_n(\theta) - I_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta) \quad \mathsf{P}_\theta - \text{a.s.}. \end{aligned} \tag{3.40}$$

where the second equality follows from (3.28). Then we have

$$\frac{I_n(\tilde{\theta}_n)}{I_n(\theta)} \sqrt{I_n(\theta)}\,(\hat{\theta}_n - \theta) \to \mathcal{N}(0,1),$$

where $\tilde{\theta}_n$ lies between $\hat{\theta}_n$ and θ. But, similarly to (3.29) we can show that

$$\frac{I_n(\tilde{\theta}_n)}{I_n(\theta)} \to 1 \quad \mathsf{P}_\theta - \text{a.s.},$$

thus, from an application of Slutsky's theorem we obtain

$$\sqrt{I_n(\theta)}\,(\hat{\theta}_n - \theta) \longrightarrow \mathcal{N}(0,1). \tag{3.41}$$

From (3.41) and (3.29) and another application of Slutsky's theorem we then obtain (3.38).

In order to prove the second part of this theorem, it suffices to show that

$$\frac{1}{n} I_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} J(\theta; \hat{\mathbf{b}}_i) \to J^*(\theta) \quad \mathsf{P}_\theta - \text{a.s}, \tag{3.42}$$

where $(\hat{\mathbf{b}}_i)_{1\le i\le n}$ are the item parameters selected according to the information-maximizing strategy (3.25). In order to prove (3.42) it suffices to show that $J(\theta; \hat{\mathbf{b}}_n) \to J^*(\theta)$ P_θ-a.s. Since $J(\theta; \mathbf{b})$ is jointly continuous

and $\hat{\theta}_n$ strongly consistent, from Lemma 1 it follows that

$$\sup_{\mathbf{b}\in\mathbb{B}} |J(\hat{\theta}_n; \mathbf{b}) - J(\theta; \mathbf{b})| \to 0 \quad \mathsf{P}_\theta - \text{a.s.} \tag{3.43}$$

and, consequently,

$$|J(\hat{\theta}_n; \hat{\mathbf{b}}_n) - J(\theta; \hat{\mathbf{b}}_n)| \leq \sup_{\mathbf{b}\in\mathbb{B}} |J(\hat{\theta}_n; \mathbf{b}) - J(\theta; \mathbf{b})| \to 0 \quad \mathsf{P}_\theta - \text{a.s.},$$

which means that it suffices to show that

$$J(\hat{\theta}_n; \hat{\mathbf{b}}_n) \to J^*(\theta) \quad \mathsf{P}_\theta - \text{a.s.}$$

But from the definition of $(\hat{\mathbf{b}}_n)$ in (3.25) we have $J(\hat{\theta}_{n-1}; \hat{\mathbf{b}}_n) = J^*(\hat{\theta}_{n-1})$, therefore from the triangle inequality we obtain:

$$\begin{aligned} |J(\hat{\theta}_n; \hat{\mathbf{b}}_n) - J^*(\theta)| &\leq |J(\hat{\theta}_n; \hat{\mathbf{b}}_n) - J(\hat{\theta}_{n-1}; \hat{\mathbf{b}}_n)| + |J^*(\hat{\theta}_{n-1}) - J^*(\theta)| \\ &\leq \sup_{\mathbf{b}\in\mathbb{B}} |J(\hat{\theta}_n; \mathbf{b}) - J(\hat{\theta}_{n-1}; \mathbf{b})| + |J^*(\hat{\theta}_{n-1}) - J^*(\theta)|. \end{aligned}$$

From (3.43) it follows that the first term in the upper bound goes to 0 P_θ − a.s.. Moreover, from the continuity of J^* (recall Lemma 2) and the strong consistency of $\hat{\theta}_n$ it follows that the second term in the upper bound goes to 0, which completes the proof.

♢

Remark: The resulting estimator is asymptotically efficient in the sense that if we could employ an oracle item selection method and select each item i so that $J(\theta; \mathbf{b}_i) = J^*(\theta)$, where $J^*(\theta)$ is the maximum Fisher information an item can achieve at the true ability level θ, the asymptotic distribution of the MLE of θ would be the same as (3.39).

3.3.3 Discussion of the design

The proposed design for a CAT based on the nominal response model is similar but not identical to that of Chang and Ying (2009) for CAT based on dichotomous logistic models. It is interesting to point out these differences in the special case of binary items ($m = 2$) in which the nominal response model reduces to the dichotomous 2PL model (recall (3.5)) and each item is characterized by two parameters, the difficulty parameter and the discrimination parameter.

Thus, for a CAT based on the 2PL model, Chang and Ying (2009) select only the difficulty parameter in order to maximize the Fisher information. Moreover, while they assume that the discrimination parameter is bounded, they allow the difficulty parameter to take any value in the real line. This is a convenient but not very realistic assumption, as items are drawn from a given bank and it is not possible to have items of arbitrary difficulty. Finally, their asymptotic analysis relies on a closed-form expression for the difficulty parameter.

On the other hand, in our approach we assume that all components of the item parameter vector are bounded and we establish the consistency of the resulting estimator for an arbitrary item selection strategy. Moreover, in the proposed information-maximizing item selection strategy (3.25), we select all components of the item parameter vector to maximize the Fisher information function. Finally, in our analysis we do not use a closed-form expression for the item parameters defined by (3.25).

3.4 CAT with response revision

3.4.1 A novel CAT

In this section we propose and analyze a novel CAT design in which examinees are allowed to revise their previous answers. As in the case of the conventional CAT presented in the previous section, we consider multiple-choice items with m categories and we assume that the total number of items that will be administered, n, is fixed. However, after each response the examinee now decides whether to revise the answer to a previous item or to proceed to a new item. The only restriction that we impose is that each item can be revised at most $m-2$ times during the test. As a result, we now need to restrict ourselves to items with $m \geq 3$ categories, unlike the previous section where the case of binary items ($m = 2$) was also included. (Note that by "revision" we strictly mean a *change* of a previous answer).

In order to formalize this setup, let us consider the time during the test at which the examinee has completed t responses and let f_t be the number of distinct items that have been administered until then. Then, the number of revisions until this time is $t - f_t$. For each item $i \in \{1, \ldots, f_t\}$, we denote g_t^i as the number of responses that have been given so far and correspond to this particular item. Since each item can be revised up to $m-2$ times, we have $1 \leq g_t^i \leq m-1$. Then, if we let C_t be the set of items that can still be revised at this time, we have $C_t = \{i \in \{1, \ldots, f_t\} : g_t^i < m-1\}$ and the decision of the examinee is

described by the following random variable:

$$d_t := \begin{cases} 0, & \text{the } t+1^{th} \text{ response corresponds to a new item} \\ i, & \text{the } t+1^{th} \text{ response is a revision of item } i \in C_t \end{cases}.$$

For each item $i \in \{1, \ldots, f_t\}$, we denote by X_j^i the category that was chosen at the j^{th} attempt on this item and by $X_{1:j}^i := (X_1^i, \ldots, X_j^i)$ the set of all distinct answers that have been chosen in the first j attempts on this item, where $1 \leq j \leq g_t^i$. Thus, we write $X_j^i = k$ if category $k \notin X_{1:j-1}^i$ was chosen on the j^{th} attempt on item i, where $1 \leq j \leq g_t^i$.

It is important to stress that, in this kind of CAT, information is coming not only from the content of the responses, but also from the decisions of the examinee to revise or not, as well as from the identity of the items that are chosen for revision. Specifically, $\mathcal{G}_t := \sigma\left(d_s,\ 1 \leq s \leq t\right)$. is the σ-algebra that contains the first t decisions of the examinee regarding revision, whereas $\mathcal{F}_t^X := \sigma\left(X_{1:g_t^i}^i\ ,\ 1 \leq i \leq f_t\right)$ is the σ-algebra that contains the first t responses, first answers *and* revisions. Then, $\mathcal{F}_t := \mathcal{G}_t \vee \mathcal{F}_t^X$ is the σ-algebra that contains all available information after t responses. Note that the number of items that have been administered until this time, f_t, is $\mathcal{G}_{t-1}$-measurable, since it can be fully recovered by $d_1, \ldots, d_{t-1}$.

For each $i \in \{1, \ldots, n-1\}$, item $i+1$ needs to be selected at the time that the examinee has answered i distinct items and does not want or is not allowed to revise any more items, that is at the $\{\mathcal{G}_t\}$-stopping time

$$\tau_i := \min\{t \geq 1 :\ f_t = i \quad \text{and} \quad d_t = 0\}.$$

Since the total number of items that will be administered is n, the test stops at the random time τ_n, which is determined by the test-taker's *revision strategy*, $(d_t)_{1 \leq t \leq \tau_n}$. Our goal is to propose a design that will guarantee the reliable estimation of the test-taker's ability *for any revision strategy*, that is no matter when and what the test-taker chooses to revise. Thus, we will not in general make any modeling assumptions about the revision strategy, $(d_t)_{1 \leq t \leq \tau_n}$. Instead, we will postulate a statistical model for the responses of the test-taker, i.e., the first answer to each item and *any subsequent revisions*.

3.4.2 The proposed design

As in the previous section, we assume that the first response to each item is governed by the nominal response model, so that for every item $i \in \{1, \ldots, n\}$ we have

$$\mathsf{P}_\theta(X_1^i = k \,|\, \mathbf{b}_i) = p_k(\theta; \mathbf{b}_i), \quad 1 \leq k \leq m, \tag{3.44}$$

where p_k is the pmf of the nominal response model defined by (3.1) - (3.3), θ an unknown scalar parameter that represents the ability of the test-taker and $\mathbf{b}_i := (a_{ik}, c_{ik})_{2\leq k\leq m}$ a $\mathbb{B}$-valued vector that characterizes item i. The item parameter $\mathbf{b}_{i+1}$ needs to be selected at time τ_i based on all the available information until this time. Thus, we will say that $(\mathbf{b}_i)_{2\leq i\leq n}$ is an *item selection strategy* if $\mathbf{b}_{i+1}$ is a $\mathbb{B}$-valued, $\mathcal{F}_{\tau_i}$-measurable random vector for every $1 \leq i \leq n-1$. The ultimate goal is to obtain a good ability estimator at the end of the test, i.e., an $\mathcal{F}_{\tau_n}$-measurable statistic $\hat{\theta}_n$ that will be close to the true ability θ under minimal assumptions on the behavior of the examinee.

