NEW ITEM SELECTION AND TEST ADMINISTRATION PROCEDURES FOR COGNITIVE DIAGNOSIS COMPUTERIZED ADAPTIVE TESTING

BY MEHMET KAPLAN

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Education

Written under the direction of

Jimmy de la Torre

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2016

ABSTRACT OF THE DISSERTATION

New Item Selection and Test Administration Procedures for Cognitive Diagnosis Computerized Adaptive Testing

by Mehmet Kaplan

Dissertation Director: Jimmy de la Torre

The significance of formative assessments has recently been underscored in the educational measurement literature. Formative assessments can provide more diagnostic information to improve teaching and learning strategies compared to summative assessments. Cognitive diagnosis models (CDMs) are psychometric models that have been developed to provide a more detailed evaluation of assessment data. CDMs aim to detect students' mastery and nonmastery of attributes in a particular content area. Another major research area in psychometrics is computerized adaptive testing (CAT). It has been developed as an alternative to paper-and-pencil tests, and widely used to deliver tests adaptively.

Although the traditional CAT seems to satisfy the needs of the current testing market by providing summative scores, the use of CDMs in CAT can produce more diagnostic information with an efficient testing design. With a general aim to address needs in formative assessments, this dissertation aims to achieve three objectives:

(1) to introduce two new item selection indices for cognitive diagnosis computerized adaptive testing (CD-CAT); (2) to control item exposure rates in CD-CAT; and (3) to propose an alternative CD-CAT administration procedure. Specifically, two new item selection indices are introduced for cognitive diagnosis. In addition, high item exposure rates that typically accompany efficient indices are controlled using two exposure control methods. Finally, a new CD-CAT procedure that involves item blocks is introduced. Using the new procedure, examinees would be able to review their responses within a block of items. The impact of different factors, namely, item quality, generating model, test termination rule, attribute distribution, sample size, and item pool size, on the estimation accuracy and exposure rates was investigated using three simulation studies. Moreover, item type usage in conjunction with the examinees' attribute vectors and generating models was also explored. The results showed that the new indices outperformed one of the most popular indices in CD-CAT, and also, they performed efficiently with the exposure control methods in terms of classification accuracy and item exposure. In addition, a new blocked-design CD-CAT procedure was promising for allowing item review and answer change during the test administration with a small loss in the classification accuracy.

Acknowledgements

I would like to express my deepest gratitude to my advisor, my mentor, and my editor, Dr. Jimmy de la Torre, for his excellent and continuous support, and also for his great patience, motivation, and immense knowledge. I feel amazingly fortunate to have such a remarkable advisor because there are only few people who can do all of these. I could not have imagined having a better advisor and mentor for my graduate study, and I hope that one day I would become an advisor as good as him. Jimmy, I will never forget the taste of the mangos you brought to our meetings.

Dr. Barrada's insightful comments and constructive criticisms helped me understand many concepts related to my dissertation's topic more deeply. I am very gratified to have him in my committee even though he lives overseas. I am very grateful to have Dr. Chia-Yi Chiu and Dr. Youngsuk Suh in my dissertation committee for their insightful comments and encouragement.

I also would like to thank the Ministry of National Education of Turkey for the grant that brought me to the U.S., and the former and current staff at the office of the Turkish Educational Attaché in New York for their support despite their immense workload. My labmates also deserve special thanks for providing excellent and peaceful working atmosphere.

Most importantly, I couldn't have come this far without my family. Doing academic research and being abroad demand a lot of love, patience, sacrifice, and understanding. I would like to thank my mom and sister for their support in all aspects. Dad, you will not be forgotten.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction and Objectives

1.1 Introduction

Interest in formative assessment has rapidly grown in the psychological and educational measurement over the past decades. It includes a range of different assessment procedures that provide more detailed feedback to improve teaching and learning rather than just giving a single score. The use of formative assessment has several advantages over summative assessment. For example, it enhances teaching and learning strategies by providing better feedback to teachers and students (DiBello & Stout, 2007). Based on the feedback that identifies individual strengths and weaknesses in a particular content, teachers can design classroom activities to optimize student learning. Huebner (2010) also stated that such assessment fulfills the demands of recent political decisions in education such as the No Child Left Behind Act (2001).

Largely to harness the benefits of the formative assessment, several cognitive diagnosis models (CDMs) have been introduced and developed in educational measurement. CDMs are latent class models that can be used to detect mastery and nonmastery of multiple fine-grained skills or attributes in a particular content domain (de la Torre, 2009). These attributes are generally binary; however, they can also have polytomous levels of mastery. Examples of binary attributes defined in the mixed fraction subtraction domains are (1) converting a whole number to a fraction, (2) separating a whole number from a fraction, (3) simplifying before subtracting (Tatsuoka, 1990). Examples of attributes with binary and polytomous levels of mastery defined in the

proportional reasoning domain are (1) prerequisite skills; (2a) comparing and (2b) ordering fractions; and (3a) constructing ratios and (3b) proportions (Tjoe & de la Torre, 2014). By identifying the presence or absence of the attributes for particular domains, CDMs can provide more diagnostic and informative feedback.

To date, a variety of models has been developed to increase the applicability of CDMs. The deterministic inputs, noisy "and" gate (DINA; de la Torre, 2009; Haertel, 1989; Junker & Sijtsma, 2001) model, the deterministic input, noisy "or" gate (DINO; Templin & Henson, 2006) model, the noisy input, deterministic "and" gate (NIDA; Maris, 1999; Junker & Sijtsma, 2001) model, the noisy input, deterministic "or" gate (NIDO; Templin & Henson, 2006) model, and fusion (Hartz, 2002; Hartz, Roussos, & Stout, 2002) model are examples of constrained CDMs. Constrained CDMs require specific assumptions about the relationship between attribute vector and task performance (Junker & Sijtsma, 2001). Nonetheless, they provide results that can easily be interpreted. In addition to constrained models, more generalized CDMs have also been proposed: the log-linear CDM (Henson, Templin, & Willse, 2009), the general diagnostic model (von Davier, 2008), and the generalized DINA model (G-DINA; de la Torre, 2011). The general models relax some of the strong assumptions in the constrained models, and provide more flexible parameterizations. However, general models are more difficult to interpret compared to constrained models because they involve more complex parametrizations. Therefore, the choice of using either a constrained or a general model depends on the particular application.

Computerized adaptive testing (CAT) has also become a popular tool in educational testing since the use of personal computers became accessible (van der Linden & Glas, 2002). It has been developed as an alternative to paper-and-pencil tests because of the following advantages: CAT offers more flexible testing schedules for individuals; the scoring procedure is faster with CAT; it makes wider range of items

with broader test contents available (Educational Testing Service, 1994); CAT provides shorter test-lengths; it enhances measurement precision; and offers tests on demand (Meijer & Nering, 1999). A pioneering application of CAT was applied by the US Department of Defense to carry out the Armed Services Vocational Aptitude Battery in the mid 1980s. However, the transition from paper-and-pencil testing to CAT truly began when the National Council of State Boards of Nursing used a CAT version of its licensing exam, and it was followed by the Graduate Record Examination (van der Linden & Glas, 2002). At present, many testing companies offer tests using within an adaptive environment (van der Linden & Glas, 2010).

A CAT procedure typically consists of three steps: "how to START", "how to CONTINUE", and "how to STOP" (Thissen & Mislevy, 2000, p. 101). First, the specification of the initial items determines the ability estimation at the early stage of the test. Second, the ability estimate is updated by giving items appropriate to the examinee's ability level. Last, the test is terminated after reaching a predetermined precision or number of items. In CAT, each examinee receives items appropriate to his/her ability level from an item bank, and the ability level is estimated during or at end of the test administration. Therefore, different tests, including different items with different lengths, can be created for different examinees.

CAT procedures are generally built upon item response theory (IRT) models, which provide summative scores based on the performance of the examinees. However, different psychometric models (i.e., CDMs) can also be used in the CAT procedures. Considering the advantages of CAT, the use of CDMs in CAT can provide better diagnostic feedback with more accurate estimates of examinees' attribute vectors. At present, most of the research in CAT has been done in the context of IRT; however, a small number of research has recently been conducted in cognitive diagnosis CAT (CD-CAT). One of the reasons behind the limited research on CD-CAT is that some of the concepts in traditional CAT (i.e., Fisher information) are not applicable in

CD-CAT because of the discrete nature of attributes.

1.2 Objectives

IRT and CAT are two well-studied research areas in psychometrics. Both have received considerable attention from a number of researchers in the field (van der Linden & Glas, 2002; Wainer et al., 1980). Although CAT in the context of IRT seems to satisfy the needs of the current testing market, it may not be sufficient in providing informative results to teachers and students to improve teaching and learning strategies. In this regard, cognitive diagnosis modeling can be used with CAT to obtain more detailed information about examinees' strengths and weaknesses with more efficient testing design. Despite its potential advantages in terms of efficiency and more diagnostic evaluations, research on CD-CAT is rather scarce. The following are examples of works in this area: Cheng (2009), Hsu, Wang, and Chen (2013), McGlohen and Chang (2008), Wang (2013), and Xu, Chang, and Douglas (2003).

Other developments in CD-CAT pertain to the test termination rules. Hsu et al. (2013) proposed two test termination rules based on the minimum of the maximum of the posterior distribution of attribute vectors in CD-CAT. They also developed a procedure based on the Sympson-Hetter method (1985) to control item exposure rates. Their procedure was capable of controlling test overlap rates using variable test-lengths. Recently, Wang (2013) proposed the mutual information item selection method in CD-CAT, and she compared the different methods (i.e., the Kullback-Leibler [K-L] information, Shannon entropy, and the posterior-weighted K-L index [PWKL]) using short test lengths. Based on this study, the PWKL was shown to have better efficiency. Additionally, the PWKL is easier to implement, thus making it a popular item selection method in CD-CAT. Despite its advantages, two shortcomings of the PWKL can be noted: the test lengths obtained from the PWKL were rather long and it produced high exposure rates. Therefore, it remains to be seen whether

other methods can be used in place of the PWKL.

This dissertation has three primary objectives: (1) to introduce two new item selection indices for CD-CAT, (2) to investigate item exposure rate control in CD-CAT, and (3) to propose a new CAT administration procedure. Of the two new item selection indices that were introduced for CD-CAT, one was based on the G-DINA model discrimination index, whereas the other one was based on the PWKL. The efficiency of the new indices was compared to the PWKL in the context of the G-DINA model. The impact of item quality, generating model, and test termination rule on the efficiency was investigated using a simulation study. In addition, high item exposure rates resulting from the different indices were controlled using the restrictive progressive and restrictive threshold methods (Wang, Chang, & Huebner, 2011). In addition to the factors, namely, item quality, generating model, and test termination rule, the impact of attribute distribution, item pool size, sample size, and prespecified desired exposure rate on the exposure rates was examined. Finally, a different CD-CAT procedure was introduced. Using the new procedure, examinees would be able to review their responses within a block of items. A successful attainment of these objectives would lead to a better understanding of CD-CAT, which in turn would increase the applicability of the procedure.

Along with these objectives, a more efficient simulation design was proposed in this dissertation. Using a small, but specific subset of the attribute vectors, and applying appropriate weights to these vectors, the new design can be used to examine how different attribute vector distributions can impact the results. With the proposed design, item type usage, in conjunction with the examinees' attribute vectors and generating models, was explored.

References

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619-632.

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

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

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

Educational Testing Service (1994). *Computer-based tests: Can they be fair to everyone?* Princeton, NJ: Educational Testing Service.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.

Hartz, S. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practice*. Unpublished doctoral thesis, University of Illinois at Urbana-Champain.

Hartz, S., Roussos, L., & Stout, W. (2002). *Skills diagnosis: Theory and practice* [User manual for Arpeggio software]. Princeton, NJ: Educational Testing Service.

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

Hsu, C.-L., Wang, W.-C., & Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement, 37*, 563-582.

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research, and Evaluation, 15*, 1-7.

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

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

McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40*, 808-821.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*, 187-194.

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

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Centre.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (p. 453-488). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer et al. (Eds.). *Computerized adaptive testing: A primer* (pp. 101-133). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

van der Linden, W. J., & Glas, C. A. W. (2002). Preface. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. Vii-Xii). Boston, MA: Kluwer.

van der Linden, W. J., & Glas, C. A. W. (2010). Preface. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. V-Viii). Boston, MA: Kluwer.

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

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1980). *Computerized adaptive testing: A Primer.* Hillsdale, NJ: Erlbaum.

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement, 73,* 1017-1035.

Wang, C., Chang, H.-H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement, 48*, 255-273.

Xu, X., Chang, H.-H., & Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Chapter 2

Study I: New Item Selection Methods for CD-CAT

Abstract

This article introduces two new item selection methods, the modified posterior-weighted Kullback-Leibler index (MPWKL) and the generalized deterministic inputs, noisy "and" gate (G-DINA) model discrimination index (GDI), that can be used in cognitive diagnosis computerized adaptive testing. The efficiency of the new methods is compared with the posterior-weighted Kullback-Leibler (PWKL) item selection index using a simulation study in the context of the G-DINA model. The impact of item quality, generating models, and test termination rules on attribute classification accuracy or test length is also investigated. The results of the study show that the MPWKL and GDI perform very similarly, and have higher correct attribute classification rates or shorter mean test lengths compared with the PWKL. In addition, the GDI has the shortest implementation time among the three indices. The proportion of item usage with respect to the required attributes across the different conditions is also tracked and discussed.

Keywords: cognitive diagnosis model, computerized adaptive testing, item selection method

---

This chapter has been published and can be referenced as: Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*, 167-188.

2.1 Introduction

Recent developments in psychometrics put an increasing emphasis on formative assessments that can provide more information to improve learning and teaching strategies. In this regard, cognitive diagnosis models (CDMs) have been developed to detect mastery and nonmastery of attributes or skills in a particular content area. In contrast to the unidimensional item response models (IRTs), CDMs provide a more detailed evaluation of the strengths and weaknesses of students (de la Torre, 2009). Computerized adaptive testing (CAT) has been developed as an alternative to paper-and-pencil test, and provides better ability estimation with a shorter and tailored test for each examinee (Meijer & Nering, 1999; van der Linden & Glas, 2002). Most of the research in CAT has been conducted in the traditional IRT context. However, a small number of research has recently been done in the context of cognitive diagnosis computerized adaptive testing (CD-CAT; Cheng, 2009; Hsu, Wang, & Chen, 2013; McGlohen & Chang, 2008; Wang, 2013; Xu, Chang, & Douglas, 2003).

One of the main components of CAT is the item selection method. By choosing more appropriate methods, better estimates of the examinees' abilities or attribute vectors can be expected. Because of the discrete nature of attributes, some of the concepts in traditional CAT such as Fisher information are not applicable in CD-CAT. The goal of this study is to introduce two new indices, the modified posterior-weighted Kullback-Leibler index (MPWKL) and the generalized deterministic inputs, noisy "and" gate (G-DINA) model discrimination index (GDI), as item selection methods in CD-CAT, and evaluate their efficiency under the G-DINA framework. Their efficiency is compared with the posterior-weighted Kullback-Leibler index (PWKL; Cheng, 2009). The effects of different factors are also investigated: The item quality is manipulated; reduced versions of the G-DINA model are used for generating item response data; and fixed-test lengths and minimum of the maximum (minimax) of

the posterior distribution of attribute vectors (Hsu et al., 2013) are used as stopping rules in the test administration. With respect to the stopping rules, the former provides a comparison of the efficiency of the three indices under different fixed-test lengths, whereas the latter provides tailored tests with different test lengths for each examinee.

The remaining sections of the article are laid out as follows: The next section gives a background in the G-DINA model and its reduced versions. In addition, the item selection indices are discussed, and the use of the GDI as an item selection method is illustrated. In the "Simulation Study" section, the design and the results of the simulation study are presented, and the efficiency of the indices under different conditions is compared. Finally, "Discussion and Conclusion" section presents with a discussion of the findings of this work and directions for future research.

2.1.1 Cognitive Diagnosis Models

CDMs aim to determine whether examinees have or have not mastered a set of specific attributes. The presence or absence of the attributes is represented by a binary vector. Let $\boldsymbol{\alpha}_i$=$\{\alpha_{ik}\}$ be the examinee's binary attribute vector for $k = 1, 2 \dots K$ attributes. The kth element of the vector is 1 when the examinee has mastered the kth attribute, and it is 0 when the examinee has not mastered it. Similarly, let $\boldsymbol{X}_i = \{x_{ij}\}$ be the binary response vector of examinee i for a set of J items in which $i = 1, 2 \dots N$, and $j = 1, 2 \dots J$. In CDM, the required attributes for each item are represented in a Q-matrix (Tatsuoka, 1983), which is a $J \times K$ matrix. The element of the jth row and the kth column, q_{jk}, is 1 if the kth attribute is required to answer the jth item correctly, and 0 otherwise.

A general CDM called *generalized deterministic inputs, noisy "and" gate* (G-DINA) model was proposed by de la Torre (2011). It is a generalization of the *deterministic inputs, noisy "and" gate* (DINA; de la Torre, 2009; Haertel, 1989; Junker

& Sijtsma, 2001) model, and it relaxes some of the strict assumptions of the DINA model. Instead of two, the G-DINA model partitions examinees into $2^{K_j^*}$ groups, where K_j^* is the number of required attributes for item j. The mathematical representation of the model consists of the combination of the baseline probability, the main effects due to the attribute k, the interaction effects due to the attributes k and k' ($k \neq k'$), and other higher-order interaction effects (for more details, see de la Torre, 2011).

A few of commonly encountered CDMs are constrained versions of, and therefore, are subsumed by the G-DINA model (de la Torre, 2011). These include the DINA model, the *deterministic input, noisy "or" gate* (DINO; Templin & Henson, 2006) model, and the *additive CDM* (*A*-CDM; de la Torre, 2011). As constrained CDMs, the DINA model assumes that lacking one of the required attributes is as the same as lacking all of the required attributes; the DINO model assumes that having one of the required attributes is as the same as having all of the required attributes; and the *A*-CDM assumes that the impacts of mastering the different required attributes are independent of each other.

2.1.2 Computerized Adaptive Testing

CAT has become a popular tool to estimate examinees' ability levels with shorter test lengths. The main goal of CAT is to construct an optimal test for each examinee. Appropriate items to each examinee's ability level are selected from an item bank, and the ability level is estimated during or end of the test administration. Therefore, different tests including different items with different lengths can be created for different examinees. Weiss and Kingsbury (1984) listed the components of CAT, which include item selection method and calibrated item pool. In addition, CAT can be used with different psychometric frameworks such as IRT or CDM. The Fisher information statistic (Lehmann & Casella, 1998) is widely used in the traditional CAT;

however, it cannot be applied in CD-CAT because it requires continuous ability levels, whereas the attribute vectors in cognitive diagnosis are discrete. Fortunately, the Kullback-Leibler (K-L) information, which is an alternative information statistic, can work under both continuous and discrete cases. This study focuses on item selection methods in the cognitive diagnosis context, which include K-L-based indices.

2.1.2.1 The Posterior-Weighted Kullback-Leibler Index

The K-L information is a measure of distance between the two probability density functions, $f(x)$ and $g(x)$, where $f(x)$ is assumed to be the true distribution of the data (Cover & Thomas, 1991). The function measuring the distance between f and g is given by

$$K(f,g) = \int \left[log\left(\frac{f(x)}{g(x)}\right)\right] f(x)dx. \tag{2.1}$$

Larger information allows easier differentiation between the two distributions or likelihoods (Lehmann & Casella, 1998). Xu et al. (2003) used the K-L information as an item selection index in CD-CAT. Cheng (2009) proposed the PWKL, which computes the index using the posterior distribution of the attribute vectors as weights. Her simulation study showed that the PWKL outperformed the K-L information in terms of estimation accuracy. The PWKL is given by

$$PWKL_j(\hat{\boldsymbol{\alpha}}_i^{(t)}) = \sum_{c=1}^{2^K}\left[\sum_{x=0}^{1} log\left(\frac{P(X_j=x|\hat{\boldsymbol{\alpha}}_i^{(t)})}{P(X_j=x|\boldsymbol{\alpha}_c)}\right) P(X_j=x|\hat{\boldsymbol{\alpha}}_i^{(t)})\pi_i^{(t)}(\boldsymbol{\alpha}_c)\right], \tag{2.2}$$

where $P(X_j=x|\boldsymbol{\alpha}_c)$ is the probability of the response x to item j given the attribute vector $\boldsymbol{\alpha}_c$, and $\pi_i^{(t)}(\boldsymbol{\alpha}_c)$ is the posterior probability of examinee i given the responses to the t items. The posterior distribution after tth response can be written as

$$\pi_i^{(t)}(\boldsymbol{\alpha}_c) \propto \pi_i^{(0)}(\boldsymbol{\alpha}_c)L(\boldsymbol{X}_i^{(t)}|\boldsymbol{\alpha}_c),$$

where $\boldsymbol{X}_i^{(t)}$ is the vector containing the responses of examinee i to the t items, $\pi_i^{(0)}(\boldsymbol{\alpha}_c)$ is the prior probability of $\boldsymbol{\alpha}_c$, and $L(\boldsymbol{X}_i^{(t)}|\boldsymbol{\alpha}_c)$ is the likelihood of $\boldsymbol{X}_i^{(t)}$ given the

attribute vector $\boldsymbol{\alpha}_c$. The $(t+1)$th item to be administered is the item that maximizes the PWKL.

2.1.2.2 The Modified Posterior-Weighted Kullback-Leibler Index

The PWKL is calculated by summing the distances between the current estimate of the attribute vector and the other possible attribute vectors using the K-L information, and it is weighted by the posterior distribution of the attribute vectors. By using the current estimate $\hat{\boldsymbol{\alpha}}_i^{(t)}$, it assumes that the point estimate is a good summary of the posterior distribution $\pi_i^{(t)}(\boldsymbol{\alpha})$. However, this may not be the case particularly when the test is still relatively short. Instead of using a point estimate, the new PWKL proposes modifying by considering the entire posterior distribution, which involves 2^K attribute vectors. The resulting new index can be referred to as the MPWKL and can be computed as

$$MPWKL_{ij}^{(t)} = \sum_{d=1}^{2^K}\left[\sum_{c=1}^{2^K}\left[\sum_{x=0}^{1} log\left(\frac{P(X_j=x|\boldsymbol{\alpha}_d)}{P(X_j=x|\boldsymbol{\alpha}_c)}\right)P(X_j=x|\boldsymbol{\alpha}_d)\pi_i^{(t)}(\boldsymbol{\alpha}_c)\right]\pi_i^{(t)}(\boldsymbol{\alpha}_d)\right]. \quad (2.3)$$

Compared with the PWKL, by using the posterior distribution, the MPWKL does not require estimating the attribute vector $\boldsymbol{\alpha}_i^{(t)}$. Using an estimate in the numerator of Equation 2.2 is tantamount to assigning a single attribute vector (i.e., $\boldsymbol{\alpha}_i^{(t)}$) a probability of 1, which may not accurately describe the posterior distribution at the early stages of the testing administration. In contrast, the numerator in Equation 2.3 considers all the possible attribute vectors, and weights them accordingly, hence, the extra summation and posterior probability. Because the MPWKL uses the entire posterior distribution $\pi_i^{(t)}(\boldsymbol{\alpha})$ rather than just an estimate $\hat{\boldsymbol{\alpha}}_i^{(t)}$, it can be expected to be more informative than the PWKL.