With respect to item selection in this novel CAT, we propose the same, information-maximizing approach as in the conventional CAT of the previous section. That is, we suggest that item $i+1$ should be selected so that

$$\hat{\mathbf{b}}_{i+1} = \underset{\mathbf{b}\in\mathbb{B}}{\operatorname{argmax}}\, J(\hat{\theta}_{\tau_i}; \mathbf{b}), \tag{3.45}$$

where J is the Fisher information function of the nominal response model given by (3.19) and $\hat{\theta}_{\tau_i}$ is an estimate of θ that is based on all the available information up to the time of selection, i.e., an $\mathcal{F}_{\tau_i}$-measurable statistic. Therefore, the item selection method (3.45) requires an estimate of θ at all the times that items are selected, $(\tau_i)_{1\leq i\leq n}$.

For the adaptive estimation of θ we will use the maximizer of the *partial* likelihood of all observed responses conditionally on the selected items and *the revision decisions of the examinee*. We will describe the proposed estimator for an arbitrary item selection strategy, not necessarily (3.45), and at every time t, not only at $(\tau_i)_{1\leq i\leq n}$. Thus, for any revision strategy $(d_t)_{1\leq t\leq \tau_n}$ and any item selection strategy $(\mathbf{b}_i)_{1\leq i\leq n}$, we suggest updating the ability of the examinee after t responses with the maximizer of

$$L_t(\theta) := \log \mathsf{P}_\theta \left(X^i_{1:g^i_t}\,,\ 1 \leq i \leq f_t \;\middle|\; \mathcal{G}_t, \mathbf{b}_{1:f_t} \right). \tag{3.46}$$

As in the case of a traditional CAT, we assume that responses coming from different items are conditionally independent so that

$$\mathsf{P}_\theta \left(X^i_{1:g^i_t}\,,\ 1 \leq i \leq f_t \;\middle|\; \mathcal{G}_t, \mathbf{b}_{1:f_t} \right) = \prod_{i=1}^{f_t} \mathsf{P}_\theta \left(X^i_{1:g^i_t} \mid \mathcal{G}_t, \mathbf{b}_i \right). \tag{3.47}$$

Moreover, we assume that the responses on a given item are independent of the revision strategy of the

examinee, in the sense that for every item $i \in \{1, \ldots, f_t\}$ we have

$$\begin{aligned} \mathsf{P}_\theta \left(X^i_{1:g^i_t} \,|\, \mathcal{G}_t, \mathbf{b}_i \right) &= \mathsf{P}_\theta \left(X^i_{1:g^i_t} \,|\, \mathbf{b}_i \right) \\ &= \mathsf{P}_\theta(X^i_1 \,|\, \mathbf{b}_i) \cdot \prod_{j=2}^{g^i_t} \mathsf{P}_\theta \left(X^i_j \,|\, X^i_{1:j-1}, \mathbf{b}_i \right). \end{aligned} \tag{3.48}$$

The second equality follows from the definition of conditional probability and it is understood that the second factor in the right-hand side is equal to 1 whenever $g^i_t = 1$. Each probability $\mathsf{P}_\theta(X^i_1 \,|\, \mathbf{b}_i)$ is determined by (3.44), according to which the first answer to each item is governed by the nominal response model. Thus, it remains to specify the contribution of the revised answers. We assume that the nominal response model also determines revisions, in the sense that

$$\mathsf{P}_\theta \left(X^i_j = k \,|\, X^i_{1:j-1}, \mathbf{b}_i \right) = \frac{p_k(\theta; \mathbf{b}_i)}{\sum_{h \notin X^i_{1:j-1}} p_h(\theta; \mathbf{b}_i)}, \quad k \notin X^i_{1:j-1}. \tag{3.49}$$

Assumptions (3.44), (3.47), (3.48) and (3.49) imply that the conditional log-likelihood function, $L_t(\theta)$, takes the form

$$L_t(\theta) = \sum_{i=1}^{f_t} \Big[\ell \left(\theta; \mathbf{b}_i, X^i_1 \right) + 1_{\{g^i_t \geq 2\}} \sum_{j=2}^{g^i_t} \ell(\theta; \mathbf{b}_i, X^i_j | X^i_{1:j-1}) \Big], \tag{3.50}$$

where $\ell(\theta; \mathbf{b}_i, X^i_1)$ is defined according to (3.4) and for every $2 \leq j \leq g^i_t$ we set

$$\ell(\theta; \mathbf{b}_i, X^i_j \,|\, X^i_{1:j-1}) := \log \mathsf{P}_\theta \left(X^i_j \,|\, X^i_{1:j-1}, \mathbf{b}_i \right).$$

The corresponding score function takes the form

$$S_t(\theta) := \frac{d}{d\theta} L_t(\theta) = \sum_{i=1}^{f_t} \Big[s \left(\theta; \mathbf{b}_i, X^i_1 \right) + 1_{\{g^i_t \geq 2\}} \sum_{j=2}^{g^i_t} s \left(\theta; \mathbf{b}_i, X^i_j \,|\, X^i_{1:j-1} \right) \Big], \tag{3.51}$$

where $s(\theta; \mathbf{b}_i, X^i_1)$ is defined according to (3.18) and for every $2 \leq j \leq g^i_t$ we have

$$s(\theta; \mathbf{b}_i, X^i_j \,|\, X^i_{1:j-1}) := \frac{d}{d\theta} \ell \left(\theta; \mathbf{b}_i, X^i_j \,|\, X^i_{1:j-1} \right). \tag{3.52}$$

Our estimate for θ after t responses, $\hat{\theta}_t$, will be the root of the score function $S_t(\theta)$. Similarly to the case of the convention CAT, the root exists and is unique for every t that is larger than some random time and a preliminary estimation procedure is needed until this time. However, similarly to Lemma 3, we can show

that $\hat{\theta}_{\tau_n}$ is the root of $S_{\tau_n}(\theta)$ for all large n with probability 1.

3.4.3 Discussion of the proposed design

Assumptions (3.44) - (3.47) are analogous to (3.23) - (3.24) in the context of a conventional CAT. The additional modeling assumptions that we impose in the novel CAT are (3.48) and (3.49). Assumption (3.48) states that the responses on any given item do not depend on the decisions of the examinee regarding revision. As a result, the statistical model for the revised answer does not depend on when the test-taker chooses to revise. Assumption (3.49) is the statistical model for the revised answers. It implies that an examinee of high ability is more likely to revise in the right direction than an examinee of low ability. However, its main advantage is that it does not introduce any additional item parameters to the ones that would be used in the conventional CAT of the previous section.

Contrary to previous CAT designs that allow for response revision, the proposed method takes into account all responses of the examinee on a given item during the test, not only the last one. Therefore, while examinees will benefit by revising wrong answers, they have to be cautious with revisions, as they cannot switch back to a previous answer during the test.

While we model and incorporate the content of revised answers in our estimation method, the decisions of the examinee regarding revision are not incorporated in the estimator. That is, we do not make any assumption regarding when and what items the test-taker chooses to revise. Incorporating such information could lead to alternative estimators and item selection methods. However, it would make the design much more vulnerable to model mispecification, as it is not at all clear how to specify universal models for the behavior of the examinee. Moreover, it could introduce additional parameters that would have to be calibrated in a pretesting period, complicating the implementation of this modified CAT design in practice.

Overall, the proposed design preserves the complexity of a conventional CAT that is based on the nominal response model, *as it does not require any additional calibration effort*, which is a very desirable property for practical implementation. Our next goal is to show that it also preserves the statistical efficiency of a conventional CAT under minimal assumptions on the revision strategy.

3.4.4 Asymptotic properties

In this section, we assume that assumptions (3.44), (3.47), (3.48) and (3.49) hold and we study the asymptotic behavior of the final ability estimator. Specifically, we establish its strong consistency for any item selection strategy and revision behavior and its asymptotic normality when the items are selected according to (3.45) and the total number of revisions is small relative to the total number of distinct items.

First, in Lemma 5 we show that the conditional score function, $S_t(\theta)$, is a martingale with predictable variation equal to the conditional Fisher information

$$I_t(\theta) := \sum_{i=1}^{f_t} \left[J(\theta; \mathbf{b}_i) + 1_{\{g_t^i \geq 2\}} \sum_{j=2}^{g_t^i} J\left(\theta; \mathbf{b}_i \,|\, X_{1:j-1}^i\right) \right] = \frac{d}{d\theta} S_t(\theta), \tag{3.53}$$

where $J(\theta; \mathbf{b}_i)$ is defined as in (3.27) and, for every $2 \leq j \leq g_t^i$, we set

$$J\left(\theta; \mathbf{b}_i \,|\, X_{1:j-1}^i\right) := \mathsf{Var}_\theta \left[s(\theta; \mathbf{b}_i, X_j^i \,|\, X_{1:j-1}^i) \right], \tag{3.54}$$

where $s(\theta; \mathbf{b}_i, X_j^i \,|\, X_{1:j-1}^i)$ is defined in (3.52).

Lemma 5. *For any item selection strategy and any revision strategy,*

(i) $\{S_t(\theta)\}_{t\in\mathbb{N}}$ *is a* $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$*-martingale under* P_θ *with bounded increments, mean zero and predictable variation*

$$\langle S(\theta) \rangle_t := \sum_{s=1}^{t} \mathsf{E}_\theta \left[\left(S_s(\theta) - S_{s-1}(\theta)\right)^2 | \mathcal{F}_{s-1} \right] = I_t(\theta).$$

(ii) $\{S_{\tau_n}(\theta)\}_{n\in\mathbb{N}}$ *is a* $\{\mathcal{F}_{\tau_n}\}_{n\in\mathbb{N}}$*-martingale with mean 0 and predictable variation* $\{I_{\tau_n}(\theta)\}_{n\in\mathbb{N}}$.