2.1.2.3 The G-DINA Model Discrimination Index

The GDI, which measures the (weighted) variance of the probabilities of success of an item given a particular attribute distribution, was first proposed by de la Torre and Chiu (2010) as an index to implement an empirical Q-matrix validation procedure. However, in this article, the index is used as an item selection method for CD-CAT. To define the index, let the first K_j^* attributes be required for item j, and define $\boldsymbol{\alpha}_{cj}^*$ as the reduced attribute vector consisting of the first K_j^* attributes, for $c = 1, \ldots, 2^{K_j^*}$. For example, if a q-vector is defined as (1,1,0,0,1) for $K_j^* = 3$ number of required attributes, the reduced attribute vector is (a_1,a_2,a_5). Also, define $\pi(\boldsymbol{\alpha}_{cj}^*)$ as the probability of $\boldsymbol{\alpha}_{cj}^*$, and $P(X_{ij} = 1|\boldsymbol{\alpha}_{cj}^*)$ as the success probability on item j given $\boldsymbol{\alpha}_{cj}^*$. The GDI for item j is defined as

$$\varsigma_j^2 = \sum_{c=1}^{2^{K_j^*}} \pi(\boldsymbol{\alpha}_{cj}^*)[P(X_{ij} = 1|\boldsymbol{\alpha}_{cj}^*) - \bar{P}_j]^2, \tag{2.4}$$

where $\bar{P}_j = \sum_{c=1}^{2^{K_j^*}} \pi(\boldsymbol{\alpha}_{cj}^*)P(X_{ij} = 1|\boldsymbol{\alpha}_{cj}^*)$ is the mean success probability. In CD-CAT applications, the posterior probability of the reduced attribute vector $\pi_i^{(t)}(\boldsymbol{\alpha}_{cj}^*)$ is used in place of $\pi(\boldsymbol{\alpha}_{cj}^*)$. This implies that the discrimination of an item is not static, and changes as the posterior distribution changes with t. The GDI measures the extent to which an item can differentiate between the different reduced attribute vectors based on their success probabilities, and is minimum (i.e., equal to zero) when $P(X_{ij} = 1|\boldsymbol{\alpha}_{1j}^*) = P(X_{ij} = 1|\boldsymbol{\alpha}_{2j}^*) = P(X_{ij} = 1|\boldsymbol{\alpha}_{2^{K_j^*}j}^*) = \bar{P}_j$ (or, trivially, when the posterior distribution is degenerate). It also attaches greater importance to reduced attribute vectors with higher $\pi(.)$. As such, a larger GDI indicates a greater ability to differentiate between reduced attribute vectors that matter. The GDI is computed for each candidate item in the pool, and the candidate item with the largest GDI is selected.

The GDI has two important properties. First, instead of the original attribute vector, $\boldsymbol{\alpha}_c$, it uses the reduced attribute vector, $\boldsymbol{\alpha}_{cj}^*$. Consequently, the GDI can be implemented more efficiently than can the PWKL or MPWKL. For example, if $K = 5$ and $K_j^* = 2$, computing the GDI involves $2^{K_j^*} = 4$ terms, whereas the PWKL and MPWKL involve $2^K = 32$ and $2^K \times 2^K = 1,024$ terms, respectively.

Second, the GDI takes both the item discrimination and the posterior distribution into account. This property is illustrated using the example in Table 2.1. It involves $K = 3$, and six items, three of which are of low discrimination (LD), and the other three are of high discrimination (HD). For the low-discriminating items, the difference between the lowest and the highest probabilities of success is 0.4; for the high-discriminating items, this difference is 0.8. In addition, these items involve one of the following q-vectors: $\boldsymbol{q}_{100}$, $\boldsymbol{q}_{110}$, and $\boldsymbol{q}_{111}$. Four distributions are considered: (1) all attribute vectors are equally probable, as in, $\pi(\boldsymbol{\alpha}_c) = 0.125$; in (2), (3), and (4), the attribute vector, namely, (1,0,0), (1,1,0), and (1,1,1), respectively, has a probability of .965 and was deemed dominant, whereas each of the remaining attribute vectors has a probability of .005. In Condition 1, the impact of the posterior distribution is discounted, whereas in Conditions 2, 3, and 4, one-attribute vector is highly dominant. In this table, the GDI was computed using the DINA model.

Table 2.1: GDIs for Different Distribution, Item Discrimination, and Q-Vectors

Condition	Dominant $\boldsymbol{\alpha}$	Low Discrimination			High Discrimination		
		$\boldsymbol{q}_{100}$	$\boldsymbol{q}_{110}$	$\boldsymbol{q}_{111}$	$\boldsymbol{q}_{100}$	$\boldsymbol{q}_{110}$	$\boldsymbol{q}_{111}$
1	None	**0.090**	0.068	0.039	**0.160**	0.120	0.070
2	(1,0,0)	**0.007**	0.004	0.002	**0.013**	0.006	0.003
3	(1,1,0)	0.007	**0.010**	0.002	0.013	**0.019**	0.003
4	(1,1,1)	0.007	0.010	**0.012**	0.013	0.019	**0.022**

Note. Numbers in bold represent the highest GDI in each condition for fixed item discrimination. GDI = G-DINA model discrimination index; G-DINA = generalized DINA; DINA = deterministic inputs, noisy "and" gate.

Several results can be noted. First, for a fixed q-vector, the high-discriminating items had higher GDI values compared to the low-discriminating items regardless of

the posterior distribution. Second, when there was no dominant attribute vector, one-attribute items had the highest GDI values for a fixed item discrimination. In contrast, when one attribute vector was highly dominant, the items with q-vectors matching the dominant attribute vectors had the highest GDI values. Finally, it can also be observed that the low-discriminating items with q-vectors that match the dominant attribute vectors can at times be preferred over the high-discriminating items with q-vectors that do not. For example, for attribute vector (1,1,0), the GDI for the low-discriminating item with $\boldsymbol{q}_{110}$ is 0.010. This is higher than the GDI for the high-discriminating item with $\boldsymbol{q}_{111}$, which is 0.003.

Based on the properties of the three indices discussed earlier, the authors expect the GDI and the MPWKL will be more informative than the PWKL. In addition, they expect the GDI to be faster than the PWKL in terms of implementation time, which in turn will be faster than MPWKL.

2.2 Simulation Study

The simulation study aimed to investigate the efficiency of the MPWKL and GDI compared to the PWKL under the G-DINA model context considering a variety of factors, namely, item quality, generating model, and test termination rule. The correct attribute and attribute vector classification rates, and a few descriptive statistics (i.e., minimum, maximum, mean, and coefficient of variation [CV]), of the test lengths were calculated based on the termination rules to compare the efficiency of the item selection indices. In addition, the time required to administer the test was also recorded for each of the item selection indices. Finally, the item usage in terms of the required attributes was tracked and reported in each condition.

2.2.1 Design

2.2.1.1 Data Generation

Different item qualities and reduced CDMs were considered in the data generation. First, due to documented impact of item quality on attribute classification accuracy (e.g., de la Torre, Hong, & Deng, 2010), different item discriminations and variances were used in the data generation. Two levels of item discrimination, HD and LD, were combined with two levels of variance, high variance (HV) and low variance (LV), in generating the item parameters. Thus, a total of four conditions, HD-LV, HD-HV, LD-LV, and LD-HV, were considered in investigating the impact of item quality on the efficiency of the indices. The item parameters were generated from uniform distributions. For HD items, the highest and lowest probabilities of success, $P(\mathbf{0})$ and $P(\mathbf{1})$, were generated from distributions with means of .1 and .9, respectively; for LD items, these means were 0.2 and 0.8. For HV and LV items, the ranges of the distribution were 0.1 and 0.2, respectively. The distributions for $P(\mathbf{0})$ and $P(\mathbf{1})$ under different discrimination and variance conditions are given in Table 2.2. The mean of the distribution determines the overall quality of the item pool, whereas the variance determines the overall quality of the administered items.

Table 2.2: Item Parameters

Item Quality	$P(\mathbf{0})$	$P(\mathbf{1})$
HD-LV	$U(0.05, 0.15)$	$U(0.85, 0.95)$
HD-HV	$U(0.00, 0.20)$	$U(0.80, 1.00)$
LD-LV	$U(0.15, 0.25)$	$U(0.75, 0.85)$
LD-HV	$U(0.10, 0.30)$	$U(0.70, 0.90)$

Note. HD-LV = high discrimination-low variance; HD-HV = high discrimination-high variance; LD-LV = low discrimination-low variance; LD-HV = low discrimination-high variance.

Second, to investigate whether the efficiency of the indices is consistent across different models, item responses were generated using three reduced models: the DINA model, the DINO model, and the A-CDM. For the DINA and DINO models,

the probability of success was set as shown in Table 2.2. For the A-CDM, in addition to the success probabilities given in Table 2.2, intermediate success probabilities were obtained by allowing each of the required attributes to contribute equally. The four item qualities and three reduced models resulted in the 12 conditions of the simulation study. The number of attributes was fixed to $K = 5$.

To design a more efficient simulation study, only a subset of the attribute vectors was considered. The six attribute vectors were $\boldsymbol{\alpha}_0 = (0, 0, 0, 0, 0)$, $\boldsymbol{\alpha}_1 = (1, 0, 0, 0, 0)$, $\boldsymbol{\alpha}_2 = (1, 1, 0, 0, 0)$, $\boldsymbol{\alpha}_3 = (1, 1, 1, 0, 0)$, $\boldsymbol{\alpha}_4 = (1, 1, 1, 1, 0)$, and $\boldsymbol{\alpha}_5 = (1, 1, 1, 1, 1)$, representing no mastery, mastery of a single attribute only, mastery of two attributes only, and so forth. For each attribute vector, 1,000 examinees were generated for a total of 6,000 examinees in each condition.

2.2.1.2 Test Termination Rules

Two test termination rules were considered in the simulation study: fixed-test lengths and minimax of the posterior distribution of the attribute vectors. The former allowed for a comparison of the efficiency of the indices with respect to classification accuracy when the CAT administration was stopped after a prespecified test length was reached for each examinee; the latter allowed for the comparison of the efficiency of the indices in terms of test lengths when the CAT administration was terminated after the largest posterior probability of an attribute vector was at least as large as a prespecified minimax value, which corresponds to the first criterion by Hsu et al. (2013). Three fixed-test lengths, 10, 20, and 40 items, were considered for the first termination rule, and four minimax values, 0.65, 0.75, 0.85, and 0.95, were used for the second rule.

2.2.1.3 Item Pool and Item Selection Methods

The Q-matrix was created to have 40 items from each of $2^K - 1 = 31$ possible q-vectors, resulting in 1,240 items in the pool. Three different item selection indices were considered: the PWKL, the MPWKL, and the GDI. For greater comparability, the first item administered to each examinee was chosen at random, and this item was fixed across the three indices. In the case of PWKL, when $\hat{\boldsymbol{\alpha}}_i^{(t)}$ was not unique, a random attribute vector was chosen from the modal attribute vectors.

Let α_{ikl} and $\hat{\alpha}_{ikl}$ be the kth true and estimated attribute in attribute vector l for examinee i, respectively. For each of the six attribute vectors considered in this design, the correct attribute classification rates (CAC), and the correct attribute vector classification rates (CVC) were computed as

$$\begin{aligned} CAC_l &= \frac{1}{1,000}\sum_{i=1}^{1,000}\sum_{k=1}^{5} I[\alpha_{ikl} = \hat{\alpha}_{ikl}], \text{ and} \\ CVC_l &= \frac{1}{1,000}\sum_{i=1}^{1,000}\prod_{k=1}^{5} I[\alpha_{ikl} = \hat{\alpha}_{ikl}], \end{aligned} \tag{2.5}$$

where $l = 0, \ldots, 5$, and I is the indicator function.Using appropriate weights (described later), the CAC and CVC were computed assuming the attributes were uniformly distributed for the fixed-test length conditions. The minimum, maximum, mean, and CV of the test lengths were calculated, again with appropriate weights where needed, when the minimax of the posterior distribution was used as the stopping criterion. This study focused on attribute vectors that were uniformly distributed. To accomplish this, the results based on the six attribute vectors needed to be weighted appropriately. For $K = 5$, the vector of the weights are 1/32, 5/32, 10/32, 10/32, 5/32, and 1/32, which represented the proportions of zero-, one-, two-, three-, four-, and five-attribute mastery vectors among the 32 attribute vectors, respectively. CV was calculated by taking the ratio of the standard deviation to the

mean.

2.2.2 Results

2.2.2.1 Fixed-Test Length

The sampling design of this simulation study can allow for results to be generalized to different distributions of the attribute vectors. This study focused on attribute vectors that were uniformly distributed. To demonstrate the efficiency of using such a design, a small study comparing two sampling procedures for the DINA model with HD-LV items was carried out. In the first procedure, which is the current sampling design, only six selected attribute vectors, each with 1,000 replicates, were used; in the second procedure, 32,000 attribute vectors were generated uniformly. The CAC and the CVC in the former and the latter were computed using weighted and simple averages, respectively. Table 2.3 shows that despite working with fewer attribute vectors, using selected attribute vectors can give the CAC and the CVC that were almost identical to those obtained using a much larger sample drawn randomly, and this was true across the different test lengths. These findings can be expected to hold across other CDMs and item qualities.

Table 2.3: Classification Accuracies Based on Two Sampling Procedures

Item		CAC		CVC	
Quality	J	Weighted	Simple	Weighted	Simple
HD-LV	10	0.969	0.969	0.875	0.876
	20	0.999	0.999	0.996	0.996
	40	1.000	1.000	1.000	1.000

Note. CAC = correct attribute classification; CVC = correct attribute vector classification; J = test length; HD-LV = high discrimination-low variance.

For all conditions, the CAC rates were, as expected, higher than the CVC rates, but both measures showed similar patterns. For this reason, only the CVC rates were reported in this article. However, the results in their entirety can be requested from

the first author. The CVC results using fixed-test lengths as a stopping rule under the different factors are presented in Table 2.4 for all the generating models. Differences in the CVC rates were evaluated using two cut points, 0.01 and 0.10. Differences below 0.01 were considered negligible, between 0.01 and 0.10 were considered slight, and above 0.10 were considered substantial.

Table 2.4: The CVC Rates using the DINA, DINO, and *A*-CDM

Item		DINA			DINO			*A*-CDM		
Quality	J	PWKL	MPWKL	GDI	PWKL	MPWKL	GDI	PWKL	MPWKL	GDI
HD-LV	10	0.752	0.878	0.887	0.749	0.855	0.849	0.839	0.817	0.826
	20	0.989	0.996	0.996	0.986	0.995	0.996	0.992	0.992	0.991
	40	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
HD-HV	10	0.854	0.979	0.981	0.870	0.979	0.981	0.963	0.967	0.962
	20	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	40	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
LD-LV	10	0.454	0.589	0.604	0.441	0.551	0.557	0.515	0.524	0.511
	20	0.814	0.892	0.890	0.803	0.872	0.871	0.855	0.857	0.859
	40	0.987	0.995	0.995	0.984	0.993	0.992	0.987	0.990	0.990
LD-HV	10	0.569	0.723	0.719	0.596	0.703	0.704	0.658	0.666	0.660
	20	0.917	0.962	0.962	0.924	0.969	0.966	0.948	0.953	0.951
	40	0.999	1.000	0.999	1.000	1.000	1.000	0.998	0.999	0.999

Note. CVC = correct attribute vector classification; DINA = deterministic inputs, noisy "and" gate; DINO = deterministic input, noisy "or" gate; *A*-CDM = additive CDM; CDM = cognitive diagnosis model; J = test length; PWKL = posterior-weighted Kullback-Leibler index; MPWKL = modified PWKL index; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; HD-LV = high discrimination-low variance; HD-HV = high discrimination-high variance; LD-LV = low discrimination-low variance; LD-HV = low discrimination-high variance.

Using the DINA and the DINO as generating models in conjunction with a short test length (i.e., 10 items), the differences in the CVC rates of the MPWKL and the GDI were mostly negligible regardless of the item quality. The only exception is the one condition, with 10 LD-LV items, where the CVC rate of the GDI was slightly higher than the MPWKL. Under the same conditions, the CVC rates of the two indices were substantially higher than the PWKL regardless of the item quality. When the test lengths were longer (i.e., 20- and 40-item tests), all of the three indices generally performed similarly using the DINA and DINO models. However, in one condition (i.e., 20-item test with LD items and the DINA model), the MPWKL and the GDI had slightly higher CVC rates compared with the PWKL.

Using the *A*-CDM as a generating model, the three indices had mostly similar

CVC rates. Interestingly, using 10-item tests with HD-LV items, the PWKL had slightly higher CVC rates compared to the MPWKL and the GDI.

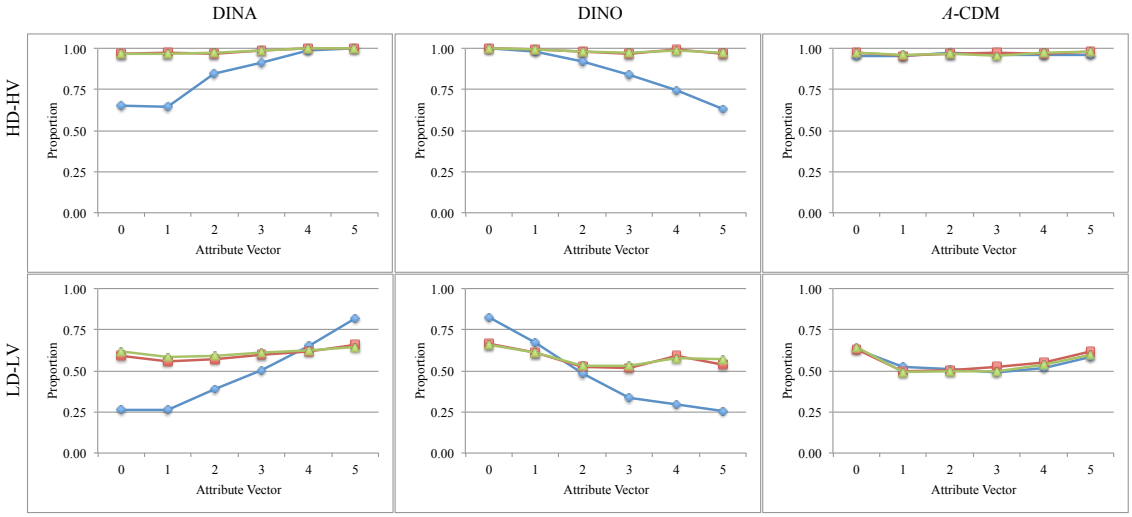
Additional findings can be culled from Table 2.4. First, as expected, increasing the test length improved the classification accuracy regardless of the item selection index, item quality, and generating model. Using a long test (i.e., 40-item test) provided a CVC rate of almost 1.00 for all of the indices. However, a clear distinction can be seen on the efficiency of the indices when shorter test lengths, in particular 10-item test, were used. For example, using the DINA model and HD-LV items, the 10-item test yielded a maximum CVC rate of 0.89 for the MPWKL and the GDI. In comparison, the PWKL had only a CVC rate of 0.75 under the same condition.

Second, the item quality had an obvious impact on the CVC rates: higher discrimination and higher variance resulted in higher classification accuracy. As can be seen from the results, HD items resulted in better rates compared with LD items regardless of the variance. Similarly, items with HV showed higher classification rates compared with LV items. Consequently, HD-HV items had the best classification accuracy, whereas LD-LV items had the worst classification accuracy regardless of the item selection index and generating model. To illustrate, using the DINA model and a 10-item test, the highest and the lowest CVC rates of 0.98 and 0.60, were obtained with HD-HV and LD-LV items, respectively, for both the MPWKL and GDI; in comparison, the CVC rates were 0.85 and 0.45 for HD-HV and LD-LV items, respectively, for the PWKL.

To investigate how the item selection indices behaved for different attribute vectors, the CVC rates for each attribute vector were calculated. Only the results for 10-item test with HD-HV and LD-LV items are presented (see Figure 2.1). Across the different item quality conditions, the CVC rates of the MPWKL and GDI were more similar for the different attribute vectors, whereas they were more varied for the PWKL. A few conclusions can be drawn from this figure. First, for HD-HV items, the

indices performed similarly for $\boldsymbol{\alpha}_4$ and $\boldsymbol{\alpha}_5$ when the DINA model was used. However, under the same condition, the MWPKL and GDI had higher CVC rates compared to the PWKL for the other four attribute vectors. Using the same item quality, the indices performed similarly for $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ when the DINO model was used; however, the CVC rates using the PWKL were lower for $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$, $\boldsymbol{\alpha}_4$, and $\boldsymbol{\alpha}_5$ compared to the other two indices. It can also be noted that the classification accuracy of the PWKL was more varied than those of the MPWKL and GDI across the attribute vectors. As can be seen from the graphs, the CVC rate of the PWKL could range from around 0.65 to 1.00, whereas these rates were mostly 1.00 for the MPWKL and GDI. The three indices had almost the same results when the A-CDM was involved.

Figure 2.1: CVC Rates for 6 Selected Attribute Vectors, $J = 10$



Note: Blue, red, and green lines represent the PWKL, MWPKL and GDI, respectively. CVC = correct attribute vector classification; PWKL = posterior-weighted Kullback-Leibler index; MPWKL = modified PWKL index; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; DINA = deterministic inputs, noisy "and" gate; DINO = deterministic input, noisy "or" gate; A-CDM = additive CDM; CDM = cognitive diagnosis model; HD-HV = high discrimination-high variance; LD-LV = low discrimination-low variance.

Second, although the CVC rates were lower, the results for LD-LV items were similar to those for HD-HV items. The MPWKL and GDI had higher CVC rates than the PWKL for $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, and $\boldsymbol{\alpha}_3$ when the DINA model was used. In contrast, the PWKL outperformed the MPWKL and GDI for $\boldsymbol{\alpha}_4$ and $\boldsymbol{\alpha}_5$ in the same condition.

Using the same item quality and the DINO model, the PWKL had higher CVC rates for $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$. However, the MPWKL and GDI had higher rates for the other four attribute vectors. Again, the CVC rates of the PWKL had higher variability (0.26-0.82) compared to those of the MPWKL and GDI (0.56-0.65). Finally, the efficiency of the indices was similar for the *A*-CDM, but the extreme attribute vectors $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_5$ can be better estimated than the remaining attribute vectors.

2.2.2.2 Minimax of the Posterior Distribution

For a fixed minimax of the posterior distribution, descriptive statistics of the test lengths are shown in Tables 2.5, 2.6, and 2.7 for the DINA, DINO and *A*-CDM, respectively. Differences in the mean were evaluated using two cut points, 0.5 and 1, and differences below 0.5 were considered negligible, between 0.5 and 1 slight, and above 1 substantial.

Table 2.5: Descriptive Statistics of Test Lengths using the DINA Model

Item		PWKL				MPWKL				GDI			
Quality	$\pi(\alpha_c\|\mathbf{X_i})$	Min	Max	Mean	CV	Min	Max	Mean	CV	Min	Max	Mean	CV
HD-LV	0.65	3	25	8.26	0.28	3	16	6.69	0.13	3	14	6.67	0.13
	0.75	3	28	8.92	0.28	4	18	7.32	0.17	4	19	7.34	0.16
	0.85	3	32	10.08	0.27	4	22	8.83	0.19	4	24	8.87	0.19
	0.95	4	35	12.05	0.25	4	26	10.99	0.19	4	31	10.99	0.19
HD-HV	0.65	2	19	7.76	0.22	2	14	6.55	0.11	2	10	6.52	0.12
	0.75	2	22	7.96	0.22	2	14	6.58	0.11	2	11	6.60	0.11
	0.85	2	23	8.45	0.22	2	14	6.72	0.10	2	14	6.73	0.10
	0.95	2	23	9.36	0.21	2	17	7.51	0.12	2	18	7.22	0.11
LD-LV	0.65	4	48	13.48	0.37	5	32	11.41	0.30	5	34	11.46	0.30
	0.75	4	50	15.21	0.36	6	40	13.02	0.29	6	38	13.08	0.29
	0.85	5	55	17.11	0.35	6	53	14.95	0.30	6	55	15.00	0.30
	0.95	5	73	21.43	0.32	7	56	19.40	0.28	7	64	19.46	0.28
LD-HV	0.65	3	35	10.45	0.32	4	28	8.60	0.24	4	28	8.57	0.24
	0.75	4	36	11.71	0.32	5	29	9.86	0.26	5	31	9.93	0.26
	0.85	4	42	13.41	0.31	5	32	11.78	0.26	5	32	11.77	0.26
	0.95	4	49	15.70	0.29	6	43	14.13	0.25	6	42	14.17	0.25

Note. DINA = deterministic inputs, noisy "and" gate; PWKL = posterior-weighted Kullback-Leibler index; MPWKL = modified PWKL index; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; CV = coefficient of variation; HD-LV = high discrimination-low variance; HD-HV = high discrimination-high variance; LD-LV = low discrimination-low variance; LD-HV = low discrimination-high variance.

Using the DINA and DINO models, the mean test lengths of the MPWKL and the GDI were generally similar (i.e., the differences were negligible), and they were

Table 2.6: Descriptive Statistics of Test Lengths using the DINO Model

Item		PWKL				MPWKL				GDI			
Quality	$\pi(\alpha_c\|\mathbf{X_i})$	Min	Max	Mean	CV	Min	Max	Mean	CV	Min	Max	Mean	CV
HD-LV	0.65	3	24	8.37	0.28	3	24	6.89	0.15	3	19	6.83	0.15
	0.75	3	27	9.08	0.28	4	27	7.58	0.18	4	21	7.58	0.18
	0.85	3	29	10.32	0.27	4	28	8.91	0.19	4	23	8.89	0.19
	0.95	4	34	12.23	0.25	4	30	11.09	0.20	4	27	11.04	0.20
HD-HV	0.65	2	17	7.78	0.23	2	10	6.60	0.11	2	10	6.53	0.11
	0.75	3	18	8.04	0.22	3	13	6.71	0.10	3	10	6.60	0.11
	0.85	3	22	8.61	0.23	3	14	7.24	0.10	3	11	6.80	0.10
	0.95	3	26	9.51	0.22	3	17	8.10	0.10	3	20	7.66	0.11
LD-LV	0.65	4	49	13.73	0.35	5	40	11.88	0.29	5	37	11.85	0.29
	0.75	4	50	15.43	0.34	6	43	13.61	0.29	6	43	13.61	0.29
	0.85	5	59	17.41	0.33	6	57	15.57	0.30	6	62	15.60	0.30
	0.95	5	67	21.83	0.31	7	69	20.10	0.29	7	68	20.07	0.29
LD-HV	0.65	3	32	10.45	0.31	4	24	8.81	0.23	4	24	8.75	0.23
	0.75	4	33	11.79	0.30	5	36	10.18	0.25	5	29	10.08	0.25
	0.85	4	36	13.45	0.30	5	36	12.13	0.24	5	42	12.11	0.25
	0.95	5	45	15.75	0.28	6	40	14.40	0.25	6	43	14.35	0.26

Note. DINO = deterministic input, noisy "or" gate; PWKL = posterior-weighted Kullback-Leibler index; MPWKL = modified PWKL index; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; DINA = deterministic inputs, noisy "and" gate; CV = coefficient of variation; HD-LV = high discrimination-low variance; HD-HV = high discrimination-high variance; LD-LV = low discrimination-low variance; LD-HV = low discrimination-high variance.

substantially shorter compared with the test lengths of the PWKL. This was true regardless of the minimax value and item quality. The largest mean test length differences occurred when LD-LV items were involved − these differences were greater than 2.0 and 1.8 for the DINA and DINO models, respectively. However, when the *A*-CDM was used, all the three indices performed similarly except in the HD-HV and 0.85 minimax value condition, where the PWKL had a slightly longer test length compared with the MPWKL and GDI.

It can also be noted that, as expected, increasing the minimax value resulted in longer test lengths regardless of the item selection index, item quality, and generating model. The change in the mean test length as a result of increasing the minimax value from 0.65 to 0.95 was substantial for all of the conditions except for one − there was only a slight change when the MPWKL and the GDI were used with HD-HV items. In addition, as in the fixed-test length, the item quality had an impact on the efficiency of the indices: Using items with higher discrimination or higher variance resulted in shorter tests. Consequently, HD-HV and LD-LV items had the shortest and the

Table 2.7: Descriptive Statistics of Test Lengths using the *A*-CDM

Item		PWKL				MPWKL				GDI			
Quality	$\pi(\alpha_c\|\mathbf{X_i})$	Min	Max	Mean	CV	Min	Max	Mean	CV	Min	Max	Mean	CV
HD-LV	0.65	6	13	6.99	0.10	6	13	6.92	0.08	6	12	6.93	0.08
	0.75	6	14	7.86	0.14	7	14	7.75	0.12	7	15	7.76	0.12
	0.85	9	18	9.98	0.14	9	19	9.74	0.13	9	20	9.79	0.13
	0.95	11	25	12.84	0.16	11	26	12.82	0.15	11	26	12.84	0.15
HD-HV	0.65	6	10	6.83	0.08	6	7	6.75	0.06	6	7	6.70	0.07
	0.75	6	14	7.18	0.12	6	8	6.83	0.06	6	8	6.80	0.06
	0.85	6	17	7.87	0.15	6	11	7.18	0.07	6	11	7.04	0.06
	0.95	6	17	8.71	0.16	6	15	8.79	0.09	7	16	8.79	0.10
LD-LV	0.65	10	32	12.67	0.21	10	30	12.65	0.22	10	29	12.62	0.22
	0.75	11	33	14.66	0.23	11	34	14.67	0.23	11	34	14.64	0.23
	0.85	12	50	17.80	0.24	12	49	17.84	0.24	12	52	17.82	0.24
	0.95	16	59	24.41	0.23	16	74	24.39	0.23	16	74	24.37	0.23
LD-HV	0.65	8	22	9.04	0.17	8	21	9.04	0.17	8	18	9.03	0.17
	0.75	9	26	11.25	0.19	9	24	11.30	0.19	9	24	11.26	0.19
	0.85	11	27	13.29	0.19	11	32	13.20	0.18	11	29	13.20	0.18
	0.95	13	39	17.43	0.20	13	38	17.49	0.19	13	41	17.48	0.20

Note. *A*-CDM = additive CDM; CDM = cognitive diagnosis model; PWKL = posterior-weighted Kullback-Leibler index; MPWKL = modified PWKL index; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; DINA = deterministic inputs, noisy "and" gate; CV = coefficient of variation; HD-LV = high discrimination-low variance; HD-HV = high discrimination-high variance; LD-LV = low discrimination-low variance; LD-HV = low discrimination-high variance.

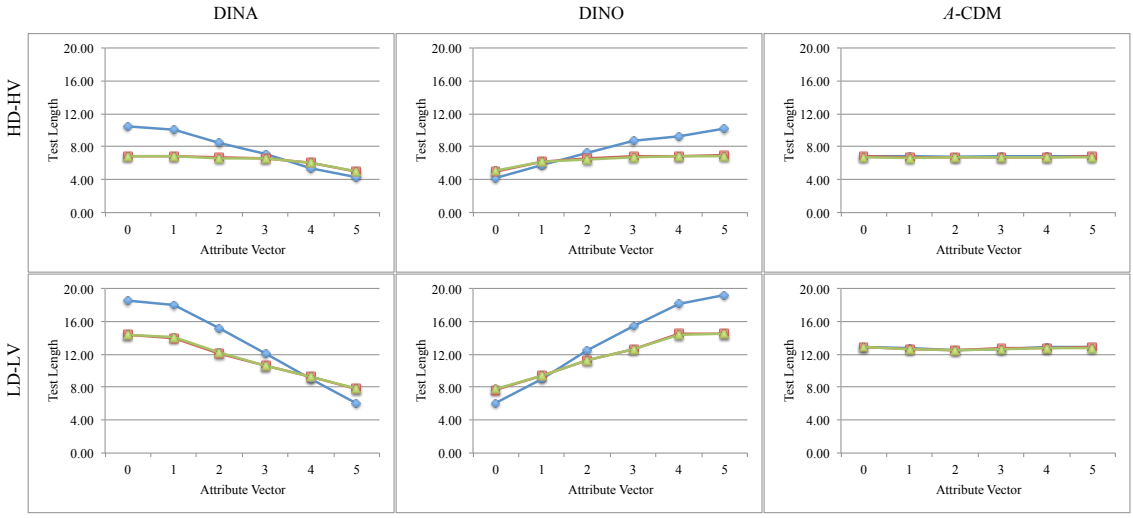
longest tests, respectively. In this study, using the minimax value of 0.95, GDI, and DINA model, HD-HV items resulted in tests with a mean of 7.22; in contrast, for LD-LV items, this mean was 19.46. Finally, generating model can have an impact on the mean test lengths, but this moderated by the choice of the item selection index − with the GDI, the DINA or DINO models consistently required shorter tests than the *A*-CDM, but this pattern was not as obvious with the other two indices.

Other findings can be gleaned from Tables 2.5, 2.6, and 2.7. First, the minimum test lengths of the three indices were similar for most of the conditions. Second, increasing the minimax of the posterior distribution generally resulted in higher minimum and maximum test lengths, especially at the two extreme minimax values. However, using HD-HV items with the DINA model, the minimum values remained the same for the three indices. Third, the item quality had an impact on the minimum, maximum, and CV of the test lengths: HD-HV items provided the smallest minimum, maximum, and CV values, whereas LD-LV items provided the largest statistics for

all of the indices. Finally, using the *A*-CDM, the indices had the smallest maximum and CV values; however, they had the highest minimum test lengths compared to the DINA and DINO models.

The mean test lengths for each attribute vector were calculated, and the results using HD-HV and LD-LV items, and 0.65 as the minimax value are shown in Figure 2.2. For the DINA model, the PWKL required longer tests, on the average, for the attribute vectors $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_1$, and $\boldsymbol{\alpha}_2$ compared to the MPWKL and GDI; however, these two indices required longer tests for $\boldsymbol{\alpha}_5$. In contrast, the MPWKL and GDI required longer tests for $\boldsymbol{\alpha}_0$, and the PWKL required longer tests for $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$, $\boldsymbol{\alpha}_4$, and $\boldsymbol{\alpha}_5$ with the DINO as the generating model. Using the *A*-CDM, the mean test lengths were similar for each attribute vector.

Figure 2.2: Mean Test Lengths for 6 Selected Attribute Vectors, $\pi(\alpha_c|\mathbf{X_i}) = 0.65$



Note: Blue, red, and green represent the PWKL, MPWKL, and GDI, respectively. PWKL = posterior-weighted Kullback-Leibler index; MPWKL = modified PWKL index; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; DINA = deterministic inputs, noisy "and" gate; DINO = deterministic input, noisy "or" gate; *A*-CDM = additive CDM; CDM = cognitive diagnosis model; HD-HV = high discrimination-high variance; LD-LV = low discrimination-low variance.

2.2.2.3 Item Usage

To gain a better understanding of how different models utilize the items in the pool, the overall item usage in terms of the number of required attributes was recorded for each condition. Only the results for the fixed-test lengths with HD-HV and LD-LV items are shown in Table 2.8.

For the DINA and DINO models, items that required one, two, and three attributes were generally used more often compared to those which required four and five attributes regardless of the item selection index and item quality. The PWKL mostly used two-attribute items for the same models except in one condition where a 10-item test with LD-LV items and the DINA were used. The MPWKL and GDI had a similar pattern of item usage (i.e., one-attribute items were mostly used for 10- and 20-item tests with LD-LV items) across different test lengths and item qualities for the DINA except in one condition where a 10-item test with HD-HV items was used. However, for the *A*-CDM, one-attribute items were mostly used with a proportion of at least 0.92 regardless of the item selection index and item quality.

Table 2.8: The Proportion of Overall Item Usage

			PWKL					MPWKL					GDI				
True	Item		Number of Required Attributes														
Model	Quality	J	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
DINA	HD-HV	10	0.25	0.45	0.23	0.06	0.01	0.38	0.27	0.31	0.02	0.01	0.34	0.34	0.28	0.03	0.01
		20	0.25	0.48	0.22	0.04	0.01	0.27	0.39	0.30	0.03	0.01	0.27	0.37	0.29	0.05	0.02
		40	0.28	0.49	0.18	0.04	0.01	0.24	0.44	0.27	0.05	0.01	0.24	0.39	0.30	0.05	0.01
	LD-LV	10	0.26	0.30	0.34	0.08	0.02	0.50	0.30	0.16	0.03	0.01	0.52	0.29	0.15	0.03	0.01
		20	0.29	0.34	0.28	0.07	0.02	0.37	0.35	0.23	0.04	0.01	0.38	0.34	0.22	0.05	0.01
		40	0.25	0.34	0.31	0.08	0.02	0.27	0.35	0.30	0.07	0.02	0.28	0.35	0.29	0.07	0.02
DINO	HD-HV	10	0.26	0.44	0.23	0.07	0.01	0.30	0.38	0.27	0.04	0.01	0.36	0.28	0.30	0.05	0.01
		20	0.25	0.44	0.24	0.06	0.01	0.28	0.40	0.26	0.06	0.01	0.23	0.41	0.26	0.08	0.02
		40	0.24	0.46	0.24	0.05	0.01	0.21	0.49	0.23	0.06	0.01	0.22	0.41	0.29	0.07	0.01
	LD-LV	10	0.23	0.32	0.32	0.11	0.02	0.46	0.32	0.17	0.04	0.01	0.49	0.30	0.16	0.04	0.01
		20	0.27	0.33	0.29	0.09	0.02	0.35	0.36	0.22	0.05	0.01	0.37	0.34	0.22	0.05	0.01
		40	0.22	0.36	0.30	0.10	0.02	0.26	0.37	0.27	0.08	0.02	0.26	0.37	0.27	0.08	0.02
A-CDM	HD-HV	10	0.92	0.03	0.03	0.02	0.00	0.92	0.03	0.03	0.02	0.00	0.92	0.03	0.03	0.02	0.00
		20	0.95	0.02	0.02	0.01	0.00	0.96	0.02	0.02	0.01	0.00	0.96	0.02	0.02	0.01	0.00
		40	0.93	0.06	0.01	0.00	0.00	0.96	0.03	0.01	0.00	0.00	0.98	0.01	0.01	0.00	0.00
	LD-LV	10	0.92	0.03	0.03	0.02	0.00	0.92	0.03	0.03	0.02	0.00	0.92	0.03	0.03	0.02	0.00
		20	0.96	0.02	0.02	0.01	0.00	0.96	0.02	0.02	0.01	0.00	0.96	0.02	0.02	0.01	0.00
		40	0.98	0.01	0.01	0.00	0.00	0.98	0.01	0.01	0.00	0.00	0.98	0.01	0.01	0.00	0.00

Note. PWKL = posterior-weighted Kullback-Leibler index; MPWKL = modified PWKL index; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; DINA = deterministic inputs, noisy "and" gate; J = test length; DINO = deterministic input, noisy "or" gate; *A*-CDM = additive CDM; CDM = cognitive diagnosis model; HD-LV = high discrimination-low variance; HD-HV = high discrimination-high variance; LD-LV = low discrimination-low variance; LD-HV = low discrimination-high variance.

To get a deeper understanding of the differences in item usage among the models, the items were grouped based on their required attributes. To accomplish this, an additional simulation study was carried out using the same factors except for one: item quality. For this study, the lowest and highest success probabilities were fixed across all of the items, specifically, $P(\mathbf{0}) = 0.1$ and $P(\mathbf{1}) = 0.9$. This design aimed to eliminate the effect of the item quality on item usage. Due to the space constraint, only the results for the GDI, 20-item test, and $\boldsymbol{\alpha}_3$ are shown in Figure 2.3. Overall, the DINA model showed the following pattern of item usage: It uses items that required the same attributes as the examinee's true attribute mastery vector, and items that required single attributes which were not mastered by the examinee. For $\boldsymbol{\alpha}_3$, the DINA model used the items that required (1,1,1,0,0), and items that required either (0,0,0,1,0) or (0,0,0,0,1). In contrast, the DINO showed a different pattern of item usage: It uses items that required the same attributes as the examinee's true nonmastery vector, and items that required single attributes which were mastered by the examinee. Again for $\boldsymbol{\alpha}_3$, the DINO model used items that required (0,0,0,1,1), and items that required (1,0,0,0,0), (0,1,0,0,0), and (0,0,1,0,0). The A-CDM used items that required single attributes regardless of the true attribute vector. The same pattern was observed for the other attribute vectors.

To further investigate how the models converged into those patterns of item usage, the test administrations were divided into periods each comparing of five items. The item usage was recorded in each period. Only the results for the GDI, 20-item test, and $\boldsymbol{\alpha}_3$ are shown (refer to Figure 2.4). In the first period, which includes the first five items, one-attribute items were used mostly regardless of the generating model and examinees' true attribute vector. In the second, third, and fourth periods (items from 6 to 10, 11 to 15, and 16 to 20, respectively), the most common item types gradually became more similar to the previous patterns of item usage for the DINA and DINO models. However, the A-CDM favored one-attribute item at the rate of almost 1.00

in each period. Again, the same pattern was observed for the other attribute vectors in this study.

2.2.2.4 Average Time

The average item administration time per examinee was recorded separately for each index. The CAT administration code was written in Ox (Doornik, 2011), and run on a computer with processor of 2.5 GHz. Only the average times in milliseconds using 10 HD-LV items and the DINA model are shown in Table 2.9. The table shows that the MPWKL was the slowest, and the GDI was the fastest index in terms of the administration time: the PWKL, MPWKL, and GDI took 6.49, 20.25, and 4.59 milliseconds, respectively. In other words, the GDI was 4.41 faster than the MPWKL, and 1.41 faster than the PWKL. As mentioned earlier, the GDI works with the reduced attribute vectors, and involves fewer terms compared to the PWKL and MWPKL. The dimensions in the PWKL and MPWKL grow exponentially as the number of attribute K increases. However, the GDI does not have the same problem as long as the number of required attributes K_j^* remains small. The advantage of the GDI can be expected to be more apparent with the A-CDM because mostly one-attribute items are picked by the different indices.

Table 2.9: Average Test Administration Time per Examinee ($J = 10$, HD-LV, and DINA)

	PWKL	MPWKL	GDI
Time	6.49	20.25	4.59
Ratio (Relative to GDI)	1.41	4.41	–

Note. HD-LV = high discrimination-low variance; DINA = deterministic inputs, noisy "and" gate; PWKL = posterior-weighted Kullback-Leibler index; MPWKL = modified PWKL index; GDI = G-DINA model discrimination index; G-DINA = generalized DINA.

2.3 Discussion and Conclusion

Compared with traditional unidimensional IRT models, CDMs provide more information that can be used to inform instruction and learning. These models can reveal examinees' strengths and weaknesses by diagnosing whether they have mastered a specific set of attributes. CAT is a tool that can be used to create tests tailored for different examinees. This allows for a more efficient determination of what students know and do not know. In this article, two new item selection indices, the MPWKL and the GDI, were introduced, and their efficiency was compared with the PWKL. In addition, a more efficient simulation design was proposed in this study. This design can allow for results to be generalized to different distributions of attribute vectors, despite involving a smaller sample size. Based on the factors manipulated in the simulation study, the two new indices performed similarly, and they both outperformed the PWKL in terms of classification accuracy and test length. The study also showed that items with HD or HV provided better classification rates or shorter test lengths. Moreover, generating models can have an impact on the efficiency of the indices: For the DINA and DINO models, the results were more distinguishable; however, the efficiency of the indices was essentially the same for the A-CDM, except in a few conditions.

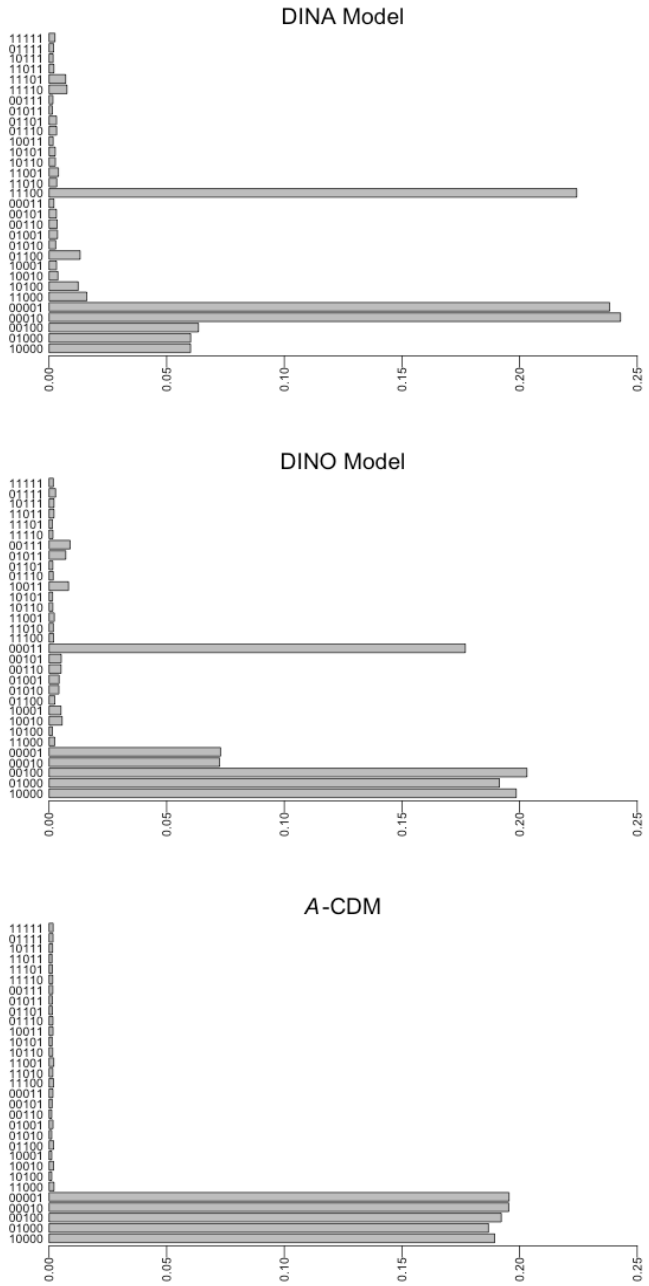
Although this study showed that the proposed indices, particularly the GDI, are promising, more research needs to be done to determine their viability. First, some constraints on the design of the Q-matrix and the size of the item pool need to be investigated. The Q-matrix in this study involved all the possible q-vectors. However, in practice, this may not be the case, particularly, when the CDMs are retrofitted to existing data. Therefore, it would be important to examinee how the indices perform when only a subset of the q-vectors exists in the pool. The current study uses a large item pool, which may not be always possible in real testing situations. Considering

smaller item pools, with or without constraints on the Q-matrix specifications, can lead to a better understanding of how the proposed indices perform under more varied conditions.