Proof. (i) After $t-1$ responses, the examinee either proceeds to a new item or revises a previous item. Therefore, the difference $S_t(\theta) - S_{t-1}(\theta)$ admits the following decomposition:

$$s\left(\theta; \mathbf{b}_{f_t}, X_1^{f_t}\right) \mathbb{1}_{\{d_{t-1}=0\}} + \sum_{i\in C_{t-1}} s\left(\theta; \mathbf{b}_i, X_{g_t^i}^i | X_{1:g_t^i-1}^i\right) \mathbb{1}_{\{d_{t-1}=i\}}, \tag{3.55}$$

where the sum in the second term is understood to be 0 when C_{t-1} is the empty set. Since d_{t-1}, C_{t-1} are $\mathcal{F}_{t-1}$-measurable, taking conditional expectations with respect to $\mathcal{F}_{t-1}$ we obtain

$$\begin{aligned} \mathsf{E}_\theta[S_t(\theta) - S_{t-1}(\theta)|\mathcal{F}_{t-1}] = {} & \mathsf{E}_\theta \left[s\left(\theta; \mathbf{b}_{f_t}, X_1^{f_t}\right) \,\middle|\, \mathcal{F}_{t-1} \right] \mathbb{1}_{\{d_{t-1}=0\}} \\ & + \sum_{i\in C_{t-1}} \mathsf{E}_\theta \left[s\left(\theta; \mathbf{b}_i, X_{g_t^i}^i | X_{1:g_t^i-1}^i\right) \,\middle|\, \mathcal{F}_{t-1} \right] \mathbb{1}_{\{d_{t-1}=i\}}. \end{aligned}$$

Since f_t and g_t^i are $\mathcal{F}_{t-1}$-measurable, it follows that

$$\mathsf{E}_\theta \left[s\left(\theta; \mathbf{b}_{f_t}, X_1^{f_t}\right) \,\middle|\, \mathcal{F}_{t-1} \right] = 0 = \mathsf{E}_\theta \left[s\left(\theta; \mathbf{b}_i, X_{g_t^i}^i | X_{1:g_t^i-1}^i\right) \,\middle|\, \mathcal{F}_{t-1} \right],$$

which proves that $S_t(\theta)$ is a zero-mean $\mathcal{F}_t$-martingale under P_θ. From (3.55) we also have

$$\begin{aligned}&\mathsf{E}_\theta[(S_t(\theta) - S_{t-1}(\theta))^2 \,|\mathcal{F}_{t-1}] \\ &= J(\theta; \mathbf{b}_{f_t})\, \mathbb{1}_{\{d_{t-1}=0\}} + \sum_{i \in C_{t-1}} J\left(\theta; \mathbf{b}_i | X^i_{1:g^i_t - 1}\right)\, \mathbb{1}_{\{d_{t-1}=i\}}\end{aligned}$$

and, consequently, the predictable variation of $S_t(\theta)$ will be

$$\begin{aligned}\langle S(\theta)\rangle_t &:= \sum_{s=1}^{t} \mathsf{E}_\theta\left[(S_s(\theta) - S_{s-1}(\theta))^2 \,|\mathcal{F}_{s-1}\right] \\ &= \sum_{s=1}^{t} \left[J(\theta; \mathbf{b}_{f_s})\, \mathbb{1}_{\{d_{s-1}=0\}} + \sum_{j \in C_{s-1}} J\left(\theta; \mathbf{b}_j | X^j_{1:g^j_{s-1}}\right)\, \mathbb{1}_{\{d_{s-1}=j\}} \right]. \\ &= I_t.\end{aligned}$$

(ii) This follows from the Optional Sampling Theorem and the fact that $(\tau_n)_{n\in\mathbb{N}}$ is a strictly increasing sequence of $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$-stopping times that are bounded, since $\tau_n \leq (m-1)n$ for every $n \in \mathbb{N}$.

♢

In the next lemma, we collect the main properties of the conditional score function and the conditional Fisher information that will be needed in our analysis.

Lemma 6. *Fix $\mathbf{b}_i \in \mathbb{B}$, $j \in \{2, \ldots, m-1\}$, and $X^i_{1:j-1}$ and let*

$$\bar{a}(\theta; \mathbf{b}_i | X^i_{1:j-1}) := \sum_{k \notin X^i_{1:j-1}} a_{ki}\, \mathsf{P}_\theta\left(X^i_j = k \,|\, X^i_{1:j-1}, \mathbf{b}_i\right).$$

(i) The conditional score in (3.52) *and the conditional Fisher information* (3.54) *admit the following representations*

$$\begin{aligned}s(\theta; \mathbf{b}_i, X^i_j = k \,|\, X^i_{1:j-1}) &= a_{ki} - \bar{a}(\theta; \mathbf{b}_i | X^i_{1:j-1}), \quad k \notin X^i_{1:j-1} \\ J\left(\theta; \mathbf{b}_i \,|\, X^i_{1:j-1}\right) &= \sum_{k \notin X^i_{1:j-1}} \left(a_k - \bar{a}(\theta; \mathbf{b}_i \,|X^i_{1:j-1})\right)^2 \mathsf{P}_\theta\left(X^i_j = k \,|\, X^i_{1:j-1}, \mathbf{b}_i\right).\end{aligned}$$

Moreover, they are both bounded by a constant that does not depend on θ or $\mathbf{b}_i$.

(ii) $\bar{a}(\theta; \mathbf{b}_i | X^i_{1:j-1}) \to a_*(\mathbf{b}_i)$ as $\theta \to -\infty$ and $\bar{a}(\theta; \mathbf{b}_i | X^i_{1:j-1}) \to a^*(\mathbf{b}_i)$ as $\theta \to +\infty$.

The proof of Lemma 6 follows by direct computation. We can now establish the strong consistency of $\hat{\theta}_{\tau_n}$ as $n \to \infty$ without any conditions on the item selection or the revision strategy.

Theorem 3.4.1. *For any item selection method and any revision strategy, as $n \to \infty$ we have*

$$\hat{\theta}_{\tau_n} \to \theta \quad \textit{and} \quad \frac{I_{\tau_n}(\hat{\theta}_{\tau_n})}{I_{\tau_n}(\theta)} \to 1 \quad \mathsf{P}_\theta\textit{-a.s.} \tag{3.56}$$

Proof. From Lemma 5 we have that $S_{\tau_n}(\theta)$ is a $\{\mathcal{F}_{\tau_n}\}$-martingale with predictable variation $I_{\tau_n}(\theta)$. Moreover, from (4.10) we have $I_{\tau_n}(\theta) \geq nJ_*(\theta) \to \infty$ and from the Martingale Strong Law of Large Numbers (Williams, D. (1991), p. 124) it follows that

$$\frac{S_{\tau_n}(\theta)}{I_{\tau_n}(\theta)} \to 0 \quad \mathsf{P}_\theta - \text{a.s.} \tag{3.57}$$

Since $S_{\tau_n}(\hat{\theta}_{\tau_n}) = 0$ for large enough n with probability 1, with a Taylor expansion around θ we have

$$\begin{aligned} 0 = S_{\tau_n}(\hat{\theta}_{\tau_n}) &= S_{\tau_n}(\theta) + S'_{\tau_n}(\tilde{\theta}_{\tau_n})(\hat{\theta}_{\tau_n} - \theta) \\ &= S_{\tau_n}(\theta) - I_{\tau_n}(\tilde{\theta}_{\tau_n})(\hat{\theta}_{\tau_n} - \theta) \quad \mathsf{P}_\theta - \text{a.s.} \end{aligned} \tag{3.58}$$

where $\tilde{\theta}_{\tau_n}$ lies between $\hat{\theta}_{\tau_n}$ and θ, and (3.57) takes the form

$$\frac{I_{\tau_n}(\tilde{\theta}_{\tau_n})}{I_{\tau_n}(\theta)} (\hat{\theta}_{\tau_n} - \theta) \to 0 \quad \mathsf{P}_\theta - \text{a.s.}$$

However, since $\tau_n \leq (m-1)n$ and $J_*(\theta) f_t \leq I_t(\theta) \leq Kt$ for every t, we have

$$\frac{I_{\tau_n}(\tilde{\theta}_{\tau_n})}{I_{\tau_n}(\theta)} \geq \frac{nJ_*(\tilde{\theta}_{\tau_n})}{\tau_n K} \geq \frac{1}{(m-1)K} J_*(\tilde{\theta}_{\tau_n})$$

and it suffices to show that

$$\limsup_n |\hat{\theta}_{\tau_n}| < \infty \quad \mathsf{P}_\theta - \text{a.s.} \tag{3.59}$$

For large n we have $S_{\tau_n}(\hat{\theta}_{\tau_n}) = 0$ and (3.57) can be rewritten as follows

$$\frac{S_{\tau_n}(\theta) - S_{\tau_n}(\hat{\theta}_{\tau_n})}{I_{\tau_n}(\theta)} \to 0 \quad \mathsf{P}_\theta - \text{a.s.} \tag{3.60}$$

But from the definition of the score function in (4.8) it follows that

$$
\begin{aligned}
&S_{\tau_n}(\theta) - S_{\tau_n}(\hat{\theta}_{\tau_n}) \\
&= \sum_{i=1}^{n} \left[\Big(s(\theta; \mathbf{b}_i) - s(\hat{\theta}_{\tau_n}; \mathbf{b}_i) \Big) + \sum_{j=2}^{g^i_{\tau_n}} \Big(s(\theta; \mathbf{b}_i, X^i_j | X^i_{1:j-1}) - s(\hat{\theta}_{\tau_n}; \mathbf{b}_i, X^i_j | X^i_{1:j-1}) \Big) \right] \\
&= \sum_{i=1}^{n} \left[\Big(\bar{\alpha}(\hat{\theta}_{\tau_n}; \mathbf{b}_i) - \bar{\alpha}(\theta; \mathbf{b}_i) \Big) + \sum_{j=2}^{g^i_{\tau_n}} \Big(\bar{\alpha}(\hat{\theta}_{\tau_n}; \mathbf{b}_i | X^i_{1:j-1}) - \bar{\alpha}(\theta; \mathbf{b}_i | X^i_{1:j-1}) \Big) \right] \\
&\geq n \inf_{\mathbf{b} \in \mathbb{B}} \Big[\bar{\alpha}(\hat{\theta}_{\tau_n}; \mathbf{b}) - \bar{\alpha}(\theta; \mathbf{b}) \Big] \\
&+ (\tau_n - n) \min_{2 \leq j \leq m-1} \min_{X_{1:j-1}} \inf_{\mathbf{b} \in \mathbb{B}} \Big[\bar{\alpha}(\hat{\theta}_{\tau_n}; \mathbf{b} \,|\, X_{1:j-1}) - \bar{\alpha}(\theta; \mathbf{b} | X_{1:j-1}) \Big],
\end{aligned}
$$

where $X_{1:j-1} := (X_1, \ldots, X_{j-1})$ is a vector of $j-1$ distinct responses on an item with parameter $\mathbf{b}$. On the other hand, $I_{\tau_n}(\theta) \leq \tau_n K$, which implies that

$$
\begin{aligned}
\frac{S_{\tau_n}(\theta) - S_{\tau_n}(\hat{\theta}_{\tau_n})}{I_{\tau_n}(\theta)} &\geq \frac{1}{K} \inf_{\mathbf{b} \in \mathbb{B}} [\bar{\alpha}(\hat{\theta}_{\tau_n}; \mathbf{b}) - \bar{\alpha}(\theta; \mathbf{b})] \\
&\quad + \frac{1}{K} \min_{2 \leq j \leq m-1} \min_{X_{1:j-1}} \inf_{\mathbf{b} \in \mathbb{B}} \Big[\bar{\alpha}(\hat{\theta}_{\tau_n}; \mathbf{b} \,|\, X_{1:j-1}) - \bar{\alpha}(\theta; \mathbf{b} | X_{1:j-1}) \Big].
\end{aligned}
$$