Second, although diagnostic assessments are primarily designed for low-stakes testing situations, their use for high-stakes decisions cannot be totally precluded. Because test security is a critical issue in high-stakes testing situations, item exposure in CD-CAT needs also to be controlled. At present, there are procedures for item exposure control in the context of CD-CAT. For example, Wang, Chang, and Huebner (2011) proposed item exposure control methods for fixed-test lengths in CD-CAT. However, the performance of these methods with the proposed indices has yet to be investigated. In addition, controlling the exposure of the items with the MPWKL and the GDI can also be examined when different termination rules are involved.

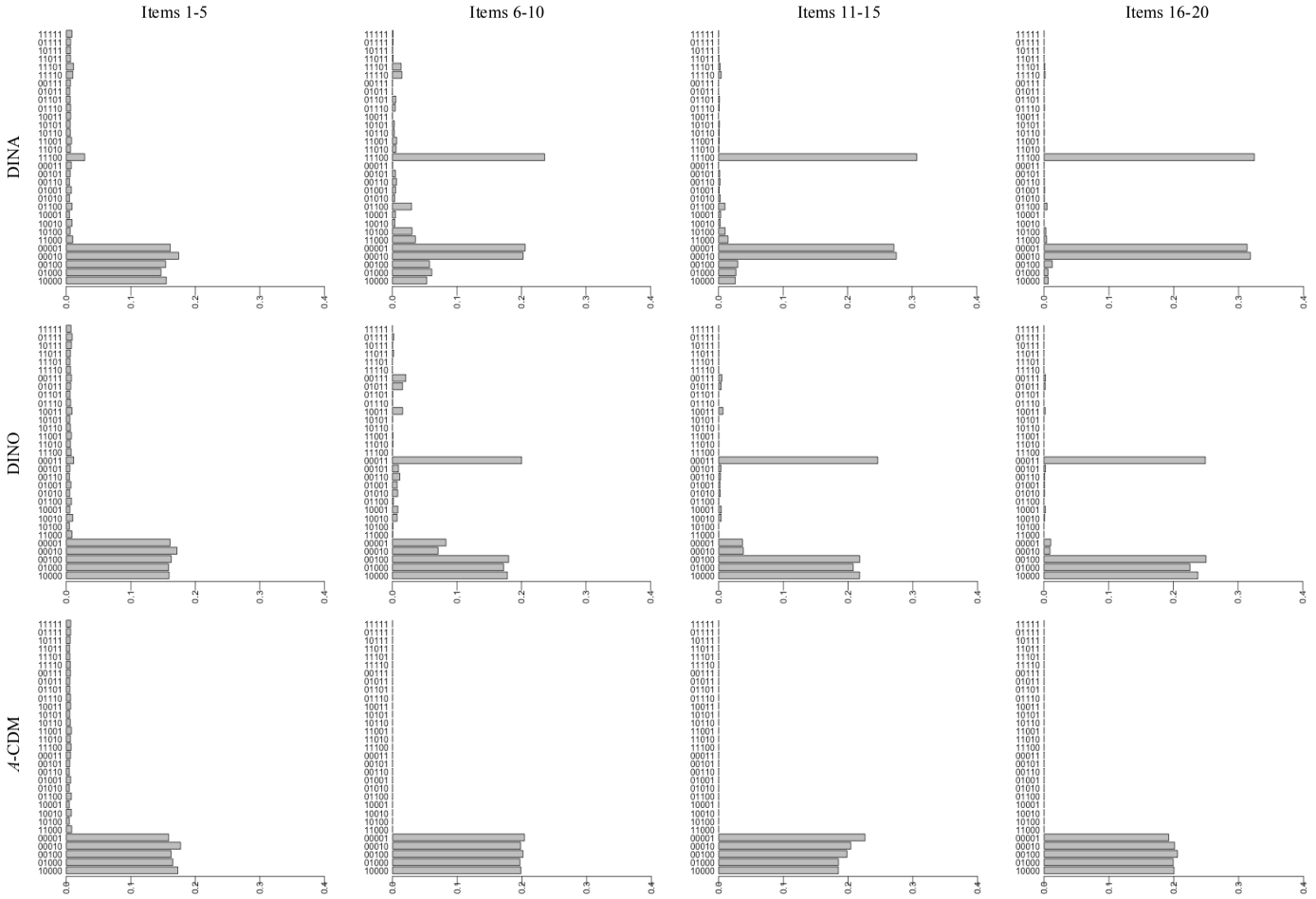
Third, each data set was generated using a single CDM in this study. However, as with previous indices, the MPWKL and the GDI are sufficiently general that it can simultaneously be applied to any CDMs subsumed by the G-DINA model. As such, it would be interesting to examine how the new indices will perform when the item pool is made up of various CDMs, which reflects what can be expected in practice − different items might require different processes (i.e., CDMs). Finally, to keep the scope of this work manageable, a few simplifications about factors affecting the performance of CD-CAT indices were made. These include fixing the number of attributes, using a single method in estimating the attribute vectors, and assuming that the item parameters were known. To obtain more generalizable conclusions, future research should consider varying these factors.

Figure 2.3: Overall Proportion of Item Usage for $\boldsymbol{\alpha}_3$, GDI, and $J = 20$



Note: GDI = G-DINA model discrimination index; G-DINA = generalized DINA; DINA = deterministic inputs, noisy "and" gate; J = test length; DINO = deterministic input, noisy "or" gate; A-CDM = additive CDM; CDM = cognitive diagnosis model.

Figure 2.4: The Proportion of Item Usage in Different Periods for $\boldsymbol{\alpha}_3$, GDI, and $J = 20$



Note: GDI = G-DINA model discrimination index; G-DINA = generalized DINA; DINA = deterministic inputs, noisy "and" gate; J = test length; DINO = deterministic input, noisy "or" gate; A-CDM = additive CDM; CDM = cognitive diagnosis model.

References

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619-632.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* New York, NY: John Wiley.

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

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

de la Torre, J., & Chiu, C.-Y. (2010, April). *General empirical method of Q-Matrix validation.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

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

Doornik, J. A. (2011). Object-oriented matrix programming using Ox (Version 6.21). [Computer software]. London: Timberlake Consultants Press.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.

Hsu, C.-L., Wang, W.-C., & Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement, 37*, 563-582.

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

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York: Springer.

McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40*, 808-821.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*, 187-194.

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

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

van der Linden, W. J., & Glas, C. A. W. (2002). Preface. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. Vii-Xii). Boston, MA: Kluwer.

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement, 73,* 1017-1035.

Wang, C., Chang, H.-H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement, 48*, 255-273.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Xu, X., Chang, H.-H., & Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Chapter 3
Study II: Item Exposure Control for CD-CAT

Abstract

This article examines the use of two item exposure control methods, namely, the restrictive progressive and restrictive threshold, in conjunction with the generalized deterministic inputs, noisy "and" gate model discrimination index (GDI) as item selection methods in cognitive diagnosis computerized adaptive testing. The efficiency of the methods is compared with the GDI using a simulation study. The impact of different factors, namely, item quality, generating model, attribute distribution, item pool size, sample size, and prespecified desired exposure rate, on classification accuracy and item exposure rates is also investigated. The results show that the GDI performed efficiently with the exposure control methods in terms of classification accuracy and item exposure. In addition, the impact of the factors on item exposure rates vary based on the methods.

Keywords: cognitive diagnosis model, computerized adaptive testing, item exposure control

3.1 Introduction

Although computerized adaptive testing (CAT) has become a popular tool in educational and psychological testing, several researchers noted that it has been criticized

for security and item bank usage problems for several years. Security problems are related to overuse of items (i.e., overexposed items; Chen, Ankenmann, & Spray, 2003), and item bank usage problems are related to the use of rarely selected items (i.e., underexposed items; Barrada, Veldkamp, & Olea, 2009). For example, in security problems, test-takers can memorize the items and distribute them publicly (H.-H. Chang, 2004). Moreover, because item selection methods tend to select items that provide the most information about ability level, the indices choose some items (e.g., high-discriminating items) more often than others (e.g., low-discriminating items). Thus, overuse of items can lead to information sharing among the examinees, and can result in items being answered correctly regardless of the examinees' ability levels (Lee, Ip, & Fuh, 2007). Also, the reliability and the validity of the test become questionable. Therefore, high item exposure rates must be controlled to lessen the impact of item sharing.

Item exposure rates can be affected by the psychometric properties of the items, the items available in the item pool, and the ability distribution of the examinees (Revuelta & Ponsoda, 1998). Two points should be considered when item exposure is controlled: preventing overexposure of some items and increasing the use of rarely selected items. A series of studies proposed methods for item exposure control in traditional CAT. Sympson and Hetter (1985) proposed an iterative procedure for controlling item exposure. In that study, an item exposure parameter, the probability of administering an item that had already been selected, was assigned to each item. If the parameter of a particular item was as low as the prespecified desired exposure rate, the item could not be administered when it was selected. However, the main drawbacks of this method involved time-consuming iterations in calculating item exposure parameters and not being able to maintain the exposure rates of all items at or below the prespecified desired exposure rate (Barrada, Abad, & Veldkamp, 2009).

Later, Davey and Parshall (1995) proposed a method that aimed to minimize a

set of items that appear together. Again, an exposure parameter was assigned to each item, but the parameter was conditioned on the administered items for a particular examinee. In a set of studies, Stocking and Lewis (1995a, 1995b) proposed two methods for controlling item exposure. The two methods, the unconditional multinomial and the conditional multinomial procedures, modified the method proposed by Sympson and Hetter and employed a multinomial model for selecting items instead of using optimal selection. S.-W. Chang and Twu (1998) conducted a study to compare these item exposure control methods. The researchers found that the Sympson and Hetter and the Stocking and Lewis unconditional methods yielded very similar results under all conditions. In addition, the Stocking and Lewis conditional method produced better results for the exposure rates; however, it also produced the highest measurement error.

The other item exposure control methods can be grouped as stratified methods. H.-H. Chang and Ying (1996) noted that selecting items based on information might be less efficient at the early stages of CAT because of the poor interim ability estimation. Moreover, item selection based on information could lead to highly skewed item usage. For example, selecting items based on the maximum Fisher information at the early stages of CAT resulted in the overuse of items with high discrimination parameters (H.-H. Chang & Ying, 1999); however, those items might not discriminate test takers well especially when the estimated ability level was not stable. To handle this issue, H.-H. Chang and Ying (1999) proposed a multistage adaptive testing approach, namely, the α-stratified strategy, for item exposure control in which the item bank was first divided into parts (i.e., strata) based on the discrimination parameter. Then, at the early stages of the test, items were selected from a stratum that had items with low discrimination. As the test progressed and the ability estimate became more stable, items with high discrimination were selected according to an optimization criterion (Georgiadou, Triantafillou, & Economides, 2007).

However, one issue with the α-stratified approach occurred when the discrimination parameters were highly correlated with the difficulty parameters. To solve that problem, H.-H. Chang, Qian, and Ying (2001) proposed another method, namely, the α-stratified strategy with b-blocking. In this method, the item bank was first divided into blocks based on the difficulty parameters in ascending order, and then each block was divided into strata based on the discrimination parameters. In this method, the discrimination parameters were distributed more evenly within each stratum, and the average discrimination increased across the strata (Georgiadou et al., 2007). However, these stratification methods did not control individual item exposure rates, and thus, some items might exceed the prespecified desired exposure rate unless other item exposure control methods (e.g., the Sympson-Hetter method) are implemented with the item selection method (Deng, Ansley, & H.-H. Chang, 2010). Different versions of the stratification methods (e.g., α-stratified strategy with content blocking, α-stratified with unequal item exposure across strata, multidimensional stratification) have been proposed in the literature, and a comprehensive literature review can be found in Georgiadou et al. (2007).

There are, however, some drawbacks with the stratified methods (Han, 2012). First, item stratification can cause a problem in limiting the number of available items in each stratum, and can result in the overuse of certain items. Second, the impact of the guessing parameter is generally ignored in those methods, and it can threaten the quality of the test administration, especially if this parameter is correlated with the difficulty and discrimination parameters (Barrada, Mazuela, & Olea, 2006). Last, the stratification methods are not effective when variable-test lengths are used (Han, 2012; Wen, H. Chang, & Hau, 2000).

In another study, Revuelta and Ponsoda (1998) proposed an item exposure control method, the restricted method, in which none of the items was allowed to be exposed for more than a prespecified desired exposure rate. This method simply assigns either

zero or one as an item exposure parameter for each item. This parameter is zero if the exposure rate of an item is greater than or equal to the prespecified desired exposure rate; otherwise, it is one (Barrada, Abad, & Veldkamp, 2009).

Another method, the progressive method, based on the maximum information was originally proposed by Revuelta (1995), and Revuelta and Ponsoda (1996). The method has two components, item information and random components. Revuelta and Ponsoda (1998) defined the method as follows: Let x be the number of administered items, and L be the number of total items in the test. Also, define a random value (H_j) between zero and the highest information value to be drawn from uniform distribution. A weight is computed considering a linear combination of random and information components as

$$\omega_j = \left(1 - \frac{x}{L}\right) H_j + \frac{x}{L} I_j, \tag{3.1}$$

where I_j is the information for item j. The impact of the random component on the item selection index is reduced, and the importance of the information increases as the test progresses.

Later, Barrada, Olea, Ponsoda, and Abad (2008) proposed two functions that can be applied to various item exposure control methods, including the progressive method. Those functions aimed to control the speed of the move from random selection to selection based on information by acceleration parameters. In other words, item selection can be mainly random at the beginning of the test, and then gradually the information part becomes more important as the test progresses. Moreover, the speed of this switch can be controlled by the functions. The researchers found that the modified methods were efficient for improving the item exposure control methods with very small losses in measurement accuracy. The idea of random selection at the beginning of the test has also been supported by other studies that noted the

random selection of items at the beginning of the test caused a very small decrease in the measurement accuracy (Barrada et al., 2008; Li & Schafer, 2005; Revuelta & Ponsoda, 1998).

Wang, Chang, and Huebner (2011) proposed a modification in which a stochastic component was added to the item selection criterion in the progressive method. In doing so, the item selection indices did not always pick the items with the most information. The restrictive progressive (RP) method using an information index is given by

$$RP\text{-}I_j = \left(1 - \frac{exp_j}{r}\right)\left[\left(1 - \frac{x}{L}\right)R_j + \frac{I_j \beta x}{L}\right], \tag{3.2}$$

where exp_j is the preliminary exposure rate, r is the prespecified desired exposure rate, x is the number of items administered, L is the test length, $R_j \sim Uniform(0, H^*)$ in which $H^* = max(I_j)$ for the remaining items in the item pool, $\beta > 0$ is a weight that will be described later, and I_j is the information index. As the items in the pool are administered, the role of the information part increases, whereas the role of the stochastic component, $1 - x/L$, decreases. The β value is an arbitrary number to give priority to test security or estimation accuracy. Small β values provide better test security, whereas high values result in better estimation accuracy. A restriction value r was also added to the model to control the maximum exposure rate. In their simulation study, the RP method was successful in controlling high exposure rates.

In addition, Wang et al. (2011) proposed the restrictive threshold (RT) method, which also has two components, restrictive and threshold components. The threshold component creates a set of items whose information is close to the largest information. Specifically, the interval for the threshold component is defined as

$$[max(I_j) - \delta, max(I_j)], \tag{3.3}$$

where δ specifies the length of the interval, and it is defined as

$$\delta = [max(I_j) - min(I_j)] \times f(x), \tag{3.4}$$

where x is the number of items administered on the test, and $f(x)$ is a monotone decreasing function. In this study, this function is defined as $f(x) = (1 - x/L)^{\beta}$ where β balances test security and estimation accuracy. The RT method was applied deterministically such that when an item's exposure rate reached the prespecified desired exposure rate, the item was removed from the item pool for the next examinees. In this article, the RP and RT methods were used with an item selection index to control the item exposure rates.

3.1.1 Cognitive Diagnosis Models

In cognitive diagnosis, a binary attribute vector typically represents the presence or absence of the specific skills or attributes in a particular content area. To achieve this, let $\boldsymbol{\alpha}_i$=$\{\alpha_{ik}\}$ be the attribute vector of examinee i, where $i = 1, 2, \ldots, N$ examinees, and $k = 1, 2, \ldots, K$ attributes. The kth element of the vector is 1 when the examinee has mastered the kth attribute, and it is 0 when the examinee has not. Similarly, the responses of the examinees to J items are represented by a binary vector, $\boldsymbol{X_i} = \{x_{ij}\}$, where x_{ij} is the ith examinee's binary response for the jth item, and $j = 1, 2, \ldots, J$. A Q-matrix (Tatsuoka, 1983), which is a $J \times K$ matrix, represents the required attributes for an item and the element of the jth row and the kth column, q_{jk}, is 1 if the kth attribute is required to answer the jth item correctly, and it is 0 otherwise.

To date, several constrained and general CDMs have been proposed in the literature. On one hand, the constrained models require specific assumptions about the relationship between attribute vector and task performance (Junker & Sijtsma,

2001). Nonetheless, they provide results that can easily be interpreted. On the other hand, the general models relax some of the strong assumptions in the constrained models, and provide more flexible parameterizations. However, general models are more difficult to interpret compared to the constrained models because they involve more complex parametrizations. One of the general models was proposed by de la Torre (2011), and it is called the generalized deterministic inputs, noisy "and" gate (G-DINA) model. A few of commonly encountered constrained CDMs can be subsumed by the G-DINA model. These are the deterministic inputs, noisy "and" gate (DINA; de la Torre, 2009; Haertel, 1989; Junker & Sijtsma, 2001) model, which assumes that lacking at least one of the required attributes is as the same as lacking all of the required attributes; the deterministic input, noisy "or" gate (DINO; Templin & Henson, 2006) model, which assumes that having at least one of the required attributes is as the same as having all of the required attributes; and the additive CDM (*A*-CDM; de la Torre, 2011), which assumes that the impacts of mastering the different required attributes are independent of each other. In Wang et al. (2011) paper, they examined the Fusion model (Hartz, 2002; Hartz, Roussos, & Stout, 2002) with the RP and RT methods; in this article, three constrained models, namely, DINA, DINO, and *A*-CDM, were used in the data generation, but the CAT administration was carried out under the G-DINA model context.

3.1.2 Computerized Adaptive Testing

CAT has become a popular tool in educational testing over the past few decades. It has been developed as an alternative to paper-and-pencil tests, and offers faster scoring and more flexible testing schedules for individuals. In addition, CAT provides shorter test-lengths and enhanced measurement precision compared to paper-and-pencil tests (Meijer & Nering, 1999). The components of CAT can be listed as calibrated item pool, starting point, item selection method, scoring procedure, and

stopping rule for the test administration (Weiss & Kingsbury, 1984). In this article, we focused on the item exposure rates that used in conjunction with efficient item selection methods.

Item selection methods based on the Fisher information are widely used in traditional CAT (Lord, 1980; Thissen & Mislevy, 2000). However, those methods are not applicable in cognitive diagnosis computerized adaptive testing (CD-CAT) because they generally work with only continuous ability levels, whereas the equivalent latent variables in cognitive diagnosis are discrete. Alternatively, the item selection methods based on the Kullback-Leibler (K-L) information can be used in CD-CAT. Xu, Chang, and Douglas (2003) first noted the issue and proposed two item selection indices based on the Kullback-Leibler (K-L) information and Shannon entropy procedure in CD-CAT, and the results showed that both indices outperformed random selection in terms of attribute classification. Later, Cheng (2009) proposed two item selection indices, namely, the posterior-weighted K-L index (PWKL) and hybrid K-L index (HKL), for CD-CAT. The calculation of the PWKL involves summing the distances between the current estimate of the attribute vector and the other possible attribute vectors weighted by the posterior distribution of attribute vectors. The results of her simulation study showed that the new indices performed similarly, and had higher classification rates compared to the K-L and Shannon entropy procedure. In another study, Kaplan, de la Torre, and Barrada (2015) proposed two new item selection indices for CD-CAT. One of them is based on the G-DINA model discrimination index (GDI), and the other one is based on the PWKL, which is called modified PWKL. The results showed that the two new indices performed very similarly and higher attribute classification rates compared to the PWKL. In addition, the GDI had the shortest administration time. In this article, the GDI was used as an item selection index with the two item exposure control methods, namely, RP and RT methods.

The GDI was originally proposed as an index to implement an empirical Q-matrix validation procedure (de la Torre & Chiu, 2015). It measures the (weighted) variance of the probabilities of success of an item given a particular attribute distribution. Later, Kaplan et al. (2015) used the index as an item selection method, and their results showed that the index was promising in CD-CAT. To give a definition of the index, let the first K_j^* attributes be required for item j, and define $\boldsymbol{\alpha}_{cj}^*$ as the reduced attribute vector consisting of the first K_j^* attributes, for $c = 1, \ldots, 2^{K_j^*}$. For example, if a q-vector is defined as (1,1,0,0,1) for $K_j^* = 3$ number of required attributes, the reduced attribute vector is (a_1,a_2,a_5). Also, define $P(X_{ij} = 1|\boldsymbol{\alpha}_{cj}^*)$ as the success probability on item j given $\boldsymbol{\alpha}_{cj}^*$. The GDI for item j is defined as

$$\varsigma_j^2 = \sum_{c=1}^{2^{K_j^*}} \pi_i^{(t)}(\boldsymbol{\alpha}_{cj}^*)[P(X_{ij} = 1|\boldsymbol{\alpha}_{cj}^*) - \bar{P}_j]^2, \tag{3.5}$$

where $\pi_i^{(t)}(\boldsymbol{\alpha}_{cj}^*)$ is the posterior probability of the reduced attribute vector and $\bar{P}_j = \sum_{c=1}^{2^{K_j^*}} \pi(\boldsymbol{\alpha}_{cj}^*)P(X_{ij} = 1|\boldsymbol{\alpha}_{cj}^*)$ is the mean success probability.

3.2 Simulation Study

The goal of this study is to investigate the efficiency of the GDI in conjunction with the RP and RT methods in terms of the item exposure rate and estimation accuracy, and also simultaneously reduce the use of overexposed items and/or increase the use of underexposed items. The design of the simulation study consisted of investigating the impact of different factors. These factors included two levels of item quality, three reduced CDMs, two attribute distributions, two sample sizes for the data generation, and two test lengths for the test termination rule. In addition, two item pool sizes, three item selection indices, including two different prespecified desired exposure rates, r^{max} values, and three different β values were used for the

item exposure control methods in the CAT administration.

3.2.1 Design

3.2.1.1 Data Generation

In the data generation, the impact of the item quality and generating model was considered. First, because recent research has shown item quality affects the accuracy of the attribute classification (e.g., de la Torre, Hong, & Deng, 2010; Kaplan et al., 2015), different item discriminations and variances were used to generate the data. Two levels of item quality, namely, low-quality (LQ) and high-quality (HQ), were considered. However, it should be noted that these two terms were used exclusively for this study, and in other studies, they have been defined differently. For the purposes of this study, HQ and LQ can also be viewed as more discriminating and less discriminating, respectively. For LQ items, the lowest and highest success probabilities (i.e., $P(\mathbf{0})$ and $P(\mathbf{1})$) were generated from uniform distributions, $U(0.15, 0.25)$ and $U(0.75, 0.85)$, respectively; and for HQ items, $P(\mathbf{0})$ and $P(\mathbf{1})$ were generated from uniform distributions, $U(0.00, 0.20)$ and $U(0.80, 1.00)$, respectively. Second, item responses were generated using three reduced models: DINA model, DINO model, and A-CDM. For the DINA and DINO models, the probabilities of success were set as discussed above. In addition to these probabilities, the intermediate success probabilities were obtained by allowing each of the required attributes to contribute equally in the A-CDM. The number of attributes was fixed at K=5.