On the event $\{\limsup_n \hat{\theta}_{\tau_n} \to \infty\}$ there is a subsequence $(\hat{\theta}_{\tau_{n_j}})$ of $(\hat{\theta}_{\tau_n})$ so that $\hat{\theta}_{\tau_{n_j}} \to \infty$ and from (3.36) we have

$$
\liminf_{n_j \to \infty} \inf_{\mathbf{b} \in \mathbb{B}} \Big[\bar{\alpha}(\hat{\theta}_{\tau_{n_j}}; \mathbf{b}) - \bar{\alpha}(\theta; \mathbf{b}) \Big] > 0.
$$

Similarly, due to Lemma 6 (ii), for any $2 \leq j \leq m-1$ and $X_{1:j-1}$ we have

$$
\liminf_{n_j \to \infty} \inf_{\mathbf{b} \in \mathbb{B}} \Big[\bar{\alpha}(\hat{\theta}_{\tau_{n_j}}; \mathbf{b} \,|\, X_{1:j-1}) - \bar{\alpha}(\theta; \mathbf{b} \,|\, X_{1:j-1}) \Big] \geq 0.
$$

Therefore,

$$
\liminf_{n_j} \frac{S_{\tau_{n_j}}(\theta) - S_{\tau_{n_j}}(\hat{\theta}_{\tau_{n_j}})}{I_{\tau_{n_j}}(\theta)} > 0
$$

and comparing with (3.60) we conclude that $\mathsf{P}(\limsup_n \hat{\theta}_{\tau_n} = \infty) = 0$. Similarly we can show that $\mathsf{P}(\limsup_n \hat{\theta}_{\tau_n} = -\infty) = 0$, which proves (3.59) and, consequently, the strong consistency of $\hat{\theta}_{\tau_n}$. In or-

der to prove the second claim of the theorem, we need to show that

$$\frac{|I_{\tau_n}(\hat{\theta}_{\tau_n}) - I_{\tau_n}(\theta)|}{I_{\tau_n}(\theta)} \to 0 \quad \mathsf{P}_\theta - \text{a.s.} \tag{3.61}$$

But $I_{\tau_n}(\theta) \geq n\, J_*(\theta)$, whereas $|I_{\tau_n}(\hat{\theta}_{\tau_n}) - I_{\tau_n}(\theta)|$ is bounded above by

$$\begin{aligned}
&\sum_{i=1}^{n} |J(\hat{\theta}_{\tau_n}; \mathbf{b}_i) - J(\theta; \mathbf{b}_i)| + \sum_{i=1}^{n} \sum_{j=2}^{g^i_{\tau_n}} \Big| J(\hat{\theta}_{\tau_n}; \mathbf{b}_i | X^i_{1:j-1}) - J(\theta; \mathbf{b}_i | X^i_{1:j-1}) \Big| \\
&\leq n \sup_{\mathbf{b} \in \mathbb{B}} |J(\hat{\theta}_{\tau_n}; \mathbf{b}) - J(\theta; \mathbf{b})| \\
&+ (\tau_n - n) \max_{2 \leq j \leq m-1} \max_{X_{1:j-1}} \sup_{\mathbf{b} \in \mathbb{B}} \Big| J(\hat{\theta}_{\tau_n}; \mathbf{b} | X_{1:j-1}) - J(\theta; \mathbf{b} | X_{1:j-1}) \Big|,
\end{aligned}$$

where again $X_{1:j-1} := (X_1, \ldots, X_{j-1})$ is a vector of $j-1$ distinct responses on an item with parameter $\mathbf{b}$. Therefore, the ratio in (3.61) is bounded above by

$$\begin{aligned}
&\frac{1}{J_*(\theta)} \sup_{\mathbf{b} \in \mathbb{B}} |J(\hat{\theta}_{\tau_n}; \mathbf{b}) - J(\theta; \mathbf{b})| \\
&+ \frac{m-2}{J_*(\theta)} \max_{2 \leq j \leq m-1} \max_{X_{1:j-1}} \sup_{\mathbf{b} \in \mathbb{B}} \Big| J(\hat{\theta}_{\tau_n}; \mathbf{b} | X_{1:j-1}) - J(\theta; \mathbf{b} | X_{1:j-1}) \Big|.
\end{aligned}$$

But similarly to (3.37) we can show that

$$\sup_{\mathbf{b} \in \mathbb{B}} |J(\hat{\theta}_{\tau_n}; \mathbf{b}) - J(\theta; \mathbf{b})| \to 0 \quad \mathsf{P}_\theta - \text{a.s.}$$

as well as that for every $2 \leq j \leq m-1$ and $X_{1:j-1}$ we have

$$\sup_{\mathbf{b} \in \mathbb{B}} |J(\hat{\theta}_{\tau_n}; \mathbf{b} \,|\, X_{1:j-1}) - J(\theta; \mathbf{b} \,|\, X_{1:j-1})| \to 0 \quad \mathsf{P}_\theta - \text{a.s.}$$

which completes the proof. ♢

In the following theorem we establish the asymptotic normality of the proposed estimator under a stability condition on the Fisher information that is satisfied when we select each item according to (3.45) and the total number of revisions is small relative to the number of distinct items.

Theorem 3.4.2. *If $I_{\tau_n}(\theta)/n$ converges in probability to a positive number, then*

$$\sqrt{I_{\tau_n}(\hat{\theta}_{\tau_n})}\,(\hat{\theta}_{\tau_n} - \theta) \to \mathcal{N}(0, 1). \tag{3.62}$$

This is true in particular when items are selected according to the information-maximizing item selection strategy (3.45) *and the number of revisions is much smaller than the number of items, in the sense that* $\tau_n - n = o_p(n)$*, in which case we also have*

$$\sqrt{n}(\hat{\theta}_{\tau_n} - \theta) \to \mathcal{N}\left(0, [J^*(\theta)]^{-1}\right). \tag{3.63}$$

Proof. Let us assume for the moment that as $n \to \infty$

$$\frac{S_{\tau_n}(\theta)}{\sqrt{I_{\tau_n}(\theta)}} \to \mathcal{N}(0,1). \tag{3.64}$$

Since $S_{\tau_n}(\hat{\theta}_{\tau_n}) = 0$ for large enough n with probability 1, with a Taylor expansion around θ we have

$$\begin{aligned} 0 = S_{\tau_n}(\hat{\theta}_{\tau_n}) &= S_{\tau_n}(\theta) + S'_{\tau_n}(\tilde{\theta}_{\tau_n})(\hat{\theta}_{\tau_n} - \theta) \\ &= S_{\tau_n}(\theta) - I_{\tau_n}(\tilde{\theta}_{\tau_n})(\hat{\theta}_{\tau_n} - \theta) \quad \mathsf{P}_\theta - \text{a.s.} \end{aligned} \tag{3.65}$$

where $\tilde{\theta}_{\tau_n}$ lies between $\hat{\theta}_{\tau_n}$ and θ. From (3.64) and (3.65) we obtain

$$\frac{I_{\tau_n}(\tilde{\theta}_{\tau_n})}{I_{\tau_n}(\theta)} \sqrt{I_{\tau_n}(\theta)}\,(\hat{\theta}_{\tau_n} - \theta) \to \mathcal{N}(0,1).$$

Thus, from (3.56) it follows that the ratio in the left-hand side goes to 1 almost surely and from Slutsky's theorem we obtain (3.62). Therefore, in order to prove the first part of the theorem, it suffices to show that if $I_{\tau_n}(\theta)/n$ converges in probability to some positive number, then (3.64) holds. In order to do so, we define the martingale-difference array

$$Y_{nt} := \frac{S_t(\theta) - S_{t-1}(\theta)}{\sqrt{n}} 1_{\{t \le \tau_n\}}, \quad t \in \mathbb{N}, \quad n \in \mathbb{N}.$$

Indeed, since $\{S_t(\theta)\}$ is an $\{\mathcal{F}_t\}$-martingale and τ_n an $\{\mathcal{F}_t\}$-stopping time, then $\{t \le \tau_n\} = \{\tau_n \le t-1\}^c \in \mathcal{F}_{t-1}$ and, consequently, we have

$$\mathsf{E}_\theta[Y_{nt}|\mathcal{F}_{t-1}] = \frac{1_{\{t \le T_n\}}}{\sqrt{n}}\, \mathsf{E}_\theta[S_t(\theta) - S_{t-1}(\theta)\,|\,\mathcal{F}_{t-1}] = 0.$$

Moreover, the increments of $\{S_t(\theta)\}_{t\in\mathbb{N}}$ are uniformly bounded, which implies that for every $\epsilon > 0$ we have

as $n \to \infty$

$$\sum_{t=1}^{\infty} \mathsf{E}_\theta[Y_{nt}^2 \, \mathbb{1}_{\{|Y_{nt}|>\epsilon\}}] \to 0. \tag{3.66}$$

Therefore, from the Martingale Central Limit Theorem (see, e.g. Theorem 35.12 in Billingsley (2008)) and Slutsky's theorem it follows that if $\sum_{t=1}^{\infty} \mathsf{E}_\theta[Y_{nt}^2 \,|\, \mathcal{F}_{t-1}]$ converges in probability to a positive number, then

$$\sqrt{\frac{n}{I_{\tau_n}(\theta)}} \sum_{t=1}^{\infty} Y_{nt} \longrightarrow \mathcal{N}(0,1).$$

But

$$\sum_{t=1}^{\infty} \mathsf{E}_\theta[Y_{nt}^2 \,|\, \mathcal{F}_{t-1}] = \frac{1}{n} \sum_{t=1}^{\tau_n} \mathsf{E}_\theta\left[(S_t(\theta) - S_{t-1}(\theta))^2 \,|\, \mathcal{F}_{t-1}\right] = \frac{I_{\tau_n}(\theta)}{n}$$

and

$$\sqrt{\frac{n}{I_{\tau_n}(\theta)}} \sum_{t=1}^{\infty} Y_{nt} = \frac{1}{\sqrt{I_{\tau_n}(\theta)}} \sum_{t=1}^{\tau_n} [S_t(\theta) - S_{t-1}(\theta)] = \frac{S_{\tau_n}(\theta)}{\sqrt{I_{\tau_n}(\theta)}},$$

which completes the proof of the first part of the theorem. In order to prove the second part, it suffices to show that

$$\frac{I_{\tau_n}(\theta)}{n} \to J^*(\theta)$$

in probability as $n \to \infty$. Recall from (3.53) that the Fisher information function can be decomposed in the following way:

$$I_{\tau_n}(\theta) = \sum_{i=1}^{n} J(\theta; \mathbf{b}_i) + I_{\tau_n}^R(\theta),$$

where $I_{\tau_n}^R(\theta)$ is the part of the information coming from revisions, i.e.,

$$I_{\tau_n}^R(\theta) := \sum_{i=1}^{n} 1_{\{g_{\tau_n}^i \geq 2\}} \sum_{j=2}^{g_{\tau_n}^i} J\left(\theta; \mathbf{b}_i \,|\, X_{1:j-1}^i\right). \tag{3.67}$$

Let $(\hat{\mathbf{b}}_i)_{2 \leq i \leq n}$ the information maximizing item selection strategy defined in (3.45). Then, from (3.42) we

have

$$\frac{1}{n}\sum_{i=1}^{n} J(\theta, \hat{\mathbf{b}}_i) \to J^*(\theta) \quad \mathsf{P}_\theta - \text{a.s.}$$

whereas from (3.67) we can see that for any revision strategy we have

$$\frac{1}{n} I^R_{\tau_n}(\theta) \leq K \frac{\tau_n - n}{n},$$

where K is some constant that does not depend on θ. The upper bound goes to 0 in probability when $\tau_n - n = o_p(n)$, which completes the proof.

♢

3.5 Numerical Examples

In this section, we present the results of two simulation studies in which we compare the proposed CAT design that allows for response revision, to which we refer as RCAT, with that of a conventional CAT that does not allow for response revision, when both designs are based on the nominal response model (3.1). Specifically, in the first study we illustrate our asymptotic results, whereas in the second study we compare the two designs in a realistic setup.

For both studies, when revision is allowed we assume that at most n_1 items can be revised during the test and that the examinee decides to revise a previous answer after the t^{th} response with probability p_t that satisfies the following recursion

$$p_{t+1} = p_t - 0.5/n_1, \quad p_1 = 0.5.$$

For n_1, we consider the following cases: $n_1/n = 0.1, 0.5, 1$. Moreover, whenever the examinee decides to revise, we assume that each of the previous items which can still be revised at time t are equally likely to be selected for revision. The revised responses were simulated according to the conditional probability model (3.49), which implies that examinees may revise either from a wrong answer to the correct one, or from the correct answer to a wrong one, depending on their true ability. We replicated the two studies for ability values in $\{-3, -2, -1, 0, 1, 2, 3\}$. For each scenario, we computed the root mean square error (RMSE) of the final ability estimator, that is $\sqrt{\mathsf{E}_\theta\left[(\hat{\theta}_n - \theta))^2\right]}$ and $\sqrt{\mathsf{E}_\theta\left[(\hat{\theta}_{\tau_n} - \theta))^2\right]}$ for CAT and RCAT respectively, on the basis of $1,000$ simulation runs (examinees).

3.5.1 An idealized item pool

In the first study we considered an idealized item pool of items that was simulated based on Passos et al. (2007). Each item has $m = 3$ categories, which means that each item can be revised at most once whenever revision is allowed. The parameters of the nominal response model are restricted in the following intervals $a_2 \in [-0.18, 4.15]$, $a_3 \in [0.17, 3.93]$, $c_2 \in [-8.27, 6.38]$ and $c_3 \in [-7.00, 8.24]$, whereas $a_1 = c_1 = 0$. The test length was $n = 50$ items and items were selected according to the information-maximizing item selection strategies (3.25) and (3.45) for CAT and RCAT, respectively.

The results are summarized in Table 3.1. We observe that the RMSE in RCAT is, typically, slightly smaller than that in CAT and slightly larger than the quantity that is suggested by our asymptotic analysis, $(\sqrt{nJ^*(\theta)})^{-1}$. The RMSE in RCAT seems to slightly outperform this benchmark in the case $\theta = -2$ when the number of revisions is large. For an examinee from this case, we plot in Figure 3.1 the evolution of the normalized total information $I_t(\theta)/f_t$, as well as the corresponding information from first responses, $\sum_{i=1}^{f_t} J(\theta; \mathbf{b}_i)/f_t$ and revisions, $I_t^R(\theta)/f_t$, where $1 \leq t \leq \tau_n$.

In Figure 3.2 we compare the approximate 95% confidence intervals, $\hat{\theta}_i \pm 1.96 \cdot (I_i(\hat{\theta}_i))^{-1/2}$ and $\hat{\theta}_{\tau_i} \pm 1.96 \cdot (I_{\tau_i}(\hat{\theta}_{\tau_i}))^{-1/2}$ that are obtained after i distinct items have been answered with a CAT and RCAT, respectively, where $1 \leq i \leq n$, when the ability parameter is $\theta = -3$. Our asymptotic results guarantee the validity of the final confidence interval ($i = n$) when n is large. The graph indicates that revision improves the estimation of θ.

Table 3.1: RMSE in CAT and RCAT in an idealized item pool

θ	$(\sqrt{nJ^*(\theta)})^{-1}$	CAT	RCAT		
			Expected Number of Revision		
			4	18	26
-3	.097	.104	.105	.107	.100
-2	.071	.075	.073	.070	.070
-1	.068	.072	.072	.072	.071
0	.068	.074	.072	.072	.072
1	.068	.077	.072	.069	.070
2	.068	.075	.072	.070	.070
3	.071	.079	.076	.073	.072

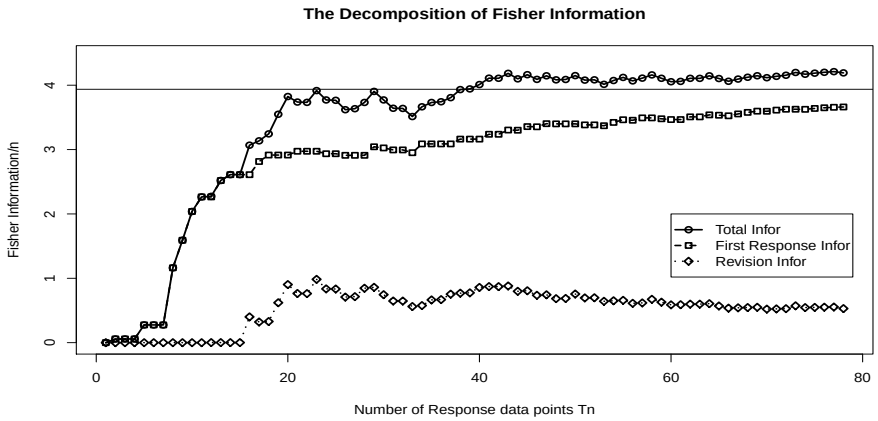

Figure 3.1: Decomposition of the Fisher information. The solid line represents the evolution of the normalized accumulated Fisher information, $\{I_t(\hat{\theta}_t)/f_t, 1 \leq t \leq \tau_n\}$, in a CAT with response revision. The dashed line with squares (diamonds) represents the corresponding information from first responses (revisions). The horizontal line represents the maximal Fisher information, $J^*(\theta)$. The true ability value is $\theta = -2$.

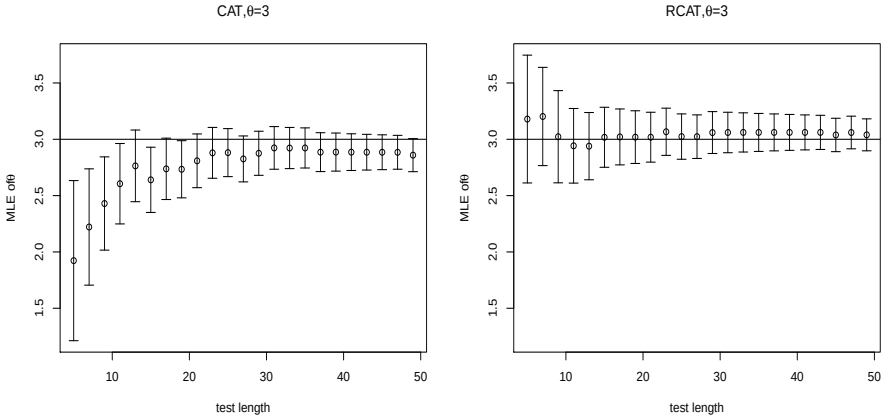

Figure 3.2: 95% Confidence Intervals. The left-hand side presents 95% confidence intervals, $\hat{\theta}_i \pm 1.96 \cdot (I_i(\hat{\theta}_i))^{-1/2}$, $1 \leq i \leq n$, in a standard CAT. The right-hand side presents the corresponding intervals $\hat{\theta}_{\tau_i} \pm 1.96 \cdot (I_{\tau_i}(\hat{\theta}_{\tau_i}))^{-1/2}$, $1 \leq i \leq n$ in the proposed RCAT design that allows for response revision. In both cases, the true value of θ is -3.