Third, the impact of attribute distribution on the efficiency of the indices was also investigated. Using different attribute distributions allows greater generalizability of the findings from the study. Two different distributions, uniform and higher-order (HO) distributions, were used to generate the examinees' attribute vectors. In the former, the examinee attribute vectors were drawn from 2^K possible attribute vectors

uniformly, whereas in the latter, the attribute vectors were drawn considering a HO latent trait. HO latent traits were introduced for cognitive diagnosis by de la Torre and Douglas (2004). In their study, a method for modeling the joint distribution of the attribute vectors based on HO specification was proposed. By positing an HO variable θ, the difficulty of mastering a specific set of attributes can be parameterized. The probability of $\boldsymbol{\alpha}$ conditional on θ can be written as

$$p(\boldsymbol{\alpha}|\theta) = \prod_{k=1}^{K} p(\alpha_k|\theta). \tag{3.6}$$

The particular model considered in the current paper expresses the logit in 3.6 as a linear function of θ, as in:

$$p(\alpha_k|\theta) = \frac{exp(\lambda_{0k} + \lambda_{1k}\theta)}{1 + exp(\lambda_{0k} + \lambda_{1k}\theta)},$$

where λ_{0k} is the difficulty parameter, λ_{1k} is the discrimination parameter, and θ is a latent continuous variable to account for the associations among the attributes. In this study, the HO parameters, difficulty and discrimination, were fixed to $\lambda_{0k} = \{-2, -1, 0, 1, 2\}$ and $\lambda_1 = 1$ across all conditions, respectively. The HO latent trait θ was drawn from a standard normal distribution. Last, the impact of the sample size on the efficiency of the indices was investigated. Two sample sizes were considered, N=500 and 1000.

3.2.1.2 Item Pool and Item Selection Methods

The impact of item pool size on the item exposure rates was investigated in this study. The Q-matrix was created from $2^K - 1 = 31$ possible q-vectors for two sizes: each with 20 and 40 items, resulting in 620 and 1240 items in the pool, respectively. For the test termination rule, two fixed-test lengths, 10 and 20, were considered. Two prespecified desired exposure rates, r^{max} of .1 and .2, were used. Three β

values in the RP and RT methods were considered to balance the exposure rate and the classification rates: β=0.5, 1.0, and 2.0. Three item selection indices were considered: the GDI, RP-GDI, and RT-GDI. For greater comparability, the first item was chosen randomly from the pool for each examinee, and this item was fixed across the indices. In addition, ten replications were performed for the RP-GDI to get more stable results.

To compare the efficiency of the indices in terms of the estimation accuracy, the correct attribute classification (CAC) rates and the correct attribute vector classification (CVC) rates were calculated. The CAC and CVC rates were computed as

$$\begin{aligned} CAC &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{5} I[\alpha_{ik} = \hat{\alpha}_{ik}], \text{ and} \\ CVC &= \frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{5} I[\alpha_{ik} = \hat{\alpha}_{ik}], \end{aligned} \tag{3.7}$$

where I is the indicator function. Different statistics (e.g., chi-square statistic and overlap rates) have been proposed to evaluate the item exposure rates associated with different indices. In this study, a chi-square statistic (H.-H. Chang & Ying, 1999) and the maximum of the item exposure rates for each condition were calculated. The statistic was calculated as

$$\chi^2 = \sum_{i=1}^{J}(r_i - \bar{r})^2/\bar{r}, \tag{3.8}$$

where r_j is the exposure rate for item j, and $\bar{r}$ is the overall mean exposure rate for the entire test. Smaller values indicate more even exposure rates. These statistics were calculated before and after the exposure rates were controlled to investigate which of the methods works better with the GDI in terms of the estimation accuracy and the item exposure rate.

3.2.2 Results

Due to space limitation, only partial results (see Table 3.1) are given; however, the results in their entirety can be requested from the first author. Several results can be noted. First, as expected, the CAC rates were higher than the CVC rates, but the measures showed similar patterns for all conditions. In addition, the GDI resulted in the highest, the RP-GDI had the second highest, and the RT-GDI had the lowest CAC and CVC rates across all conditions. Last, the GDI yielded the highest, the RT-GDI yielded the second highest, and the RP-GDI yielded the lowest maximum and chi-square value of the item exposure rates for all conditions. Only the CVC rates for the classification accuracy and only the chi-square values for the item exposure rate were discussed in the following sections. In addition, the impact of the factors on the attribute classification and item exposure rates were discussed in detail.

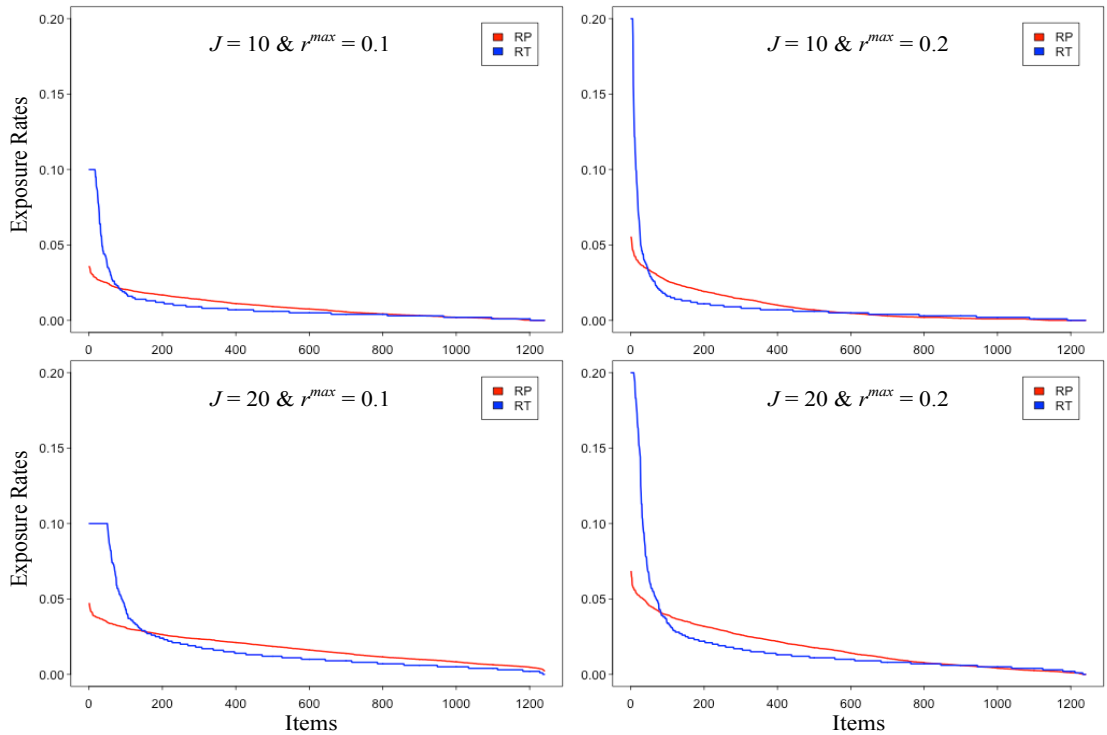
Table 3.1: The CVC rates, and the Maximum and the Chi-Square Values of Item Exposure Rates Using the DINA, 10-Item Test, $\beta = 0.5$, and $r^{max} = 0.1$

				GDI			RP-GDI			RT-GDI		
IQ	J	N	AD	CVC	Max	χ^2	CVC	Max	χ^2	CVC	Max	χ^2
LQ	620	500	U	0.55	0.95	269.88	0.44	0.03	1.04	0.40	0.10	8.02
			HO	0.61	0.91	251.58	0.44	0.03	1.05	0.45	0.10	7.29
		1000	U	0.56	0.96	265.60	0.43	0.03	0.93	0.40	0.10	7.48
			HO	0.58	0.90	255.09	0.44	0.03	0.95	0.43	0.10	6.69
	1240	500	U	0.55	0.96	608.93	0.47	0.02	2.83	0.40	0.10	13.12
			HO	0.61	0.94	571.14	0.49	0.03	2.94	0.43	0.10	13.54
		1000	U	0.60	0.95	590.22	0.47	0.02	2.70	0.43	0.10	11.19
			HO	0.61	0.95	560.46	0.48	0.03	2.72	0.47	0.10	11.99
HQ	620	500	U	0.87	0.93	251.03	0.74	0.03	0.77	0.68	0.10	7.46
			HO	0.92	0.86	235.40	0.72	0.03	0.96	0.71	0.10	6.85
		1000	U	0.87	0.93	251.87	0.76	0.03	0.64	0.74	0.10	6.50
			HO	0.90	0.84	230.22	0.72	0.03	0.89	0.73	0.10	6.59
	1240	500	U	0.86	0.94	523.39	0.78	0.02	1.96	0.71	0.10	12.63
			HO	0.87	0.92	494.70	0.77	0.02	2.43	0.68	0.10	13.46
		1000	U	0.89	0.94	512.25	0.78	0.02	1.78	0.71	0.10	10.74
			HO	0.89	0.91	473.53	0.76	0.02	2.25	0.71	0.10	11.85

Note. CVC = correct attribute vector classification; DINA = deterministic inputs, noisy "and" gate; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; RP-GDI = restrictive progressive GDI; RT-GDI = restrictive threshold GDI; IQ = item quality; J = pool size; N = sample size; AD = attribute distribution; LQ = low-quality; HQ high-quality; U = uniform; HO = higher-order.

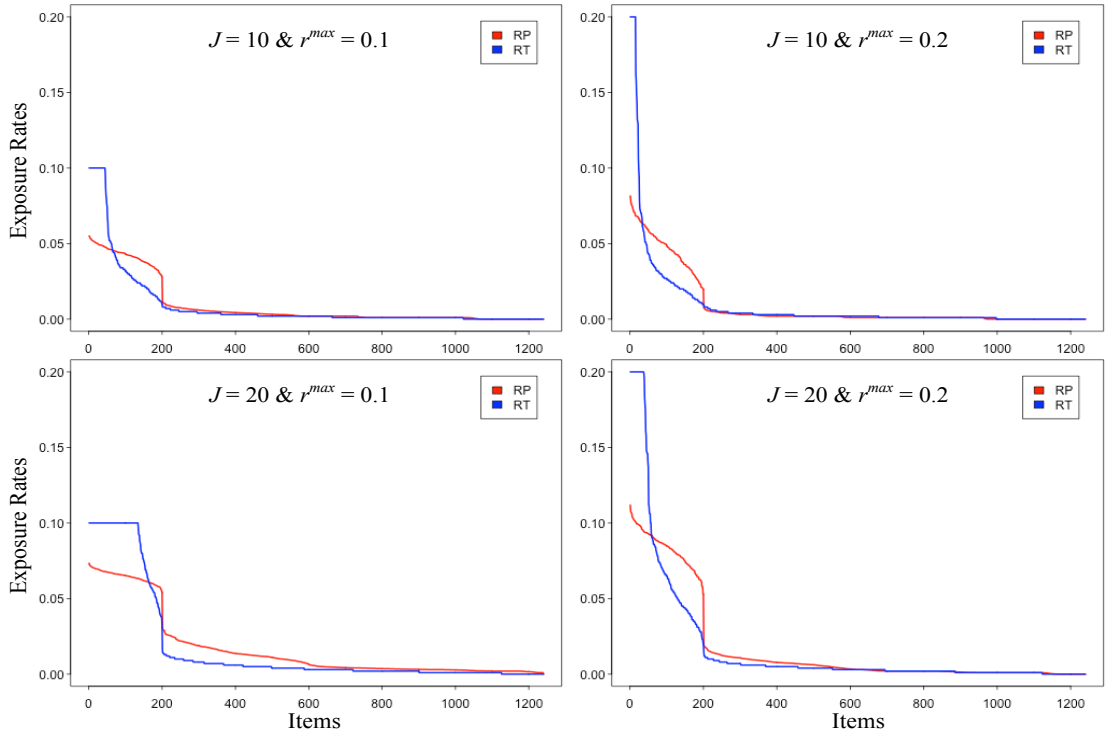
To gain a better understanding of how different exposure control methods behaved in different conditions, the item exposure rates are shown in Figures 3.1 and 3.2 for the 10-item test with the RP-GDI and RT-GDI using the DINA model and the *A*-CDM, respectively. Several conclusions can be gleaned from the figures. First, the RP method resulted in more uniform item exposure rates because of its probabilistic nature, and the RT method yielded more skewed rates because it was implemented deterministically. Second, the maximum exposure rates were always lower than the desired r^{max} value when the RP method was used, whereas the maximum exposure rates were equal to the desired r^{max} when the RT method was used. Third, more items reached the desired r^{max} value using the *A*-CDM compared to the DINA and DINO models, and those items were mostly one-attribute items. Last, using the *A*-CDM resulted in more skewed item exposure rates compared to the DINA (or DINO) model.

Figure 3.1: Item Exposure Rates for the DINA model



Note: Red and blue lines represent the RP and RT, respectively; RP = restrictive progressive; RT = restrictive threshold; DINA = deterministic inputs, noisy "and" gate; J = test length.

Figure 3.2: Item Exposure Rates for the *A*-CDM



Note: Red and blue lines represent the RP and RT, respectively; RP = restrictive progressive; RT = restrictive threshold; *A*-CDM = additive CDM; CDM = cognitive diagnosis model; J = test length.

In general, there is a trade-off between estimation accuracy and item exposure rate (Way, 1998). In other words, reducing high item exposure rates will result in lower classification rates, and vice versa. To better examine the impact of the different factors, differences in the CVC rates were evaluated using a cut point of 0.05. Differences below 0.05 were considered negligible, whereas differences above 0.05 were considered substantial. In addition, the chi-square statistic ratios were calculated to compare the efficiency of the indices under different factors, and the ratios were evaluated using two cut points, 0.15 and 0.25. If the ratio was equal to one, then the two chi-square values were considered equal to each other. If the ratio was within the range of (0.85,1.15), it was considered negligible; within (0.75,0.85) or (1.15,1.25), it was considered moderate; otherwise, it was considered substantial.

3.2.2.1 The Impact of the Item Quality

As expected, using HQ items instead of LQ items resulted in higher classification rates across different factors (e.g., generating model, item selection index). Moreover, the increases in the CVC rates were greater when short tests (i.e., 10-item tests) were used compared to long tests (i.e., 20-item tests). For example, the increases were around 0.30 and 0.10 on average for the 10- and 20-item tests, respectively.

However, the impact of the item quality on the item exposure rates varied based on the other factors. The chi-square ratios using the RP-GDI and RT-GDI are shown in Table 3.2 for the DINA and DINO models. Several results can be noted. First, for the GDI with the DINA and DINO models, the use of LQ items instead of HQ items resulted in negligible differences in the chi-square values regardless of the other factors except for short tests with a large pool in the DINA model, and short tests with a large pool and the uniform distribution in the DINO model, where the differences were moderate. Second, for the RT-GDI with the DINA and DINO models, the use of LQ items instead of HQ items generally resulted in negligible to moderate differences in the chi-square values regardless of the other factors except for some conditions. For example, the differences were substantial when a small pool was used with the HO distribution, a small β, and an r^{max} of .2 regardless of the test length and sample size. Third, for the RP-GDI with the DINA and DINO models, the use of LQ items mostly yielded larger chi-square values than HQ items, and there were more cases where the differences in chi-square values were substantial compared to the RT-GDI. However, there were some exceptions. For example, the differences were negligible to moderate when the HO distribution was used with a small β regardless of the pool size, sample size, test length, and r^{max} value. Last, for the A-CDM, the use of LQ items instead of HQ items resulted in negligible differences in the chi-square values across all the conditions.

Table 3.2: The Chi-Square Ratios Comparing LQ vs. HQ

				DINA								DINO							
				RP-GDI				RT-GDI				RP-GDI				RT-GDI			
				Test Length															
				10		20		10		20		10		20		10		20	
				r^{max}															
J	N	AD	β	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
620	500	U	0.5	**1.36**	**1.46**	**1.45**	**1.63**	1.07	1.21	1.21	1.20	**1.35**	**1.44**	**1.53**	**1.64**	1.09	1.20	1.14	1.08
			2	**1.59**	**1.42**	**2.00**	**1.74**	**1.28**	**1.31**	1.19	**1.31**	**1.59**	**1.42**	**2.05**	**1.76**	1.24	**1.32**	1.25	**1.27**
		HO	0.5	1.09	1.21	0.81	0.99	1.06	1.12	0.98	1.02	1.09	1.23	0.87	1.05	1.03	1.01	0.95	0.94
			2	**1.32**	**1.33**	1.15	**1.32**	1.16	1.23	1.14	1.15	**1.35**	**1.34**	1.19	**1.34**	1.17	1.23	1.14	1.16
	1000	U	0.5	**1.45**	**1.47**	**1.54**	**1.75**	1.15	1.20	1.15	1.12	**1.45**	**1.49**	**1.59**	**1.76**	1.12	1.12	1.20	1.14
			2	**1.63**	**1.45**	**2.13**	**1.75**	1.22	**1.34**	1.21	**1.27**	**1.65**	**1.44**	**2.11**	**1.76**	1.24	**1.33**	1.21	**1.29**
		HO	0.5	1.07	1.20	0.81	1.01	1.01	1.18	0.98	0.99	1.08	1.24	0.82	1.06	1.05	1.02	0.95	0.95
			2	**1.35**	**1.33**	1.17	**1.32**	1.16	1.25	1.16	1.17	**1.37**	**1.33**	1.20	**1.35**	1.18	1.24	1.15	1.15
1240	500	U	0.5	**1.45**	**1.44**	**1.67**	**1.67**	1.04	1.24	1.21	1.22	**1.48**	**1.44**	**1.71**	**1.70**	1.10	1.10	1.12	1.12
			2	**1.47**	**1.38**	**1.73**	**1.48**	1.18	1.20	1.22	1.23	**1.45**	**1.37**	**1.74**	**1.47**	1.19	**1.27**	1.24	1.23
		HO	0.5	1.21	**1.28**	0.99	1.22	1.01	0.96	1.00	1.06	1.21	**1.26**	1.05	1.24	0.98	1.02	0.98	1.06
			2	**1.33**	**1.30**	**1.33**	**1.32**	1.16	1.23	1.16	1.14	**1.34**	**1.30**	**1.32**	**1.32**	1.19	1.14	1.16	1.17
	1000	U	0.5	**1.52**	**1.46**	**1.78**	**1.69**	1.04	1.03	1.05	1.08	**1.49**	**1.45**	**1.74**	**1.70**	1.10	1.15	1.15	1.17
			2	**1.46**	**1.38**	**1.75**	**1.48**	1.20	1.24	1.22	1.25	**1.45**	**1.37**	**1.74**	**1.48**	1.19	1.25	1.24	1.24
		HO	0.5	1.21	**1.30**	1.03	1.21	1.01	1.02	1.02	1.08	1.23	**1.30**	1.02	1.22	1.02	0.98	1.07	1.05
			2	**1.34**	**1.31**	**1.34**	**1.32**	1.15	1.17	1.16	1.15	**1.35**	**1.32**	**1.34**	**1.33**	1.19	1.13	1.18	1.18

Note. Substantial differences are shown in bold. LQ = low-quality; HQ = high-quality; DINA = deterministic inputs, noisy "and" gate; DINO = deterministic input, noisy "or" gate; RP-GDI = restrictive progressive GDI; RT-GDI = restrictive threshold GDI; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; J = pool size; N = sample size; AD = attribute distribution; U = uniform; HO = higher-order.

3.2.2.2 The Impact of the Sample Size

Increasing the sample size resulted in negligible differences in the classification rates regardless of the other factors (e.g., item selection index, generating model, and item quality) except for some conditions using the RT-GDI with the DINA model, where the differences were substantial. For example, a large sample (i.e., N=1000) yielded higher classification rates compared to a small sample (i.e., N=500) when the uniform distribution and a small β, and the HO distribution and a large β were used with short tests, HQ items, a small pool, and an r^{max} of .1.

Similarly, the impact of the sample size on the item exposure rates were negligible across the different factors, based on the chi-square ratios shown in Table 3.3. However, there were some conditions where the differences in the chi-square values were either moderate or substantial. For example, the differences were moderate for the RP-GDI when the DINA, short tests and HQ items were used with an r^{max} of .1, a small β, a small pool, and the uniform distribution; and when the DINO, long tests were used with an r^{max} of .1, a small β, a small pool, and the uniform distribution regardless of the item quality; for the RT-GDI when the DINA, LQ items were used with a small β, a large pool, and the uniform distribution regardless of the test length and the r^{max}; and when the DINO, short tests and HQ items were used with an r^{max} of .1, and a small β, a large pool, and the HO distribution. In addition, the differences were substantial for the DINA and DINO models when the RP-GDI, long tests, and HQ items were used with an r^{max} of .1, a small β, a small pool, and the uniform distribution; and when the RT-GDI, long tests, and HQ items were used with an r^{max} of .1, a small β, a large pool, and the HO distribution.

Table 3.3: The Chi-Square Ratios Comparing Small vs. Large Sample Size

				DINA								DINO							
				RP-GDI				RT-GDI				RP-GDI				RT-GDI			
				Test Length															
				10		20		10		20		10		20		10		20	
				r^{max}															
IQ	J	AD	β	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
LQ	620	U	0.5	1.12	1.10	1.21	1.14	1.07	1.01	1.13	1.08	1.09	1.05	1.17	1.07	1.11	1.08	1.16	1.10
			2	1.07	1.02	1.10	1.06	1.04	1.00	1.03	1.03	1.03	1.02	1.08	1.03	1.02	0.99	1.06	1.01
		HO	0.5	1.10	1.07	1.11	1.06	1.09	1.02	1.03	1.02	1.06	1.04	1.10	1.02	1.07	1.03	1.02	1.01
			2	1.03	1.03	1.04	1.03	1.01	1.01	1.00	1.00	1.00	1.02	1.00	1.01	1.03	0.99	1.02	0.98
	1240	U	0.5	1.05	1.03	1.11	1.05	1.17	1.23	1.24	1.20	1.09	1.06	1.10	1.03	1.10	0.97	1.07	1.06
			2	1.02	1.01	1.02	1.01	1.04	1.03	1.05	1.03	1.02	1.01	1.02	1.00	1.04	1.02	1.01	1.01
		HO	0.5	1.08	1.04	1.05	1.04	1.13	1.05	1.15	1.18	1.08	1.03	1.09	1.06	1.14	1.11	1.16	1.20
			2	1.02	1.01	1.01	1.01	1.05	1.05	1.04	1.02	1.02	1.01	1.01	1.01	1.03	1.03	1.04	1.01
HQ	620	U	0.5	1.19	1.11	**1.29**	1.22	1.15	1.00	1.07	1.01	1.18	1.09	1.22	1.15	1.15	1.01	1.21	1.16
			2	1.10	1.04	1.17	1.07	0.99	1.03	1.04	1.00	1.06	1.04	1.12	1.03	1.02	1.00	1.02	1.03
		HO	0.5	1.08	1.07	1.12	1.07	1.04	1.08	1.03	0.99	1.05	1.06	1.03	1.03	1.09	1.04	1.01	1.02
			2	1.05	1.03	1.06	1.03	1.02	1.02	1.02	1.02	1.02	1.01	1.02	1.02	1.04	1.00	1.02	0.97
	1240	U	0.5	1.10	1.04	1.18	1.07	1.18	1.03	1.08	1.07	1.10	1.07	1.12	1.03	1.11	1.01	1.09	1.11
			2	1.01	1.01	1.03	1.01	1.05	1.06	1.04	1.04	1.02	1.01	1.02	1.01	1.04	1.00	1.01	1.01
		HO	0.5	1.08	1.06	1.09	1.03	1.14	1.11	1.17	1.20	1.09	1.06	1.07	1.05	1.18	1.07	**1.28**	1.20
			2	1.03	1.02	1.01	1.01	1.04	1.00	1.04	1.04	1.03	1.02	1.03	1.01	1.02	1.02	1.06	1.02

Note. Substantial differences are shown in bold. DINA = deterministic inputs, noisy "and" gate; DINO = deterministic input, noisy "or" gate; RP-GDI = restrictive progressive GDI; RT-GDI = restrictive threshold GDI; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; IQ = item quality; J = pool size; AD = attribute distribution; LQ = low-quality; HQ = high-quality; U = uniform; HO = higher-order.