3.5.2 A discrete item pool

In the second study, we consider an item pool with 135 items that was constructed from a large scale standardized test in China. The item parameters were calibrated based on 10,000 examinees' responses. The MULTILOG (Thissen (1991)) was used to calibrate the item parameters of the nominal response model, which are described by Figure 3.3 of the supplementary material. Each item has $m = 4$ categories, which means that each item can be revised at most twice when revision is allowed. As before, for both designs we

select the item that has the maximum Fisher information, but now items are selected without replacement for each examinee, i.e. no item will be administered to the same examinee twice. We considered 3 levels for the test length, n, and 3 levels for the maximum number of items that can be revised, to which were referred as "small","medium" and "large". Specifically, the following cases were considered $n = 20$ ($n_1 = 5, 10, 20$), 30 ($n_1 = 5, 15, 30$) and 40 ($n_1 = 5, 20, 40$). The results are documented in Table 3.2 and show that the positive effect of revisions in the ability estimation is much more intense than in the case of an idealized item pool, especially when the number of revisions is large. However, as expected due to the discreteness of item pool, the RMSEs from both designs were much larger than $(\sqrt{nJ^*(\theta)})^{-1}$.

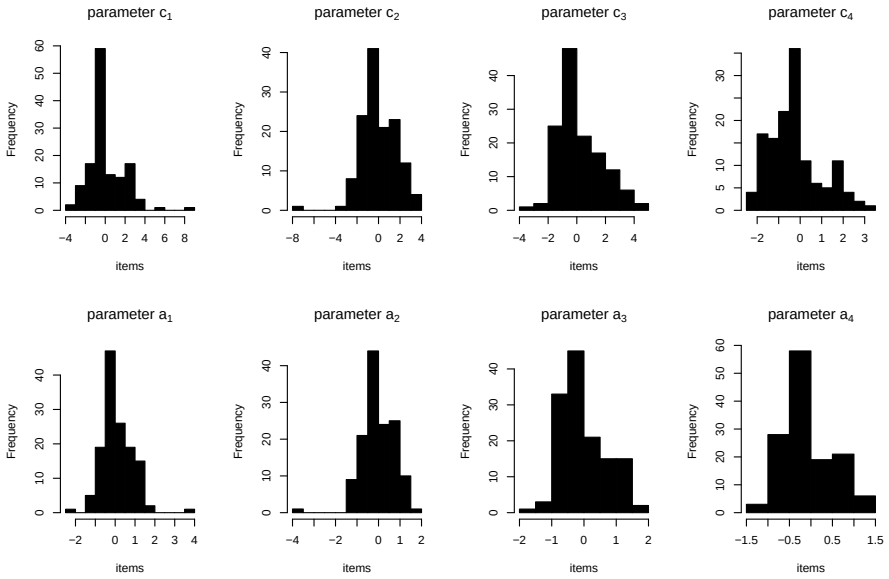

Figure 3.3: Calibrated item parameters of the nominal response model in a pool with 134 items, each having $m = 4$ categories.

Table 3.2: RMSE of CAT and RCAT in a realistic item pool

θ			-3	-2	-1	0	1	2	3
n=20	Design	Condition							
		$(\sqrt{nJ^*(\theta)})^{-1}$	.084	.059	.059	.080	.081	.107	.165
	CAT		.283	.230	.301	.346	.338	.300	.333
		small	.264	.211	.258	.342	.324	.276	.315
	RCAT	medium	267	.194	.278	.342	.328	.276	.292
		large	.248	.181	.259	.333	.320	.250	.289
n=30	Design	Condition							
		$(\sqrt{nJ^*(\theta)})^{-1}$	.068	.048	.048	.065	.066	.087	.135
	CAT		.256	.200	.236	.303	.291	.265	.314
		small	.246	.178	.227	.296	.287	.253	.300
	RCAT	medium	.232	.181	.221	.276	.278	.224	.274
		large	.222	.159	.205	.283	.275	.217	.267
n=40	Design	Condition							
		$(\sqrt{nJ^*(\theta)})^{-1}$	.059	.042	.042	.057	.058	.075	.117
	CAT		.243	.178	.213	.279	.289	.260	.309
		small	.247	.173	.208	.269	.277	.257	.300
	RCAT	medium	209	.149	.190	.260	.253	.215	.271
		large	.206	.145	.186	.257	.251	.202	.247

3.6 Two test-taking strategies in CAT

We now discuss two famous test-taking strategies that can take advantage of the feature of response revision in CAT. For both of them, examinees are assumed to know the adaptive design mechanism, which is the general principle that a correct response leads to a more difficult item, whereas a wrong response leads to an easier item.

The Wainer strategy was originally proposed by Wainer (1993), and later reformulated by Stocking (1997).Its main idea is that examinees can intentionally design an easier test by first answering all items incorrectly on purpose and then trying to correct their answers at the end of the test. Although Stocking (1997) described this strategy as an "unrealistic worst-case model", lots of simulation studies and real experiments have been conducted to evaluate modified CAT designs that allow for response revision when the examinees are simulated or trained to adopt the Wainer strategy (Bowles and Pommerich, 2001; Gershon and Bergstrom, 1995; Stocking, 1997; Vispoel et al., 1999; Wang and Wingersky, 1992).

The Kingsbury strategy (K) was proposed by Green et al. (1984) and was later extended by Wise et al.

(1999) to the so-called generalized Kingsbury Strategy (GK). The main idea of both strategies is that an examinee may choose to revise his/her answer to the previous item if he/she feels that the currently received item is much easier than the previous one. Like Wainer's strategy, these two strategies have also been evaluated extensively through a series of simulation studies and real testing experiments (Kingsbury, 1996; Vispoel et al., 2002). For a more detailed discussion we refer to Han (2013).

3.7 Simulation Studies Regarding Three test-taking behaviors

Three simulation studies were conducted to investigate the performance of the proposed CAT design, which from now on will be called RCAT. Recall that in RCAT, all responses to a certain item during the test contribute to the ability estimation, i.e., the first answer to this item as well as any subsequent revisions. A modification of the RCAT design in which only the most recent answer to each item is used for ability estimation is also studied. We refer to this design as RCAT-Naive. In the first study, we simulate a scenario in which the examinee revises only a number of careless errors. In the second and third study, we simulate a scenario in which the examinee adopts the Wainer strategy and the GK strategy, respectively. All studies were based on a realistic item pool presented in 3.5.2. For simplicity, none of the three studies incorporated item exposure control or non-statistical constraints in the item selection algorithm. Definitely, these factors will be included in a future study. The test length was fixed at 40 items. For each simulation condition, a sample of 5000 examinees was simulated at 13 θ nodes from -3 to 3 with step size 0.5. The evaluation criteria were bias and the root mean square error (RMSE).

3.7.1 Correcting careless errors.

The goal of this simulation study is to evaluate whether the RCAT design reduces the measurement error that is caused by careless mistakes, such as typing errors, by allowing the examinees to correct these errors. In order to do so, we selected a small number of items to which the response of the examinee was simulated to be incorrect. The incorrect answer was equally likely to be one of the three distracters of this item. Each of these items was not selected at random, but was chosen so that the probability of a correct answer to this item is high, according to the Nominal Response Model item response function. We refer to the errors to these particular items as "typing errors". The responses to all other items were simulated according to the Nominal Response Model item response function. Two types of CAT designs were simulated. The first is a standard CAT where response revisions are not allowed. The second is the RCAT design and in which examinees were simulated to revise only the above typing errors, but not the responses to any other items.

Table 3.3: The conditional bias from the three designs

Ability	Number of Typos								
	1			3			5		
	CAT	RCAT		CAT	RCAT		CAT	RCAT	
		I	II		I	II		I	II
-3	**-.10**	-.03	-.04	**-.25**	-.08	-.07	**-.35**	-.10	-.10
-2.5	**-.08**	-.05	-.05	**-.23**	-.08	-.09	**-.36**	-.12	-.13
-2	**-.08**	-.05	-.05	**-.18**	-.10	-.10	**-.27**	-.14	-.14
-1.5	**-.12**	-.09	-.09	**-.25**	-.19	-.20	**-.36**	-.26	-.26
-1	**-.12**	-.10	-.09	**-.31**	-.25	-.26	**-.46**	-.36	-.38
-0.5	**-.12**	-.09	-.09	**-.37**	-.30	-.32	**-.58**	-.45	-.51
0	**-.13**	-.10	-.11	**-.44**	-.37	-.39	**-.74**	-.55	-.67
0.5	**-.14**	-.11	-.10	**-.49**	-.41	-.44	**-.90**	-.65	-.82
1	**-.12**	-.08	-.08	**-.43**	-.33	-.35	**-.90**	-.58	-.79
1.5	**-.08**	-.04	-.04	**-.28**	-.17	-.20	**-.66**	-.35	-.52
2	**-.07**	-.02	-.03	**-.22**	-.10	-.11	**-.44**	-.22	-.27
2.5	**-.10**	-.04	-.04	**-.24**	-.12	-.13	**-.43**	-.20	-.25
3	**-.14**	-.10	-.10	**-.36**	-.23	-.24	**-.53**	-.32	-.36

Moreover, two cases were considered for the time of this revision. In the first one, examinees were simulated to correct each typing error as soon as possible during the test. This is denoted as RCAT- I. In the second one, examinees were simulated to correct all typing errors at the end of the test. This is denoted as RCAT-II. The number of errors each examinee made was set to 1, 3 and 5 respectively.

The bias for the three designs at each of the 13 θ nodes is documented in Table 3.3. The bias of the RCAT-I and RCAT-II designs is smaller than that of the regular CAT design at all nodes, even when there is only 1 typing error in the 40 items. As expected, the reduction in the bias becomes more substantial as the number of typing errors increases. This suggests that when the examinees correct their careless mistakes, the RCAT design reduces measurement error and increases test validity, even though the original typing errors are not completely erased. The RCAT-I design had slightly smaller bias than the RCAT-II design at most of the nodes. This indicates that examinees can benefit more by correcting their errors as soon as possible after they realize them than correcting them at the end of the test. Moreover, it provides an argument against the hypothesis that revision should only take place at the end and not during the test. The corresponding RMSEs are documented in Figure 3.4. The RCAT-I design has the smallest RMSEs at each of the nodes, and is followed by the RCAT-II design. The regular CAT design where response revisions are not allowed always has the largest RMSE. This again supports the hypothesis that allowing for response revision according to the RCAT design can reduce measurement error.

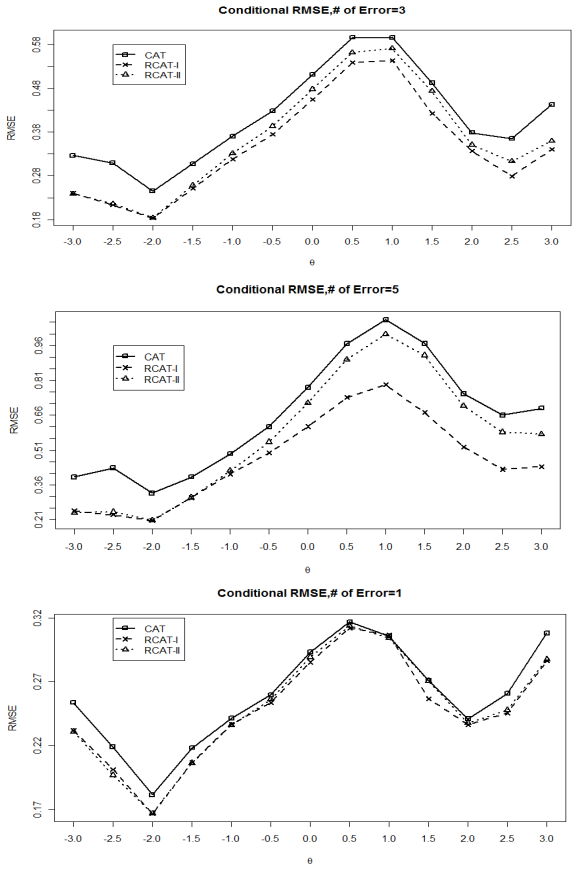

Figure 3.4: The conditional RMSEs at different scenarios for number of errors

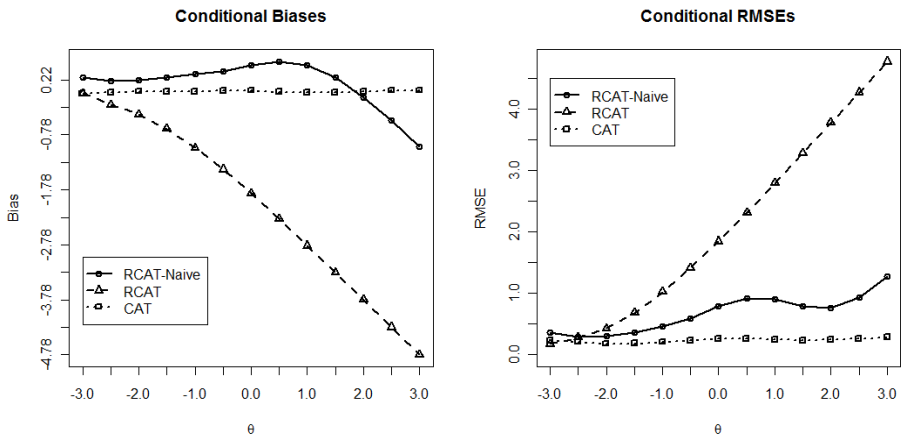

Figure 3.5: . The conditional biases and RMSEs under the Wainer strategy

3.7.2 The Wainer Strategy

The goal of this simulation study is to compare the performance of the proposed RCAT design and the RCAT-Naive design when the examinee adopts the Wainer strategy. To be comparable with earlier researches, this study mimics the "unrealistic worst case" scenario used in Stocking (1997). Therefore, the examinee was simulated to randomly select one of the three distracters in each of the 40 items and then to go back and re-do each question based on the conditional probability in equation (3.49). The RCAT-Naive design estimates each examinee's ability using only the 40 revised responses, whereas the RCAT design includes in the ability estimation the first 40 wrong answers, in addition to the 40 revisions. Note here that the examinee receives exactly the same items in both designs. Finally, for comparison purposes, a standard CAT design that does not allow for response revision, and in which the examinee cannot follow the Wainer strategy, is simulated as the baseline.

The conditional bias and the conditional root mean square error (RMSE) of the three designs are summarized in Figure 3.5. We can see that in the RCAT-Naive design where the first attempt to each item is ignored, examinees with ability levels from -3 to 1 can achieve a positive bias. On the other hand, in the proposed RCAT design where all incorrect responses of the first round are taken into account, all examinees receive much lower scores compared to their true abilities. Moreover, this negative bias increases with the ability level. In fact, even if the examinee was able to correct all the wrong answers in the second round, his/her final score at the end of the test would still be much lower than the true ability level. Thus, the proposed RCAT design penalizes the Wainer strategy, unlike the RCAT-Naive design that is quite vulnerable to it.

3.7.3 The GK Strategy

The purpose of this study is to investigate the performance of the proposed RCAT design, as well as to compare it with the RCAT-Naive design and a regular CAT design, when the examinee adopts the GK strategy. The idea of the GK strategy is that if the examinee realizes that the current item is less difficult than the previous one, then he/she might go back to change his/her answer to the previous item. In order to simulate the GK strategy, we assume that there is a parameter that describes the difficulty of each item. Then, following Wise Wise et al. (1999) we simulate the GK strategy as follows: if the examinee receives a new item whose difficulty parameter is lower than that of the previous administered item by a certain amount, say x , then he/she goes back to change his/her answer to the previous item with some probability y . Following Wise et al. (1999), we set $x = 0.5$ and $y = 0.73$. For the revised responses, we consider two models. In the first one, revisions are simulated according to the conditional probability (3.49). In the second one, the revised answer is selected randomly among the remaining three options. Thus, two versions for RCAT and RCAT-Naive Design are considered in this study and are summarized in Table 3.4 below:

Table 3.4: Four types of CAT designs that allow for response revision

Design	Model for revised answer	Name
RCAT	Conditional Probability (7)	RCAT-C
	Random Guess	RCAT-R
RCAT-Naive	Conditional Probability (7)	RCAT-NC
	Random Guess	RCAT-NR

Unlike the logistic dichotomous models, the polytomous Nominal Response Model, on which all the above designs are based, does not have a natural difficulty parameter. For this reason, the difficulty index for each item was developed according to the nonparametric transformation based on the correct proportion and the point-biseral correlation (Richardson, 1936; Tucker, 1946). Note that this difficulty index is used only in order to simulate the examinee's revision strategy, and it is not used in the ability estimation algorithm.

The bias of the five designs is documented in Table 3.5. For the two RCAT-Naive designs, some low ability examinees were underestimated, while some medium to high ability examinees were overestimated. This means that if the first attempts are not included in the ability estimation, an examinee with a relatively high ability may benefit by adopting the GK strategy. On the other hand, the bias of the RCAT designs is almost 0 at all nodes, similarly to the standard CAT design. This indicates that the proposed RCAT design, unlike the RCAT-Naive modification, is robust against the GK strategy. The RMSEs from all designs

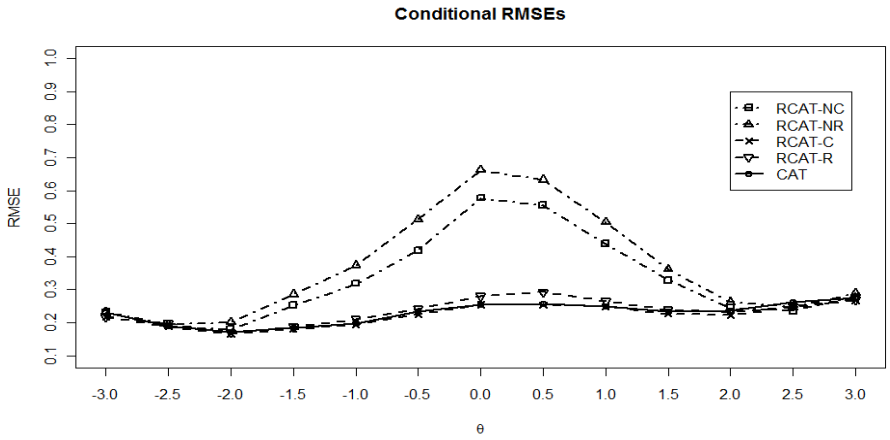

Figure 3.6: The conditional RMSEs from six designs under the GK strategy

are presented in Figure 3.6. Again, the two RCAT-Naive designs have larger RMSEs for examinees whose abilities fall into the range of -1.5 to 1.5. The RMSEs of the two RCAT designs are quite similar to those of the standard CAT design.

Table 3.5: The bias from five designs under the GK strategy

Design	θ												
	-3	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	2.5	3
RCAT-NC	.06	-.02	**-.08**	**-.16**	**-.11**	**-.04**	**.16**	**.33**	**.31**	**.21**	**.10**	.03	-.03
RCAT-NR	.09	-.02	**-.09**	**-.17**	**-.14**	**.00**	**.23**	**.39**	**.37**	**.27**	**.14**	.01	-.11
RCAT-C	-.02	-.01	.00	.01	.01	.02	.02	.00	-.01	-.01	-.01	.02	.02
RCAT-R	.03	-.03	-.01	-.02	-.01	.00	.03	.04	.05	.04	.03	.01	-.05
CAT	-.02	-.01	.00	.01	.01	.02	.02	.00	-.01	-.01	.01	.02	.02

3.8 Conclusion and Discussion

In Computerized Adaptive Testing (CAT), the administered items are taylored to the examinee's ability, which is learned on-line as the test progresses. However, unlike in a traditional paper-pencil test, examinees are not allowed to revise their answers. Response revision is widely believed to harm statistical efficiency and, for this reason, modified CAT designs that have been proposed in order to incorporate this feature postulate quite restrictive revision rules.

In this work, we have proposed a novel CAT design whose goal is to mitigate the clash between adaptivity

and flexibility, preserving the statistical efficiency of the traditional CAT, while allowing the examinees to revise their previous answers at any time during the test. In order to do so, the proposed design relies on a polytomous IRT model, unlike the majority of CAT designs that are based on dichotomous models. Thus, we are able to postulate a joint probability model for all responses on an item, the first answer and any subsequent revisions, which are all used for the update of the ability parameter. On the other hand, we do not model the decision of the examinee to revise or not and we propose exactly the same item selection method as in the corresponding conventional CAT: selecting the item with the maximum Fisher information at the current estimate of the ability parameter.