3.2.2.3 The Impact of the Attribute Distribution

Using the uniform distribution instead of the HO distribution in generating attribute vectors resulted in negligible differences in the classification rates across different factors (e.g., item selection index, generating model, and item quality). However, in some conditions, the HO distribution yielded higher classification rates than the uniform distribution, and the differences in the CVC rates were substantial. For example, the HO distribution resulted in higher CVC rates in the following conditions: using the GDI with short tests, LQ items, and a small sample regardless of the pool size for the DINA model; using the RT-GDI with short tests, LQ items, a large pool, a small sample, an r^{max} of .1, and a small β for the DINO model; and using the RP-GDI with short tests, a small pool, a large sample, an r^{max} of .1, and a small β regardless of the item quality for the *A*-CDM.

Likewise, the impact of the attribute distribution on the item exposure rates was mostly negligible regardless of the other factors. The chi-square ratios using the RP-GDI and RT-GDI are shown in Table 3.4 for the DINA and DINO models. However, the differences in the chi-square values were moderate to substantial in some conditions. Specifically, for the DINA and DINO models, the differences in the chi-square values were substantial when the RP-GDI was used with long tests and an r^{max} of .1 regardless of the item quality, pool size, sample size, and β, and those differences were moderate when the RP-GDI was used with short tests, HQ items, a small pool, a small sample, an r^{max} of .1, and a large β. In addition, there were fewer cases where the differences in the chi-square values were substantial when the RT-GDI was used instead of the RP-GDI.

Table 3.4: The Chi-Square Ratios Comparing HO vs. Uniform Distribution

				DINA								DINO							
				RP-GDI				RT-GDI				RP-GDI				RT-GDI			
				Test Length															
				10		20		10		20		10		20		10		20	
				r^{max}															
IQ	J	N	β	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
LQ	620	500	0.5	1.00	0.99	**1.36**	**1.29**	0.91	0.88	0.85	0.87	0.98	0.99	**1.33**	**1.27**	0.92	0.90	0.89	0.93
			2	1.04	1.03	**1.45**	1.15	0.95	0.95	1.00	1.03	1.02	1.02	**1.39**	1.14	1.01	0.96	0.99	1.00
		1000	0.5	1.02	1.02	**1.48**	**1.38**	0.89	0.87	0.93	0.93	1.01	1.00	**1.42**	**1.34**	0.96	0.94	1.01	1.01
			2	1.07	1.03	**1.53**	1.18	0.98	0.95	1.02	1.07	1.04	1.02	**1.50**	1.16	0.99	0.96	1.03	1.04
	1240	500	0.5	1.04	1.02	**1.36**	1.22	1.03	0.92	0.98	1.06	1.02	1.00	**1.41**	1.25	1.06	1.09	1.02	1.11
			2	1.04	1.02	1.20	1.09	1.00	1.04	1.04	1.03	1.04	1.03	1.19	1.10	1.04	0.97	1.07	1.05
		1000	0.5	1.01	1.01	**1.44**	1.24	1.07	1.09	1.06	1.08	1.04	1.03	**1.42**	1.21	1.03	0.96	0.94	0.98
			2	1.04	1.02	1.21	1.09	0.99	1.02	1.05	1.03	1.04	1.03	1.20	1.09	1.05	0.97	1.04	1.05
HQ	620	500	0.5	**1.26**	1.20	**2.44**	**2.11**	0.92	0.95	1.05	1.02	1.22	1.16	**2.34**	**1.99**	0.98	1.07	1.06	1.07
			2	1.24	1.10	**2.52**	**1.51**	1.05	1.01	1.04	1.18	1.20	1.08	**2.41**	**1.49**	1.06	1.03	1.09	1.10
		1000	0.5	**1.38**	1.24	**2.82**	**2.39**	1.01	0.89	1.09	1.04	**1.36**	1.20	**2.75**	**2.23**	1.03	1.03	**1.27**	1.21
			2	**1.30**	1.12	**2.79**	**1.56**	1.03	1.02	1.06	1.16	1.25	1.10	**2.64**	**1.52**	1.04	1.03	1.09	1.16
	1240	500	0.5	1.24	1.15	**2.30**	**1.67**	1.07	1.19	1.18	1.22	1.25	1.15	**2.31**	**1.71**	1.18	1.17	1.17	1.18
			2	1.15	1.08	**1.56**	1.22	1.02	1.02	1.10	1.11	1.13	1.08	**1.57**	1.23	1.04	1.09	1.15	1.11
		1000	0.5	**1.27**	1.14	**2.48**	**1.74**	1.10	1.11	1.09	1.09	**1.26**	1.16	**2.41**	**1.69**	1.11	1.11	1.00	1.09
			2	1.13	1.08	**1.59**	1.22	1.03	1.08	1.10	1.12	1.12	1.07	**1.55**	1.22	1.05	1.07	1.09	1.10

Note. Substantial differences are shown in bold. HO = higher-order; DINA = deterministic inputs, noisy "and" gate; DINO = deterministic input, noisy "or" gate; RP-GDI = restrictive progressive GDI; RT-GDI = restrictive threshold GDI; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; IQ = item quality; J = pool size; N = sample size; LQ = low-quality; HQ = high-quality.

3.2.2.4 The Impact of the Test Length

As expected, increasing the test length resulted in higher classification rates. In addition, the differences in the CVC rates were always substantial regardless of the factors (e.g., item selection index, generating model). Moreover, using LQ items yielded greater differences in the CVC rates compared to HQ items. For example, the differences across all conditions were on average 0.31 and 0.11 for LQ and HQ items when the GDI and the DINA model were used, respectively. Also, those differences were greater when the RT-GDI was used instead of the RP-GDI.

The impact of the test length on the item exposure rates was mostly negligible to moderate when the GDI and RT-GDI were used as the item selection indices across different conditions. The chi-square ratios are shown in Table 3.5 for the DINA, DINO, and *A*-CDM. However, interestingly, using short test lengths (i.e., 10-item test) resulted in smaller chi-square values than long test lengths (i.e., 20-item test) in some conditions when the RT-GDI was used with the *A*-CDM. For example, using the RT-GDI yielded substantial differences in the chi-square values when a large pool was used with a small β and an r^{max} of .1 regardless of the item quality, sample size, and attribute distribution.

Using short test lengths (i.e., 10-item test) resulted in larger chi-square values compared to the long test lengths (i.e., 20-item test), and the differences in the chi-square values were generally substantial when the RP-GDI was used regardless of the different factors. However, there were some conditions where the differences were negligible to moderate when the RP-GDI was used as the item selection index. For example, for the DINA and DINO models, the differences were negligible when an r^{max} of .2 and a large β were used with a large pool and the HO distribution regardless of the sample size and the item quality, and the differences were moderate when an r^{max} of .2 and a large β were used with LQ items, a large pool, and the uniform

distribution regardless of the sample size. Again, using long test lengths resulted in larger chi-square values than short test lengths in some conditions where the RP-GDI was used with the A-CDM.

3.2.2.5 The Impact of the Pool Size

The impact of the pool size on the classification rates was mostly negligible across different factors (e.g., item selection index, generating model) except for some conditions. For example, the difference in the CVC rates was substantial for the GDI and the DINO model when short tests were used with LQ items, a small sample, and the uniform distribution, where a large pool (i.e., J=1240) resulted in higher CVC rates compared to a small pool (i.e., J=620), and the difference was 0.09. In addition, the differences in the CVC rates were generally substantial regardless of the other factors, and especially when the A-CDM was used.

For all the CDMs, increasing the pool size resulted in higher item exposure rates, and the differences in the chi-square values were substantial across different conditions, as shown by the chi-square ratios are shown in Table 3.6. However, there were some conditions where the differences were moderate when the RT-GDI, long tests, an r^{max} of .2, and a small β were used regardless of the item quality, sample size, and attribute distribution.

3.2.2.6 The Impact of the Desired r^{max} Value

Increasing the r^{max} value generally resulted in negligible differences in the classification rates regardless of the other factors. However, there were some conditions where the differences in the CVC rates were substantial especially when the RP-GDI and RT-GDI were used with the A-CDM. For example, the differences were substantial for the RP-GDI and RT-GDI with the A-CDM, when long tests were used regardless of the item quality, pool size, sample size, attribute distribution, and β

(with a large β and HQ items as an exception, where the differences were negligible); and when long tests were used with a small pool regardless of the item quality, sample size, and attribute distribution.

As expected, increasing the r^{max} value resulted in higher item exposure rates across different conditions. The chi-square ratios are shown in Table 3.7 for the DINA, DINO, and A-CDM. Several results can be noted. First, long tests (i.e., 20-item tests) yielded higher chi-square values than short tests (i.e., 10-item tests) with respect to the r^{max} value. Second, the differences in the chi-square values were always substantial when the RP-GDI was used as an item selection index. Third, using a large r^{max} value mostly yielded substantial differences in the chi-square values when RT-GDI was used; however, there were some conditions where the differences were negligible to moderate. For example, for the DINA model, the differences were negligible when short tests and a small β were used with HQ items, a small pool, and a small sample regardless of the attribute distribution. Also, for the DINA model, the differences were moderate when long tests and a small β were used with a large sample regardless of the item quality, pool size, and attribute distribution.

3.2.2.7 The Impact of β

Increasing the β value resulted in negligible differences in the classification rates for the RP-GDI using the DINA and DINO models; however, it generally yielded substantial differences for the same index using the A-CDM. In addition, an increase in the β value generally resulted in substantial differences in the CVC rates for the RT-GDI regardless of the other factors (e.g., generating model, test length). For the DINA model and the RP-GDI, the differences were substantial when short tests and an r^{max} of .1 were used with LQ items, a small pool, a small sample, and the HO distribution, where increasing the β value from 0.5 to 1.0 resulted in higher CVC rates (i.e., the difference was 0.06); when short tests and an r^{max} of .2 were used with LQ

items, a small pool, a small sample, and the uniform distribution where increasing the β value from 0.5 to 1 resulted in higher CVC rates (i.e., the difference was 0.07); and when long tests and an r^{max} of .2 were used with LQ items, a small pool, a small sample, and the HO distribution where increasing the β value from 0.5 to 1.0 resulted in higher CVC rates (i.e., the difference was 0.06).

Increasing the β value resulted in substantial differences in the chi-square values regardless of the other factors. Moreover, for the RP-GDI, increasing the β value from 0.5 to 1.0 yielded greater differences in the chi-square values compared to increasing the β value from 1.0 to 2.0. However, for the RT-GDI, increasing the β value from 1.0 to 2.0 yielded greater differences in the chi-square values compared to increasing the β value from 0.5 to 1.0.

3.3 Discussion and Conclusion

In this article, the efficiency of the new index, the GDI, was investigated in terms of the classification accuracy and the item exposure using two item exposure control methods, namely, RP and RT methods. In addition, the impact of different factors on the item exposure was also examined. Based on the factors manipulated in the simulation study, as expected, the RP method resulted in more uniform item exposure rates compared to the RT method because of the method's probabilistic nature. Moreover, the factors, namely, the item quality, attribute distribution, test length, pool size, prespecified desired exposure rate, and β, generally had a substantial impact on the exposure rates when the RP method was used; however, fewer factors, such as the pool size, prespecified desired exposure rate, and β, generally had a substantial impact on the exposure rates when the RT method was used. The other factors had moderate or negligible effects on the item exposure rates with some exceptions. Overall, the results of this study suggest that, relative to other methods examined, the RP-GDI is a more promising method for use in practice.

This study showed that the new index performed efficiently with the item exposure control methods in terms of attribute classification accuracy and item exposure rates. Nonetheless, more research must be done to ensure the index is practical. First, the results were obtained using 10 replications in the RP method because it did not yield stable results across the conditions. In more detail, the increase in the β value did not increase the classification rates in all conditions because of the random component. Results that are more stable without any replication in practice must be obtained. Second, at present, the efficiency of only a limited number of item exposure control methods has been examined in the context of CD-CAT. It would, therefore, be instructive to examine the applicability of the other item exposure control methods in traditional CAT (e.g., multiple maximum exposure rates; Barrada et al., 2009)

in the context of CD-CAT. Third, some constraints in the design of the Q-matrix should be investigated. The Q-matrix in this study involved all possible q-vectors. However, in practice, this may not be the case, particularly when the CDMs are retrofitted to existing data. Also, the impact of Q-matrix misspecifications needs to be investigated in the CAT framework. Third, only the item exposure control methods that can work for fixed-length tests were used in this study. It would be interesting to examine the efficiency of the methods when variable-length tests are used. Finally, a few simplifications were made in the design of this study to keep the scope of this work manageable. These simplifications include fixing the number of attributes and assuming that the item parameters were known. To obtain more generalizable conclusions, these factors should be varied in future research.

Table 3.5: The Chi-Square Ratios Comparing Short vs. Long Test Length

					DINA				DINO				*A*-CDM			
					RP-GDI		RT-GDI		RP-GDI		RT-GDI		RP-GDI		RT-GDI	
					r^{max}											
IQ	J	N	AD	β	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
LQ	620	500	U	0.5	**3.27**	**2.32**	0.87	0.96	**3.27**	**2.33**	0.86	0.96	**2.66**	**1.67**	1.01	**0.67**
				2	**2.47**	**1.45**	1.14	1.00	**2.47**	**1.47**	1.08	0.99	**1.58**	0.95	**1.36**	0.86
			HO	0.5	**2.40**	**1.79**	0.92	0.97	**2.43**	**1.83**	0.90	0.93	**2.70**	**1.69**	1.00	**0.66**
				2	**1.76**	**1.30**	1.08	0.92	**1.80**	**1.32**	1.10	0.96	**1.59**	0.96	**1.35**	0.87
		1000	U	0.5	**3.53**	**2.40**	0.91	1.03	**3.52**	**2.39**	0.90	0.98	**2.67**	**1.68**	1.00	**0.66**
				2	**2.54**	**1.50**	1.12	1.03	**2.60**	**1.49**	1.13	1.02	**1.59**	0.95	**1.36**	0.86
			HO	0.5	**2.43**	**1.77**	0.88	0.96	**2.50**	**1.78**	0.86	0.91	**2.71**	**1.69**	0.98	**0.66**
				2	**1.77**	**1.31**	1.07	0.91	**1.81**	**1.31**	1.08	0.94	**1.60**	0.96	**1.36**	0.86
	1240	500	U	0.5	**2.42**	**1.75**	1.00	1.15	**2.48**	**1.78**	1.05	1.13	**1.71**	1.17	**0.69**	0.79
				2	**1.55**	1.19	1.03	0.93	**1.54**	1.18	1.03	0.95	0.98	0.78	0.89	0.91
			HO	0.5	**1.84**	**1.46**	1.05	1.01	**1.80**	**1.43**	1.08	1.10	**1.71**	1.17	**0.67**	0.79
				2	**1.34**	1.11	0.99	0.94	**1.35**	1.11	1.01	0.88	0.99	0.78	0.89	0.89
		1000	U	0.5	**2.56**	**1.79**	1.05	1.13	**2.50**	**1.73**	1.02	1.22	**1.71**	1.17	**0.67**	0.80
				2	**1.55**	1.18	1.04	0.93	**1.54**	1.17	1.01	0.94	0.98	0.78	0.89	0.90
			HO	0.5	**1.80**	**1.46**	1.07	1.14	**1.82**	**1.47**	1.11	1.19	**1.72**	1.17	**0.66**	0.79
				2	**1.32**	1.11	0.98	0.92	**1.33**	1.11	1.02	0.87	0.99	0.78	0.88	0.89
HQ	620	500	U	0.5	**3.46**	**2.60**	0.97	0.95	**3.73**	**2.67**	0.91	0.86	**2.63**	**1.57**	1.02	**0.63**
				2	**3.10**	**1.78**	1.05	1.00	**3.17**	**1.82**	1.10	0.96	**1.56**	0.90	**1.39**	0.82
			HO	0.5	**1.78**	**1.48**	0.85	0.89	**1.94**	**1.56**	0.84	0.86	**2.68**	**1.64**	1.03	**0.64**
				2	**1.53**	**1.29**	1.06	0.86	**1.58**	**1.32**	1.07	0.90	**1.58**	0.92	**1.39**	0.82
		1000	U	0.5	**3.76**	**2.85**	0.91	0.96	**3.86**	**2.83**	0.96	0.99	**2.61**	**1.58**	1.00	**0.62**
				2	**3.31**	**1.81**	1.11	0.97	**3.33**	**1.82**	1.10	0.99	**1.57**	0.91	**1.39**	0.81
			HO	0.5	**1.84**	**1.48**	0.85	0.81	**1.91**	**1.52**	0.78	0.85	**2.71**	**1.62**	0.99	**0.63**
				2	**1.54**	**1.30**	1.07	0.86	**1.58**	**1.32**	1.05	0.88	**1.59**	0.93	**1.39**	0.82
	1240	500	U	0.5	**2.79**	**2.02**	1.16	1.14	**2.87**	**2.09**	1.07	1.15	**1.61**	1.08	**0.64**	0.80
				2	**1.82**	**1.27**	1.06	0.95	**1.84**	**1.27**	1.08	0.92	0.93	**0.73**	0.84	0.85
			HO	0.5	**1.51**	**1.39**	1.05	1.12	**1.56**	**1.40**	1.08	1.14	**1.66**	1.10	**0.65**	0.78
				2	**1.34**	1.13	0.98	0.88	**1.33**	1.12	0.98	0.91	0.95	**0.73**	0.84	0.85
		1000	U	0.5	**3.00**	**2.07**	1.07	1.18	**2.91**	**2.02**	1.06	1.25	**1.61**	1.08	**0.64**	0.79
				2	**1.85**	**1.26**	1.05	0.94	**1.84**	**1.27**	1.05	0.93	0.93	**0.73**	0.83	0.85
			HO	0.5	**1.53**	**1.35**	1.08	1.21	**1.52**	**1.39**	1.17	**1.28**	**1.66**	1.09	**0.64**	0.79
				2	**1.32**	1.12	0.98	0.91	**1.33**	1.11	1.01	0.91	0.95	**0.74**	0.83	0.85

Note. Substantial differences are shown in bold. DINA = deterministic inputs, noisy "and" gate; DINO = deterministic input, noisy "or" gate; *A*-CDM = additive CDM; CDM = cognitive diagnosis model; RP-GDI = restrictive progressive GDI; RT-GDI = restrictive threshold GDI; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; IQ = item quality; J = pool size; N = sample size; AD = attribute distribution; LQ = low-quality; HQ = high-quality; U = uniform; HO = higher-order.

Table 3.6: The Chi-Square Ratios Comparing Large vs. Small Pool Size

				DINA								DINO								*A*-CDM							
				RPGDI				RTGDI				RPGDI				RTGDI				RPGDI				RTGDI			
				Test Length																							
				10		20		10		20		10		20		10		20		10		20		10		20	
				r^{max}																							
IQ	N	AD	β	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
LQ	500	U	0.5	**2.71**	**2.22**	**3.66**	**2.95**	**1.64**	**1.78**	**1.42**	**1.49**	**2.80**	**2.30**	**3.69**	**3.03**	**1.54**	**1.52**	**1.27**	**1.30**	**3.32**	**2.30**	**5.15**	**3.27**	**1.38**	**1.47**	**2.02**	1.25
			2	**2.16**	**1.94**	**3.44**	**2.38**	**1.89**	**1.77**	**2.09**	**1.90**	**2.17**	**1.94**	**3.48**	**2.41**	**1.91**	**1.80**	**2.01**	**1.88**	**2.00**	**1.53**	**3.23**	**1.88**	**1.72**	**1.75**	**2.63**	**1.67**
		HO	0.5	**2.81**	**2.28**	**3.66**	**2.81**	**1.86**	**1.88**	**1.63**	**1.82**	**2.91**	**2.33**	**3.92**	**2.98**	**1.76**	**1.84**	**1.47**	**1.55**	**3.29**	**2.28**	**5.18**	**3.30**	**1.37**	**1.48**	**2.06**	1.24
			2	**2.17**	**1.92**	**2.84**	**2.26**	**1.99**	**1.94**	**2.18**	**1.90**	**2.22**	**1.96**	**2.97**	**2.33**	**1.98**	**1.82**	**2.16**	**1.97**	**2.00**	**1.54**	**3.21**	**1.89**	**1.73**	**1.76**	**2.61**	**1.71**
	1000	U	0.5	**2.90**	**2.37**	**4.00**	**3.18**	**1.50**	**1.46**	**1.30**	**1.34**	**2.79**	**2.28**	**3.93**	**3.15**	**1.55**	**1.69**	**1.37**	**1.35**	**3.30**	**2.29**	**5.16**	**3.29**	**1.33**	**1.47**	**2.00**	1.22
			2	**2.26**	**1.97**	**3.71**	**2.50**	**1.89**	**1.73**	**2.05**	**1.91**	**2.19**	**1.96**	**3.70**	**2.48**	**1.87**	**1.74**	**2.09**	**1.90**	**2.00**	**1.54**	**3.24**	**1.88**	**1.72**	**1.77**	**2.64**	**1.70**
		HO	0.5	**2.86**	**2.35**	**3.87**	**2.86**	**1.79**	**1.84**	**1.47**	**1.56**	**2.86**	**2.35**	**3.93**	**2.85**	**1.66**	**1.71**	**1.28**	**1.31**	**3.30**	**2.29**	**5.21**	**3.29**	**1.35**	**1.47**	**2.00**	1.23
			2	**2.19**	**1.95**	**2.93**	**2.30**	**1.92**	**1.87**	**2.10**	**1.85**	**2.17**	**1.97**	**2.95**	**2.33**	**1.99**	**1.76**	**2.11**	**1.92**	**2.00**	**1.54**	**3.23**	**1.89**	**1.71**	**1.76**	**2.64**	**1.70**
HQ	500	U	0.5	**2.56**	**2.24**	**3.17**	**2.88**	**1.69**	**1.75**	**1.42**	**1.46**	**2.54**	**2.30**	**3.30**	**2.94**	**1.53**	**1.66**	**1.30**	1.25	**3.18**	**2.18**	**5.20**	**3.19**	**1.33**	**1.47**	**2.12**	1.16
			2	**2.33**	**2.00**	**3.98**	**2.80**	**2.05**	**1.93**	**2.04**	**2.03**	**2.38**	**2.00**	**4.09**	**2.87**	**1.99**	**1.86**	**2.02**	**1.94**	**1.87**	**1.45**	**3.15**	**1.79**	**1.62**	**1.66**	**2.70**	**1.59**
		HO	0.5	**2.53**	**2.16**	**2.99**	**2.29**	**1.96**	**2.20**	**1.60**	**1.74**	**2.61**	**2.27**	**3.26**	**2.52**	**1.85**	**1.83**	**1.43**	**1.39**	**3.17**	**2.18**	**5.14**	**3.26**	**1.34**	**1.44**	**2.13**	1.18
			2	**2.15**	**1.97**	**2.46**	**2.26**	**1.99**	**1.95**	**2.15**	**1.91**	**2.24**	**2.01**	**2.67**	**2.36**	**1.94**	**1.97**	**2.13**	**1.96**	**1.89**	**1.44**	**3.14**	**1.81**	**1.63**	**1.65**	**2.71**	**1.60**
	1000	U	0.5	**2.77**	**2.38**	**3.47**	**3.28**	**1.65**	**1.71**	**1.41**	**1.38**	**2.71**	**2.33**	**3.59**	**3.27**	**1.58**	**1.65**	**1.43**	**1.31**	**3.18**	**2.18**	**5.17**	**3.20**	**1.33**	**1.47**	**2.09**	1.15
			2	**2.53**	**2.06**	**4.51**	**2.96**	**1.93**	**1.88**	**2.03**	**1.95**	**2.48**	**2.06**	**4.49**	**2.95**	**1.95**	**1.85**	**2.05**	**1.97**	**1.87**	**1.45**	**3.16**	**1.79**	**1.63**	**1.66**	**2.73**	**1.58**
		HO	0.5	**2.53**	**2.18**	**3.06**	**2.39**	**1.80**	**2.13**	**1.41**	**1.44**	**2.51**	**2.26**	**3.15**	**2.47**	**1.70**	**1.78**	1.13	1.19	**3.19**	**2.18**	**5.22**	**3.24**	**1.36**	**1.48**	**2.09**	1.17
			2	**2.21**	**1.98**	**2.57**	**2.30**	**1.94**	**1.99**	**2.10**	**1.88**	**2.22**	**2.00**	**2.65**	**2.38**	**1.98**	**1.92**	**2.05**	**1.86**	**1.88**	**1.45**	**3.15**	**1.82**	**1.63**	**1.66**	**2.73**	**1.59**

Note. Substantial differences are shown in bold. DINA = deterministic inputs, noisy "and" gate; DINO = deterministic input, noisy "or" gate; *A*-CDM = additive CDM; CDM = cognitive diagnosis model; RPGDI = restrictive progressive GDI; RTGDI = restrictive threshold GDI; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; IQ = item quality; N = sample size; AD = attribute distribution; LQ = low-quality; HQ = high-quality; U = uniform; HO = higher-order.