We performed a rigorous asymptotic analysis of the proposed method, which to our knowledge is the first in the literature of CAT designs that allow for response revision. We showed that the resulting estimator is strongly consistent and, under a stability assumption on the cumulative Fisher information, asymptotically normal. This assumption is satisfied when the number of revisions is much smaller than the number of distinct items, in which case the asymptotic variance of the resulting estimator is the same as the one that is obtained by the corresponding conventional CAT. However, as our simulation study corroborated, a large number of revisions can lead to more efficient estimation.

The only restriction that we impose on the examinee is that any given item with $m > 2$ categories can be revised at most $m - 2$ times during the test. However, the examinee does not recover the freedom of a paper-pencil test, as all responses on an item, and not only the last one, contribute to the estimation of the ability parameter. This is a feature that helps protect the resulting ability estimator against cheating strategies, which is an issue that we explore in our current research.

The most desirable feature of our approach from a practical point of view is that it *does not require any additional calibration effort* to the one that is needed by the corresponding conventional CAT that is based on the nominal response model. Thus, a traditional CAT system that is based on the nominal response model can be very easily modified to allow response revision.

The results of three simulation studies regarding three test-taking strategies provide important insight to our proposed design. First of all, it was observed that examinees who only correct careless errors during the test have the chance to improve their scores. Moreover, in this scenario, the estimation bias is smaller and the estimation accuracy higher than that in a standard CAT design where the examinees do not have the opportunity to correct such mistakes. Moreover, it was observed that it is more beneficial for the examinee, as well as for the accuracy of the test, to correct these errors immediately than at the end of the test. This reveals an additional benefit of allowing the examinees can revise at any time during the test, and not only at the end.

Second, it was observed that examinees who adopt the Wainer strategy have a very high risk of getting a much lower score than the one that corresponds to their true ability levels. This is the case because in the proposed design all responses during the test contribute to the ability estimation. On the other hand, if the design is modified so that only the last answer to each item contributes to the ability estimation, then it was show that examinees of low and medium ability can artificially inflate their scores by adopting the Wainer strategy.

Third, it was shown from simulation studies that examinees do not seem to gain anything by adopting the generalized Kingsbury strategy, as their ability estimates are very similar to those that they would obtain in a standard CAT where they could not change their answers. On the other hand, if the design is modified so that only the last answer to each item contributes to the ability estimation, then examinees of high ability can improve their scores by adopting the generalized Kingsbury strategy.

Our work opens a number of research directions. First of all, since items in reality are drawn without replacement, this may call for modifications of the item selection strategy in the spirit of Chang and Ying (1999). Moreover, more empirical and theoretical work is required in order to understand the effect of different revision behaviors to the final ability estimation, which is another issue under investigation. In fact, it would be beneficial to model and incorporate in the ability estimation and the item selection the decisions of the examinee to revise or not at each step. However, obtaining reliable, universal models for the behavior of the examinee is a challenge that can be better addressed as soon as the proposed design is implemented in practice and relevant data can be obtained. Another interesting direction is to take the point of the test-taker and understand if there is an optimal revision strategy that could be employed. Finally, it remains an open problem to incorporate response revision in a CAT that is based on binary items, where a dichotomous IRT model must be used and our approach is not applicable.

References

Ayers, E., Nugent, R., and Dean, N. (2008). Skill set profile clustering based on student capability vectors computed from online tutoring data. *Educational Data Mining 2008*, page 210.

Barnes, T. (2010). Novel derivation and application of skill matrices: The q-matrix method. *Handbook on educational data mining*, pages 159–172.

Bartroff, J., Finkelman, M., and Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika*, 73(3):473–486.

Berk, R. H. et al. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58.

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459.

Bowles, R. and Pommerich, M. (2001). An examination of item review on a cat using the specific information item selection algorithm.

Chang, H.-H. and Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3):211–222.

Chang, H.-H. and Ying, Z. (2007). Computerized adaptive testing. *The Sage Encyclopedia of Measurement and Statistics*, 1:170–173.

Chang, H.-H. and Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, 37(3):1466–1488.

Chen, J., Torre, J., and Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, 50(2):123–140.

Chiu, C.-Y. and Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, 30(2):225–250.

Chiu, C.-Y., Douglas, J. A., and Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74(4):633–665.

Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, 2005(2):i–35.

Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61(2):287–307.

De La Torre, J. (2011). The generalized dina model framework. *Psychometrika*, 76(2):179–199.

De La Torre, J., van der Ark, L. A., and Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*.

Gershon, R. and Bergstrom, B. (1995). Does cheating on cat pay: Not! In *American Educational Research Association Annual Meeting*, volume 1995.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., and Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21(4):347–360.

Han, K. T. (2013). Item pocket method to allow response review and change in computerized adaptive testing. *Applied Psychological Measurement*, 37(4):259–275.

Hartz, S. (2002). A bayesian framework for the unified model for assessing cognitive abilities: lending theory with practicality. *Unpublished doctoral dissertation*.

Hartz, S. and Roussos, L. (2008). The fusion model for skills diagnosis: Blending theory with practicality. *ETS Research Report Series*, 2008(2):i–57.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2):191–210.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233.

Jaeger, J., Tatsuoka, C., Berns, S., Varadi, F., Czobor, P., and Uzelac, S. (2006). Associating functional recovery with neurocognitive profiles identified using partially ordered classification models. *Schizophrenia research*, 85(1):40–48.

Junker, B. W. and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3):258–272.

Kingsbury, G. (1996). Item review and adaptive testing. In *Annual Meeting of the National Council on Measurement in Education, New York, NY*.

Lai, T. L. and Robbins, H. (1979). Adaptive design and stochastic approximation. *The Annals of Statistics*, 7(6):1196–1221.

Law, P. (2002). No child left behind act of 2001. *Public Law*, 107:110.

Liu, J., Ying, Z., and Zhang, S. (2015). A rate function approach to computerized adaptive testing for cognitive diagnosis. *Psychometrika*, (2):468–490.

Lord, F. M. (1971). Robbins-monro procedures for tailored testing. *Educational and Psychological Measurement*, 31(1):3–31.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Luecht, R. M. and Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, 35(3):229–249.

Macready, G. B. and Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational and Behavioral Statistics*, 2(2):99–120.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64(2):187–212.

Park, Y. and Lee, Y. (2011). Diagnostic cluster analysis of mathematics skills. *IERI monograph series: Issues and methodologies in large-scale assessments*, 4:75–107.

Passos, V. L., Berger, M. P., and Tan, F. E. (2007). Test design optimization in cat early stage with the nominal response model. *Applied Psychological Measurement*, 31(3):213–232.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* ERIC.

Richardson, M. W. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika*, 1(2):33–49.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Robitzsch, A., Kiefer, T., George, A. C., Uenlue, A., and Robitzsch, M. A. (2014). Package cdm.

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

Roussos, L. A., Templin, J. L., and Henson, R. A. (2007). Skills diagnosis using irt-based latent class models. *Journal of Educational Measurement*, 44(4):293–311.

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

Rupp, A. A. and Templin, J. (2008a). The effects of q-matrix misspecification on parameter estimates and classification accuracy in the dina model. *Educational and Psychological Measurement*, 68(1):78–96.

Rupp, A. A. and Templin, J. L. (2007). Unique characteristics of cognitive diagnosis models. In *The Annual Meeting of the National Council for Measurement in Education, Chicago.*

Rupp, A. A. and Templin, J. L. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, 6(4):219–262.

Schmidt, F. L., Urry, V. W., and Gugel, J. F. (1978). Computer assisted tailored testing: Examinee reactions and evaluations. *Educational and Psychological Measurement*, 38(2):265–273.

Sie, H., Finkelman, M. D., Bartroff, J., and Thompson, N. A. (2015). Stochastic curtailment in adaptive mastery testing improving the efficiency of confidence interval–based stopping rules. *Applied Psychological Measurement*, 39(4):278–292.

Stocking, M. L. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. *Applied Psychological Measurement*, 21(2):129–142.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(3):337–350.

Tatsuoka, C. and Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):143–157.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4):345–354.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational and Behavioral Statistics*, 10(1):55–73.

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

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory.* Scientific Software International.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, 11(1):1–13.

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of educational measurement*, 35(4):328–345.

Vispoel, W. P., Clough, S. J., Bleiler, T., Hendrickson, A. B., and Ihrig, D. (2002). Can examinees use judgments of item difficulty to improve proficiency estimates on computerized adaptive vocabulary tests? *Journal of Educational Measurement*, 39(4):311–330.

Vispoel, W. P. and Coffman, D. D. (1992). Computerized adaptive testing of music-related skills. *Bulletin of the Council for Research in Music Education*, (112):29–49.

Vispoel, W. P., Hendrickson, A. B., and Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement*, 37(1):21–38.

Vispoel, W. P., Rocklin, T. R., Wang, T., and Bleiler, T. (1999). Can examinees use a review option to obtain positively biased ability estimates on a computerized adaptive test? *Journal of Educational Measurement*, 36(2):141–157.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, 12(1):15–20.

Wainer, H. and Wright, B. D. (1980). Robust estimation of ability in the rasch model. *Psychometrika*, 45(3):373–391.

Wang, M. and Wingersky, M. (1992). Incorporating post-administration item response revision into cat. In *annual meeting of the American Educational Research Association, San Francisco, CA*.

Wang, S. and Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika*, 80(1):85–100.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25.

Willse, J., Hensen, R., and Templin, J. (2007). Using sum scores or irt in place of cognitive diagnosis models: can existing or more familiar models do the job? In *The Annual Meeting of the National Council for Measurement in Education, Chicago*.

Wise, S. L. (1996). A critical analysis of the arguments for and against item review in computerized adaptive testing. In *Annual Meeting of the National Council on Measurement in Education (NCME)*, volume 1996.

Wise, S. L., Finney, S. J., Enders, C. K., Freeman, S. A., and Severance, D. D. (1999). Examinee judgments of changes in item difficulty: Implications for item review in computerized adaptive testing. *Applied Measurement in Education*, 12(2):185–198.

Wu, C. J. (1985). Efficient sequential designs with binary data. *Journal of the American Statistical Association*, 80(392):974–984.

Wu, C. J. (1986). Maximum likelihood recursion and stochastic approximation in sequential designs. *Lecture Notes-Monograph Series*, 8:298–313.

Ying, Z. and Wu, C. J. (1997). An asymptotic theory of sequential designs based on maximum likelihood recursions. *Statistica Sinica*, 7(1):75–91.