Table 3.7: The Chi-Square Ratios Comparing r^{max} of .1 vs. .2

					DINA				DINO				*A*-CDM			
					RPGDI		RTGDI		RPGDI		RTGDI		RPGDI		RTGDI	
					Test Length											
IQ	J	N	AD	β	10	20	10	20	10	20	10	20	10	20	10	20
LQ	620	500	U	0.5	**2.84**	**3.99**	1.25	1.12	**2.80**	**3.93**	1.25	1.12	**3.32**	**5.30**	**1.31**	**1.96**
				2	**2.17**	**3.69**	**1.62**	**1.84**	**2.19**	**3.67**	**1.64**	**1.78**	**1.96**	**3.27**	**1.69**	**2.68**
			HO	0.5	**2.81**	**3.77**	1.20	1.14	**2.83**	**3.75**	1.21	1.17	**3.32**	**5.29**	**1.32**	**2.00**
				2	**2.16**	**2.92**	**1.62**	**1.89**	**2.19**	**2.99**	**1.57**	**1.80**	**1.96**	**3.26**	**1.69**	**2.63**
		1000	U	0.5	**2.90**	**4.26**	**1.32**	1.17	**2.91**	**4.29**	**1.28**	1.18	**3.36**	**5.35**	**1.31**	**1.97**
				2	**2.26**	**3.82**	**1.68**	**1.83**	**2.20**	**3.85**	**1.68**	**1.86**	**1.98**	**3.30**	**1.67**	**2.65**
			HO	0.5	**2.89**	**3.97**	**1.28**	1.16	**2.88**	**4.04**	1.25	1.18	**3.35**	**5.38**	**1.32**	**1.97**
				2	**2.17**	**2.95**	**1.62**	**1.91**	**2.16**	**2.99**	**1.63**	**1.87**	**1.98**	**3.30**	**1.69**	**2.65**
	1240	500	U	0.5	**2.33**	**3.23**	**1.36**	1.17	**2.31**	**3.22**	1.23	1.14	**2.30**	**3.36**	**1.39**	1.21
				2	**1.96**	**2.55**	**1.52**	**1.68**	**1.96**	**2.54**	**1.54**	**1.67**	**1.51**	**1.90**	**1.72**	**1.70**
			HO	0.5	**2.29**	**2.89**	1.22	**1.27**	**2.26**	**2.85**	**1.27**	1.24	**2.30**	**3.37**	**1.42**	1.20
				2	**1.91**	**2.32**	**1.58**	**1.65**	**1.93**	**2.35**	**1.44**	**1.64**	**1.51**	**1.92**	**1.72**	**1.72**
		1000	U	0.5	**2.37**	**3.39**	**1.29**	1.21	**2.38**	**3.43**	**1.40**	1.16	**2.33**	**3.41**	**1.44**	1.21
				2	**1.97**	**2.58**	**1.54**	**1.71**	**1.97**	**2.58**	**1.56**	**1.68**	**1.52**	**1.91**	**1.72**	**1.70**
			HO	0.5	**2.37**	**2.93**	**1.31**	1.23	**2.37**	**2.94**	**1.30**	1.21	**2.33**	**3.40**	**1.44**	1.21
				2	**1.94**	**2.31**	**1.58**	**1.68**	**1.96**	**2.36**	**1.44**	**1.70**	**1.52**	**1.93**	**1.74**	**1.71**
HQ	620	500	U	0.5	**2.66**	**3.54**	1.10	1.13	**2.62**	**3.66**	1.13	1.18	**3.13**	**5.23**	**1.28**	**2.08**
				2	**2.43**	**4.25**	**1.59**	**1.67**	**2.46**	**4.28**	**1.54**	**1.76**	**1.84**	**3.17**	**1.59**	**2.70**
			HO	0.5	**2.53**	**3.06**	1.14	1.10	**2.50**	**3.12**	1.23	1.19	**3.15**	**5.17**	**1.30**	**2.09**
				2	**2.15**	**2.54**	**1.52**	**1.88**	**2.21**	**2.65**	**1.49**	**1.77**	**1.86**	**3.20**	**1.60**	**2.72**
		1000	U	0.5	**2.86**	**3.77**	**1.26**	1.20	**2.84**	**3.87**	**1.28**	1.24	**3.18**	**5.24**	**1.30**	**2.10**
				2	**2.55**	**4.66**	**1.53**	**1.74**	**2.52**	**4.62**	**1.57**	**1.74**	**1.85**	**3.20**	**1.59**	**2.73**
			HO	0.5	**2.57**	**3.19**	1.10	1.15	**2.50**	**3.14**	**1.28**	1.17	**3.19**	**5.33**	**1.31**	**2.07**
				2	**2.20**	**2.61**	**1.51**	**1.89**	**2.22**	**2.66**	**1.56**	**1.87**	**1.87**	**3.21**	**1.59**	**2.72**
	1240	500	U	0.5	**2.34**	**3.22**	1.14	1.16	**2.37**	**3.25**	1.22	1.15	**2.15**	**3.21**	**1.41**	1.13
				2	**2.09**	**2.99**	**1.50**	**1.66**	**2.07**	**3.01**	**1.44**	**1.68**	**1.42**	**1.80**	**1.62**	**1.59**
			HO	0.5	**2.17**	**2.34**	**1.28**	1.20	**2.18**	**2.42**	1.21	1.15	**2.17**	**3.28**	**1.40**	1.16
				2	**1.96**	**2.33**	**1.49**	**1.68**	**1.98**	**2.34**	**1.52**	**1.63**	**1.42**	**1.85**	**1.62**	**1.61**
		1000	U	0.5	**2.46**	**3.56**	**1.30**	1.17	**2.45**	**3.52**	**1.34**	1.13	**2.18**	**3.24**	**1.43**	1.16
				2	**2.08**	**3.05**	**1.49**	**1.67**	**2.09**	**3.03**	**1.48**	**1.68**	**1.44**	**1.81**	**1.62**	**1.58**
			HO	0.5	**2.21**	**2.49**	**1.31**	1.17	**2.24**	**2.46**	**1.34**	1.23	**2.18**	**3.31**	**1.43**	1.16
				2	**1.98**	**2.34**	**1.56**	**1.69**	**2.00**	**2.38**	**1.51**	**1.69**	**1.44**	**1.86**	**1.62**	**1.59**

Note. Substantial differences are shown in bold. DINA = deterministic inputs, noisy "and" gate; DINO = deterministic input, noisy "or" gate; *A*-CDM = additive CDM; CDM = cognitive diagnosis model; RPGDI = restrictive progressive GDI; RTGDI = restrictive threshold GDI; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; IQ = item quality; J = pool size; N = sample size; AD = attribute distribution; LQ = low-quality; HQ = high-quality; U = uniform; HO = higher-order.

References

Barrada, J. R., Abad, F. J., & Veldkamp, B. P. (2009). Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema, 21*, 313-320.

Barrada, J. R., Mazuela, P., & Olea, J. (2006). Maximum information stratification method for controlling item exposure in computerized adaptive testing. *Psicothema, 18*, 156-159.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *The British Journal of Mathematical and Statistical Psychology, 61*, 493-513.

Barrada, J. R., Veldkamp, B. P., & Olea, J. (2009). Multiple maximum exposure rates in computerized adaptive testing. *Applied Psychological Measurement, 58-73*, 313-320.

Chang, H.-H. (2004). Understanding computerized adaptive testing-From Robbins-Monro to Lord and beyond. In D. Kaplan (Eds.), *The Sage handbook of quantitative methodology for the social sciences* (p. 117-133). Thousand Oaks, CA: Sage.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

Chang, H.-H., & Ying, Z. (1999). α-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222.

Chang, H.-H., Qian, J., & Ying, Z. (2001). α-Stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement, 25*, 333-341.

Chang, S.-W. & Twu, B. (1998). *A comparative study of item exposure control methods in computerized adaptive testing.* (Research Report 98-3). Iowa City, IA: American College Testing.

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*, 129-145.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619-632.

Davey, T., & Parshall, C. (1995, April). *New algorithms for item selection and exposure control with computer adaptive testing*. Paper presented at the annual meeting of the American Education Research Association, San Francisco, CA.

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

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

de la Torre, J., & Chiu, C.-Y. (2015). A general method of empirical Q-matrix validation. *Psychometrika*. Advance online publication. doi:10.1007/s11336-015-9467-8

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

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

Deng, H., Ansley, T., & Chang, H.-H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement, 47*, 202-226.

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2007). A Review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning, and Assessment , 5* (8). (Retrieved May 1, 2007, from http://www.jtla.org)

Han, K. T. (2012). An efficiency balanced information criterion for item selection in computerized adaptive testing. *Journal of Educational Measurement, 46*, 225-246.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.

Hartz, S. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practice*. Unpublished doctoral thesis, University of Illinois at Urbana-Champain.

Hartz, S., Roussos, L., & Stout, W. (2002). *Skills diagnosis: Theory and practice* [User manual for Arpeggio software]. Princeton, NJ: Educational Testing Service.

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

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*, 167-188.

Lee, Y., Ip, E. H., & Fuh, C. (2007). A strategy for item exposure in multidimensional computerized adaptive testing. *Educational and Psychological Measurement, 68*, 215-232.

Li, Y. H., & Schafer, W. D. (2005). Increasing the homogeneity of CATs item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests. *Journal of Educational Measurement, 42*, 245-269.

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

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*, 187-194.

Revuelta, J. (1995). *El control de la exposición de los items en tests adaptativos informatizados* [Item exposure control in computerized adaptive tests]. Unpublished master's dissertation, Universidad Autónoma de Madrid, Spain.

Revuelta, J., & Ponsoda, V. (1996). Métodos sencillos para el control de las tasas de expósicion en tests adaptativos informatizados [Simple methods for item exposure control in CATs]. *Psicologica, 17*, 161-172.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Stocking, M. L., & Lewis, C. (1995a). *A new method of controlling item exposure in computerized adaptive testing* (Research Report 95-25). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995b). *Controlling item exposure conditional on ability in computerized adaptive testing* (Research Report 95-24). Princeton, NJ: Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Centre.

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

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

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer et al. (Ed.). *Computerized adaptive testing: A primer* (pp. 101-133). Hillsdale: Erlbaum.

Wang, C., Chang, H.-H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement, 48*, 255-273.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17-27.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Wen, J., Chang, H., & Hau, K. (2000, April). *Adaption of α-stratified method in variable length computerized adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Xu, X., Chang, H.-H., & Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Chapter 4
Study III: A Blocked-CAT Procedure for CD-CAT

Abstract

This paper introduces a blocked-design procedure for cognitive diagnosis computerized adaptive testing (CD-CAT), which allows examinees to review items and change their answers during test administration. Four blocking versions of the new procedure were proposed. In addition, the impact of several factors, namely, item quality, generating model, block size, and test length, on the classification rates was investigated. Two popular item selection indices in CD-CAT were used and their efficiency was compared using the new procedure. The results showed that the new procedure is promising for allowing item review with a small loss in attribute classification accuracy under some conditions. This indicates that, as in traditional CAT, that the use of block design in CD-CAT has the potential to address certain issues in practical testing situations (e.g., correcting careless errors, reducing student anxiety).

Keywords: cognitive diagnosis model, computerized adaptive testing, item review

4.1 Introduction

The debate over whether to provide examinees the options to review items and change answers during test administration has continued for several years. Test takers and test developers have different attitudes toward these options. Test takers want to benefit from item review and answer change, which reduce test anxiety and thus

increase test scores legitimately. However, test developers are reluctant to provide these options to test takers for several reasons (Vispoel, Clough, & Bleiler, 2005). The two most common concerns are decreased testing efficiency (e.g., longer testing times) and illegitimate score gains (e.g., test-taking strategies).

Wise (1996) classified score gains as legitimate and illegitimate. The former refers to a score gain in which examinees, who possess the required knowledge to answer an item correctly, increase their scores after they review the item and change their answer. The latter refers to a score gain in which examinees, who do not possess the required knowledge, are somehow able to answer the item correctly because they get a clue from other items, for example. On one hand, examinees can obtain legitimate score gains by using item review and answer change. In turn, the validity of the test increases, and therefore, inferences from the test results become more meaningful and appropriate. On the other hand, providing these options can decrease testing efficiency by lengthening testing times and testing precision with higher errors in ability estimates because of the illegitimate score gains (Vispoel, Rocklin, & Wang, 1994; Wise, 1996).

There is a common belief among examinees and college instructors that changing initial responses to items about which examinees are uncertain might lower the examinees' test scores (Benjamin, Cavell, & Schallenberger, 1984). In contrast to this belief, researchers have shown that most examinees changed their answers when they were allowed, and those changes were generally from incorrect to correct. So much, those who made changes improved their test scores (Benjamin et al., 1984). Moreover, the results in those studies showed that examinees changed their answers for only a very small percentage of answers, but a large number of examinees made changes for at least a few items.

Researchers have investigated the impact of item review and answer change on paper-and-pencil tests for nearly 100 years (e.g., Benjamin et al., 1984; Crocker &

Benson, 1980; Mathews, 1929; Mueller & Wasser, 1977; Smith, White, & Coop, 1979; Waddell & Blankenship, 1995), and the impact on computerized adaptive testing (CAT) for the last two decades (e.g., Han, 2013; Liu, Bridgeman, Lixiong, Xu, & Kong, 2015; Olea, Revuelta, Ximenez, & Abad, 2000; Stocking, 1997; Vispoel, Hendrickson, & Bleiler, 2000; Wainer, 1993; Wise, 1996). However, there is still doubt about providing item review and answer change to examinees in CAT. Researchers have recently suggested that reviewable CAT might introduce bias in the ability estimation and an increase in the standard error of measurement (Papanastasiou & Reckase, 2007). In addition, providing these options in CAT requires more complicated item selection algorithms and longer testing time, and results in lower measurement precision and an increase in the possibility of artificially inflated scores (Wise, 1996).

Reviewable CAT requires more complicated item selection algorithms because most of the item selection algorithms in CAT rely on a provisional ability estimate to select the next item. Changing the answer of an item during the test administration can make the following items no longer appropriate for estimating the ability level (Yen, Ho, Liao, & Chen, 2012). In addition, structures that are more flexible must be developed for examinees' diverse review styles. For example, some examinees like to review item by item sequentially; however, others mark some of the items and review them later (Wise, 1996). Researchers have also suggested that reviewable CAT requires longer testing times. For example, study results showed that item review in computer-based testing increased the average testing time by about 25% (Revuelta, Ximenez, & Olea, 2003; Vispoel, Wang, de la Torre, Bleiler, & Dings, 1992). Moreover, Wise (1996) noted another concern related to testing time: he postulated that only examinees who can quickly complete the test benefit from item review if the testing time is limited.

As noted before, examinees can have illegitimate score increases even if they do

not possess the required knowledge to answer the item correctly. This happens when they can get a clue from the characteristics of other items on the test, or they can use specific testing strategies (e.g., the Wainer strategy, the Kingsbury strategy, and the generalized Kingsbury strategy). Although examinees may not be able to successfully implement these strategies, standardized test preparation companies can teach them how to do so (Vispoel et al., 2000).

In the Wainer strategy (Wainer, 1993), examinees intentionally give incorrect answers to all items on the first pass, which leads them to gradually encounter relatively easier items. After item review, the examinees replace all the answers with the correct ones. To get the full benefit of using this strategy, an examinee must have all the knowledge required to give correct answers on the second pass. Therefore, examinees with high proficiency can generally benefit from this strategy, and increase their test scores (Wise, 1996). However, using this strategy involves some risk. Failure on even a single item might result in underestimation of the examinee's ability level (Gershon & Bergstrom, 1995). Also, researchers suggested that examinees who used the Wainer strategy can be detected from the number of items whose answers changed and the size of standard error of the estimates (Vispoel, Rocklin, Wang, & Bleiler, 1999). Similarly, restricted item review can be used to safe guard against the use of the Wainer strategy.

Stocking (1997) proposed a blocked-design CAT in which item review was allowed within a block of items and investigated the impact of the Wainer strategy with and without item review on the test. She noted that the bias in the estimates and the standard errors were at acceptable levels using this method. Later, several studies supported the finding that there was no significant difference in the accuracy of ability estimation between limited review and no review procedures when using the block design (Vispoel, 2000; Vispoel, Clough, Bleiler, Hendrickson, & Ihrig, 2002; Vispoel et al., 2005). Moreover, researchers have also shown that testing time increased by only

5-11% on average with the majority of examinees indicating that they had adequate opportunity for item review and answer change in the blocked-design CAT (Vispoel et al., 2005). However, Han (2013) noted that the blocked-design CAT still did not allow test takers to skip items. He proposed an item pocket method in which examinees had the option to skip items in addition to reviewing items and changing answers.

In the Kingsbury strategy (Green, Bock, Humphreys, Linn, & Reckase, 1984; Kingsbury, 1996), the examinees know that the difficulty of an item depends on the response to the previous item, and understand the correctness of their response is based on the difficulty level of the current item. In other words, if the current item is less difficult than the previous item, then the answer to the previous item is likely incorrect. Kingsbury (1996) investigated the impact of this strategy in CAT and found that using this strategy resulted in substantial score gains especially for low-proficiency examinees, modest score gains for moderate-proficiency examinees, and very small gains for high-proficiency examinees. However, it is not clear how accurately examinees can detect the difficulty levels of the items. Green et al. (1984) conducted a study in which examinees judged the difficulty of items, and the results showed examinees did not successfully distinguish item difficulty. Moreover, Wise, Finney, Enders, Freeman, and Severance (1999) found that examinees judged item difficulty poorly without actually solving the items. Similar to the Kingsbury strategy, in the generalized Kingsbury strategy (Wise et al., 1999), examinees distinguish the difficulty of all item pairs on the test.

Having the options to review items and change answers during test administration has several benefits for examinees. These options are beneficial for correcting typing/careless errors, misreading of items, temporary lapses in memory, reconceptualization of answers to previously administered items, and test validity (Vispoel, 1998). The results of studies on item review clearly showed that examinees highly endorsed item review in computer-based test administration (e.g., Gershon & Bergstrom, 1995;

Legg & Buhr, 1992; Vispoel, 1998; Vispoel & Coffman, 1994; Vispoel, 2000; Vispoel et al., 2002, 2005). In addition, these options can alter careless errors made by examinees, and relax the testing environment for examinees who have high test anxiety (Vispoel, 2000). Careless errors can result in inaccurate measurement of the examinee's ability, and this is a threat to test validity (Stone & Lunz, 1994).

Another procedure which allows item review and answer change is multistage testing (MST). In MST, the test adaption occurs at the sets of item level or the testlet level instead of the item level. In MST, items are preassembled into modules *prior* to the test administration. In contrast, in the blocked-design CAT, items are grouped into blocks on the fly or *during* the actual test administration. Hendrickson (2007) summarized the MST procedure, which involves several adaptive stages within the test administration. In the first stage, different items with a broad range of difficulty levels are given to obtain initial estimates. Based on the results from the first stage, a block of items with difficulty levels appropriate for the examinee's ability level is given in the next stage. When appropriate, this block includes different content domains. Depending on the test, this stage can be repeated. The stage is useful in differentiating ability within a narrower range. Several testing companies have started using MST in their exams (e.g., Medical Licensure Examination, Graduate Record Examination; Robin, Steffen, & Liang, 2014).

The advantages of using MST can be summarized as follows: it can increase test construction and test form quality, control exposure rates of test materials, provide better test security, obtain greater assurance of local independence, minimize item ordering and context effects, and allow item review during the test administration (Hendrickson, 2007). MST provides better test quality and security because several blocks of items can be created in which content balance and item difficulty can be considered within the block. Local independence requires the examinee's response to the current item should not have a relationship with previous items. If local item

dependence exists among some items in the tests, it violates the local independence assumption in IRT, which is widely used in item-level adaptive testing. MST also allows examinees to review items and skip them within a block during the test administration. Given these advantages, MST is a promising option in adaptive testing that offers a more efficient testing design and environment compared to traditional CAT.

However, several issues regarding MST have been reported. These issues include identifying the optimal number of stages and the range of difficulty within the stages, obtaining feasible statistical information for psychometric and exposure concerns, and differentiating between scores and decisions based on number-correct and IRT scoring procedures (Hendrickson, 2007). For example, the purpose of the test and the characteristics of the population to be tested help determine the length and difficulty of MST tests. Generally, more items are needed in MST to obtain sufficient measurement precision compared to CAT. In addition, constructing the blocks of items and combining them under MST require more work for content domain experts, item developers, and psychometricians than for blocks of items in CAT. Finally, replacing items that are independent within the same block to control item exposure can be difficult in MST (Wainer & Kiely, 1987).

MST applications beyond unidimensional models are also limited. However, most models used for diagnostic testing require a multidimensional latent trait. More specifically, cognitive diagnosis modeling requires the estimation of a set of discrete attributes, an attribute vector, which consists of several dimensions. Constructing the blocks in MST for cognitive diagnosis can be challenging because there are no difficulty parameters for every relevant dimension in CDMs. Multistage testing using CDMs (CD-MST) was first noted by von Davier and Cheng (2014). They discussed several heuristics that can be applied in the CD-MST selection stage and suggested Shannon entropy for selecting the next block of items. However, the authors did not

investigate further how a block of items in the context of cognitive diagnosis for MST can be created.

To date, no research has been done to investigate the impact of item review and answer change in the context of cognitive diagnosis CAT (CD-CAT). The goal of this study was to propose a new CD-CAT administration procedure. In this procedure, a block of items is the unit of administration. Because there were no difficulty parameters to partition the test into blocks in cognitive diagnosis, blocking was performed based on the information using item selection indices. Therefore, different from MST, content balancing and item difficulty were not applicable in the new procedure. Using this blocked design, examinees have an opportunity to review and change their answers within the block.

4.1.1 Cognitive Diagnosis Models

CDMs aim to determine whether or not examinees have a mastery of a set of typically binary attributes. A binary attribute vector represents the presence or absence of the skill or attribute. Let $\boldsymbol{\alpha}_i=\{\alpha_{ik}\}$ be the attribute vector of examinee i, where $i = 1, 2, \ldots, N$ examinees, and $k = 1, 2, \ldots, K$ attributes. The kth element of the vector is 1 when the examinee has mastered the kth attribute, and it is 0 when the examinee has not. In cognitive diagnosis, examinees are classified into latent classes based on the attribute vectors. Each attribute vector corresponds to a unique latent class. Therefore, K attributes create 2^K latent classes or attribute vectors. Similarly, the responses of the examinees to J items are represented by a binary vector. Let $\boldsymbol{X_i} = \{x_{ij}\}$ be the ith examinee's binary response vector for a set of $j = 1, 2, \ldots, J$ items. The required attributes for an item are represented in a Q-matrix (Tatsuoka, 1983), which is a $J \times K$ matrix. The element of the jth row and the kth column, q_{jk}, is 1 if the kth attribute is required to answer the jth item correctly, and it is 0 otherwise.

To date, a variety of general CDMs has been proposed to increase their applicability. For example, the log-linear CDM (Henson, Templin, & Willse, 2009), the general diagnostic model (von Davier, 2008), and the generalized deterministic inputs, noisy "and" gate model (G-DINA; de la Torre, 2011) are examples of general CDMs. The G-DINA model relaxes some of the strict assumptions of the deterministic inputs, noisy "and" gate (DINA; de la Torre, 2009; Haertel, 1989; Junker & Sijtsma, 2001) model, and it partitions examinees into $2^{K_j^*}$ groups, where $K_j^* = \sum_{k=1}^{K} q_{jk}$ is the number of required attributes for item j. A few constrained CDMs can be derived from the G-DINA model using different constraints (de la Torre, 2011). These include the DINA model, which assumes that lacking one of the required attributes is as the same as lacking all of the required attributes; the deterministic input, noisy "or" gate (DINO; Templin & Henson, 2006) model, which assumes that having one of the required attributes is as the same as having all of the required attributes; and the additive CDM (A-CDM; de la Torre, 2011), which assumes that the impacts of mastering the different required attributes are independent of each other.

4.1.2 Computerized Adaptive Testing

CAT has become a popular tool in testing because it allows examinees to receive different tests, with possibly different lengths. Compared to paper-and-pencil tests, the mode of test administration changes from paper to computer, and the test delivery algorithms change from linear to adaptive (van der Linden & Pashley, 2010). Therefore, it provides a tailored test for each examinee, and better ability estimation with shorter test lengths (Meijer & Nering, 1999). A typical CAT procedure involves selecting appropriate items to each examinee's ability level from an item pool, estimating the ability level during or end of the test, and scoring the examinee's performance.

One of the crucial components of CAT is the item selection methods. In traditional

CAT, item selection methods based on the Fisher information are widely used (Lord, 1980; Thissen & Mislevy, 2000); however, those methods are not applicable in CD-CAT because the equivalent latent variables in cognitive diagnosis are discrete. This issue was first noted by Xu, Chang, and Douglas (2003), and they proposed two item selection indices for CD-CAT based on the Kullback-Leibler (K-L) information and Shannon entropy procedure. The efficiency of these indices was compared to random selection using a simulation study. The results of their study showed that both indices outperformed random selection in terms of attribute classification accuracy. Later, Cheng (2009) proposed two item selection indices in CD-CAT, and both were based on the K-L information, namely, the posterior-weighted K-L index (PWKL) and hybrid K-L index (HKL). The results showed that the new indices performed similarly, but both had higher classification rates than the K-L and Shannon entropy procedure. Therefore, the PWKL has become popular in the research of CD-CAT because of its better classification rates and easier implementation.

Recently, Kaplan, de la Torre, and Barrada (2015) proposed two new item selection indices based on the PWKL and the G-DINA model discrimination index (GDI) for CD-CAT. The results showed that the two new indices performed very similarly and higher attribute classification rates compared to the PWKL. In addition, the GDI had the shortest administration time. In this article, the PWKL and GDI will be used as item selection indices with the new CD-CAT administration procedure.

4.1.2.1 Item Selection Methods

4.1.2.1.1 The Kullback-Leibler Information Index

The K-L information is a non-symmetric measure of distance between the two probability distributions, X and Y, where X is assumed to be the true distribution of the data (Cover & Thomas, 1991). The function measuring the distance between

the two distributions is given by

$$K(f,g) = \int \left[log\left(\frac{f(x)}{g(x)}\right)\right] f(x)dx, \tag{4.1}$$

where $f(x)$ and $g(x)$ are the probability density functions of the distributions X and Y, respectively. Larger information gives easier differentiation between the two distributions (Lehmann & Casella, 1998). The item selection methods based on the K-L information have been used in traditional and nontraditional CAT (Chang & Ying, 1996; Xu & Douglas, 2006; McGlohen & Chang, 2008; Xu et al., 2003). All findings showed that the item selection methods based on the K-L information produced good estimation accuracy, and they can work under both continuous and discrete variables (i.e., attribute vectors in CDMs). Thus, the K-L information can be used as an alternative to the Fisher information in CD-CAT.

4.1.2.1.2 The Posterior-Weighted Kullback-Leibler Index

The developments in CD-CAT required more detailed evaluation of the CAT components by the researchers. Therefore, Cheng (2009) proposed the PWKL as an item selection method based on the K-L information. The PWKL is a modified version of the K-L information using the posterior distribution of the attribute vectors as weights. The calculation of the PWKL involves summing the distances between the current estimate of the attribute vector and the other possible attribute vectors, and it is based on the K-L information. The PWKL is given by

$$PWKL_j(\hat{\boldsymbol{\alpha}}_i^{(t)}) = \sum_{c=1}^{2^K}\left[\sum_{x=0}^{1} log\left(\frac{P(X_j = x|\hat{\boldsymbol{\alpha}}_i^{(t)})}{P(X_j = x|\boldsymbol{\alpha}_c)}\right) P(X_j = x|\hat{\boldsymbol{\alpha}}_i^{(t)})\pi_i^{(t)}(\boldsymbol{\alpha}_c)\right], \tag{4.2}$$

where $P(X_j = x|\boldsymbol{\alpha}_c)$ is the probability of the response x to item j given the attribute vector $\boldsymbol{\alpha}_c$, and $\pi_i^{(t)}(\boldsymbol{\alpha}_c)$ is the posterior probability of examinee i given the responses to the t items. The $(t+1)$th item to be administered is the item that maximizes the

PWKL.

4.1.2.1.3 The G-DINA Model Discrimination Index

The G-DINA model discrimination index (GDI) was first proposed by by de la Torre and Chiu (2015) as an index to implement an empirical Q-matrix validation procedure. It measures the (weighted) variance of the probabilities of success of an item given a particular attribute distribution. Later, Kaplan et al. (2015) used it as an item selection index for CD-CAT. To give a summarized definition of the index, define $\boldsymbol{\alpha}_{cj}^{*}$ as the reduced attribute vector consisting of the first K_j^* attributes, for $c = 1, \ldots, 2^{K_j^*}$. For example, if a q-vector is defined as (1,1,0,0,1) for $K_j^* = 3$ number of required attributes, the reduced attribute vector is (a_1,a_2,a_5). Also, define $P(X_{ij} = 1|\boldsymbol{\alpha}_{cj}^{*})$ as the success probability on item j given $\boldsymbol{\alpha}_{cj}^{*}$. The GDI for item j is defined as

$$\varsigma_j^2 = \sum_{c=1}^{2^{K_j^*}} \pi_i^{(t)}(\boldsymbol{\alpha}_{cj}^{*})[P(X_{ij} = 1|\boldsymbol{\alpha}_{cj}^{*}) - \bar{P}_j]^2, \tag{4.3}$$

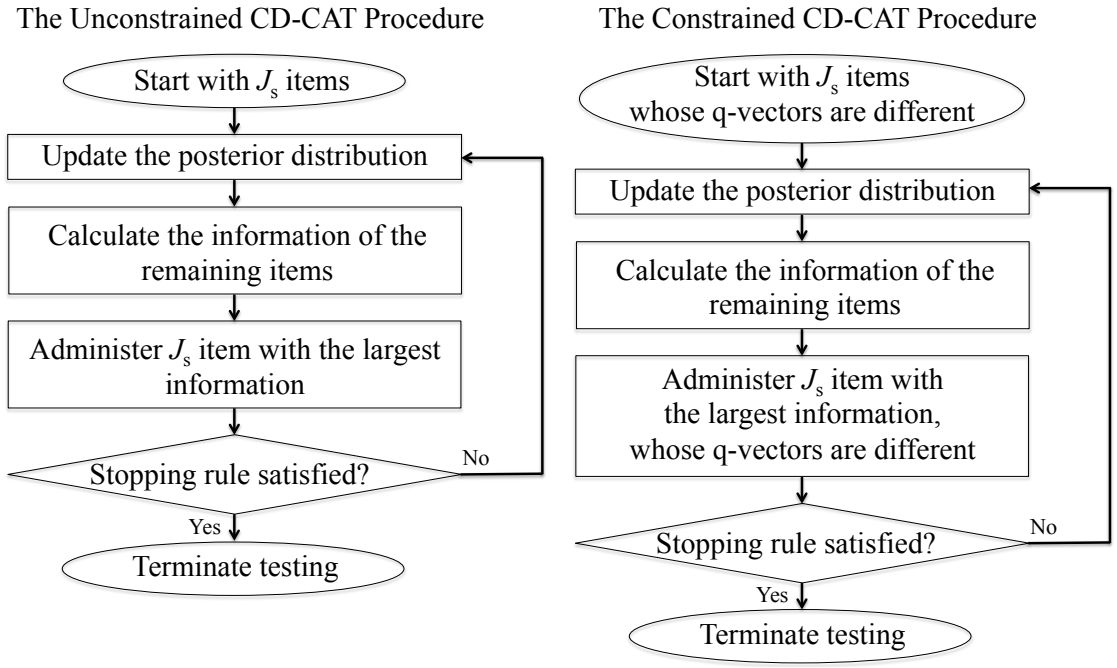
where $\pi_i^{(t)}(\boldsymbol{\alpha}_{cj}^{*})$ is the posterior probability of the reduced attribute vector and $\bar{P}_j = \sum_{c=1}^{2^{K_j^*}} \pi(\boldsymbol{\alpha}_{cj}^{*})P(X_{ij} = 1|\boldsymbol{\alpha}_{cj}^{*})$ is the mean success probability.

4.2 Simulation Study

One of the most important issues with traditional CAT administration is that examinees cannot review their responses to previous items. In this article, a new CD-CAT administration procedure was proposed. In this procedure, a block of items, instead of one item, is administered at a time. Examinees then can review their responses within the same block. Four methods (unconstrained, constrained, hybrid-1, and hybrid-2) were considered in this blocked-design CAT. In the unconstrained version, a block of J_s items was randomly administered first to calculate the examinee's posterior distribution, which was needed to compute the item selection indices. The

most informative J_s items remaining in the pool were administered together, and the posterior distribution was updated. This cycle continued until the test termination rule was satisfied. The unconstrained version of the new procedure is shown in the left panel of Figure 4.1.

Figure 4.1: The New CD-CAT Procedures



In the constrained version, items were selected based on constraint on the q-vectors. Specifically, none of the items within the same block are allowed to have the same q-vector. A previous study showed that item selection indices did not provide relevant information when the same type of items (e.g., the same q-vector) were administered repeatedly (Kaplan et al., 2015). As with the unconstrained version, the first J_s items were randomly selected from the pool; however, the q-vectors of the items were constrained to be different from each other. Again, the posterior distribution was calculated, and the next J_s items were selected from the pool based on the item selection index, with the same constraint. This procedure continued until the termination criterion was satisfied. The right panel of Figure 4.1 shows the constrained version of the proposed procedure.

In the hybrid-1 version, a block of J_s items with the same constraint as in the constrained version was administered during the first half of the test, and the second half of the test was performed without constraint. In the hybrid-2 version, no constraint was applied during the first half of the test, but the constraint was applied in the second half. The viability of the new procedure was examined in a simulation study. The impact of different factors on the attribute classification accuracy of the new procedures were investigated.

4.2.1 Design

4.2.1.1 Data Generation

The impact of the item quality and generating model was considered in the data generation. In addition, a subset of attribute vectors was used to generate the examinees' attribute vectors. Two levels of item quality, namely, low-quality (LQ) and high-quality (HQ), were considered. However, it should be noted that these two terms were used exclusively for this study, and in other studies, they have been defined differently. For the purposes of this study, HQ and LQ can also be viewed as more discriminating and less discriminating, respectively. For LQ items, the lowest and highest success probabilities (i.e., $P(\mathbf{0})$ and $P(\mathbf{1})$) were generated from uniform distributions, $U(0.15, 0.25)$ and $U(0.75, 0.85)$, respectively; and for HQ items, $P(\mathbf{0})$ and $P(\mathbf{1})$ were generated from uniform distributions, $U(0.00, 0.20)$ and $U(0.80, 1.00)$, respectively. Item responses were generated using three reduced models: the DINA model, the DINO model, and the A-CDM. The probability of success was set as discussed above for the DINA and DINO. In addition to these probabilities, the intermediate success probabilities were obtained by allowing each required attribute to contribute equally for the A-CDM. The number of attributes was fixed at $K = 5$.

A more efficient simulation design from Kaplan et al. (2015)'s paper was also used

in this study. One representative of each attribute vector (i.e., no mastery, mastery of a single attribute only, mastery of two attributes only, and so forth) was used and the appropriate weights are applied. Two thousand examinees were generated for each attribute vector, resulting in a total of 12,000 examinees.

4.2.1.2 Item Pool and Item Selection Methods

The Q-matrix was created from $2^K - 1 = 31$ possible q-vectors, each with 40 items. The pool then totaled 1240 items. Only the fixed test lengths were used as a test termination rule. The test lengths were set to 8, 16, and 32 items, and the size of the blocks was set to J_s=1, 2, and 4. In fact, J_s=1 corresponds to traditional CD-CAT administration. Two item selection indices were considered: the PWKL and the GDI. For greater comparability, a uniform distribution of the attribute vectors was used as the prior distributions for the indices across all conditions. In the case of the PWKL, when the estimate of the attribute vector was not unique, a random attribute vector was chosen from the modal attribute vectors.

To compare the efficiency of the indices, the means of the correct attribute classification (CAC) rate and the correct attribute vector classification (CVC) rate were computed for each condition when the fixed test length was used as the termination rule. For each of the six attribute vectors considered in the design, let α_{ikl} and $\hat{\alpha}_{ikl}$ be the kth true and estimated attribute in attribute vector l, $l = 0, 1 \ldots 5$, for examinee i, respectively. The CAC and CVC rates were computed as

$$\begin{aligned} CAC_l &= \frac{1}{2,000} \sum_{i=1}^{2,000} \sum_{k=1}^{5} I[\alpha_{ikl} = \hat{\alpha}_{ikl}], \text{ and} \\ CVC_l &= \frac{1}{2,000} \sum_{i=1}^{2,000} \prod_{k=1}^{5} I[\alpha_{ikl} = \hat{\alpha}_{ikl}], \end{aligned} \tag{4.4}$$

where I is the indicator function. Using appropriate weights (described below), the CAC and the CVC were computed assuming the attributes were uniformly distributed

for the fixed test-length conditions. This study focused on uniformly distributed attribute vectors. Thus, the results based on the six attribute vectors had to be weighted appropriately. For $K = 5$, the vector of the weights were 1/32, 5/32, 10/32, 10/32, 5/32, and 1/32, which represented the proportions of zero-, one-, two-, three-, four-, and five-attribute mastery vectors among the 32 attribute vectors, respectively.

4.2.2 Results

4.2.2.1 Classification Accuracy

This study focused on attribute vectors that were uniformly distributed; however, the sampling design of the study can allow for results to be generalized to different distributions of the attribute vectors (e.g., higher-order; de la Torre & Douglas, 2004). The CAC and CVC rates were computed using appropriate weighted averages. For all conditions, the CAC rates were, as expected, higher than the CVC rates, but the measures showed similar patterns. Thus, only the CVC rates are discussed. The CVC rates under the different factors are presented in Tables 4.1, 4.2, and 4.3 for the DINA, the DINO, and the *A*-CDM, respectively. In Kaplan et al. (2015), differences in the classification rates were evaluated using different cut points to better summarize the findings. Similarly, in this study, differences in the CVC rates were evaluated using two cut points, 0.03 and 0.10. Differences below 0.03 were considered negligible, differences between 0.03 and 0.10 were considered moderate, and differences above 0.10 were considered substantial. In addition, 8-item tests were considered as short, 16-item tests were considered as medium-length, and 32-item tests were considered as long tests.

Using the PWKL with the DINA and the DINO as the generating models, the constrained version with the PWKL had the best classification accuracy among the other blocking versions, whereas the unconstrained version with the PWKL had the

Table 4.1: The CVC Rates Using the DINA Model

			PWKL				GDI			
IQ	J	J_s	UC	H2	H1	C	UC	H2	H1	C
LQ	8	1	0.41	0.41	0.41	0.41	0.53	0.53	0.53	0.53
		2	0.28	0.30	0.33	0.36	0.52	0.53	0.52	0.53
		4	0.20	0.26	0.32	0.35	0.45	0.45	0.46	0.49
	16	1	0.75	0.75	0.75	0.75	0.83	0.83	0.83	0.83
		2	0.58	0.65	0.70	0.73	0.80	0.81	0.80	0.81
		4	0.42	0.58	0.63	0.71	0.71	0.76	0.77	0.79
	32	1	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98
		2	0.91	0.94	0.96	0.96	0.98	0.98	0.98	0.98
		4	0.80	0.90	0.94	0.95	0.97	0.97	0.97	0.97
HQ	8	1	0.85	0.85	0.85	0.85	0.98	0.98	0.98	0.98
		2	0.54	0.60	0.68	0.73	0.98	0.98	0.97	0.97
		4	0.37	0.49	0.59	0.70	0.96	0.96	0.96	0.96
	16	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		2	0.97	0.99	0.99	0.99	1.00	1.00	1.00	1.00
		4	0.84	0.96	0.97	0.99	0.99	1.00	1.00	1.00
	32	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. CVC = correct attribute vector classification; DINA = deterministic inputs, noisy "and" gate; PWKL = posterior-weighted Kullback-Leibler index; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; UC = unconstrained; H2 = hybrid-2; H1 = hybrid-1; C = constrained; IQ = item quality; J = test length; J_s = block size; LQ = low-quality; HQ = high-quality.

worst classification accuracy regardless of the block size, test length, and item quality except on the 32-item test with the HQ item conditions where the classification accuracy was perfect. However with the GDI, using different blocking versions with the GDI resulted in different classification accuracies based on the factors; however, those differences were mostly negligible. In the following section, the impact of the factors (block size, test length, and item quality) on the CVC rates is discussed.

4.2.2.1.1 The Impact of the Block Size

4.2.2.1.1.1 Short Tests with LQ Items

For short tests with LQ items, increasing the block size generally resulted in lower classification rates regardless of the blocking version and item selection index except

Table 4.2: The CVC Rates Using the DINO Model

			PWKL				GDI			
IQ	J	J_s	UC	H2	H1	C	UC	H2	H1	C
LQ	8	1	0.42	0.42	0.42	0.42	0.59	0.59	0.59	0.59
		2	0.26	0.31	0.36	0.40	0.53	0.53	0.53	0.53
		4	0.20	0.28	0.31	0.40	0.52	0.51	0.48	0.46
	16	1	0.74	0.74	0.74	0.75	0.84	0.84	0.84	0.84
		2	0.58	0.65	0.70	0.74	0.83	0.83	0.82	0.83
		4	0.45	0.59	0.67	0.73	0.78	0.80	0.78	0.79
	32	1	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98
		2	0.92	0.94	0.96	0.97	0.98	0.98	0.98	0.98
		4	0.81	0.91	0.95	0.96	0.97	0.97	0.97	0.97
HQ	8	1	0.85	0.85	0.85	0.85	0.99	0.99	0.99	0.99
		2	0.61	0.69	0.73	0.78	0.98	0.98	0.98	0.98
		4	0.45	0.60	0.65	0.74	0.96	0.95	0.96	0.97
	16	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		2	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00
		4	0.86	0.97	0.97	0.99	1.00	1.00	1.00	1.00
	32	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. CVC = correct attribute vector classification; DINA = deterministic inputs, noisy "and" gate; PWKL = posterior-weighted Kullback-Leibler index; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; UC = unconstrained; H2 = hybrid-2; H1 = hybrid-1; C = constrained; IQ = item quality; J = test length; J_s = block size; LQ = low-quality; HQ = high-quality.

for several conditions. First, for the DINA model and the PWKL, increasing the block size from one to two resulted in substantial differences for the unconstrained and hybrid-2 versions, and moderate differences for the hybrid-1 and constrained versions. For the DINO model and the PWKL, this increase resulted in substantial differences for the unconstrained, hybrid-1, and hybrid-2 versions; the differences were negligible for the constrained version. For example, in the DINA model, the differences were 0.13 and 0.11 for the unconstrained and hybrid-2 versions, and 0.05 and 0.08 for the hybrid-1 and constrained versions, respectively. Moreover, for the DINA model, increasing the block size from two to four resulted in moderate differences for the unconstrained and hybrid-2 versions; however, the differences were negligible when the constrained and hybrid-1 versions were used. For the DINO model, increasing the size from two

Table 4.3: The CVC Rates Using the A-CDM

			PWKL				GDI			
IQ	J	J_s	UC	H2	H1	C	UC	H2	H1	C
LQ	8	1	0.46	0.46	0.46	0.46	0.50	0.50	0.50	0.50
		2	0.45	0.45	0.45	0.45	0.50	0.45	0.50	0.45
		4	0.45	0.44	0.45	0.42	0.39	0.44	0.47	0.38
	16	1	0.81	0.81	0.81	0.81	0.79	0.79	0.79	0.79
		2	0.79	0.78	0.79	0.77	0.78	0.77	0.78	0.78
		4	0.73	0.73	0.70	0.70	0.74	0.73	0.73	0.73
	32	1	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
		2	0.97	0.96	0.97	0.96	0.97	0.96	0.97	0.96
		4	0.96	0.93	0.95	0.92	0.96	0.93	0.95	0.93
HQ	8	1	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
		2	0.93	0.93	0.96	0.93	0.97	0.96	0.97	0.96
		4	0.72	0.85	0.82	0.90	0.96	0.95	0.96	0.95
	16	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00
		4	0.99	0.99	0.99	0.99	1.00	1.00	0.99	0.99
	32	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. CVC = correct attribute vector classification; DINA = deterministic inputs, noisy "and" gate; PWKL = posterior-weighted Kullback-Leibler index; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; UC = unconstrained; H2 = hybrid-2; H1 = hybrid-1; C = constrained; IQ = item quality; J = test length; J_s = block size; LQ = low-quality; HQ = high-quality.

to four resulted in moderate differences for the unconstrained and hybrid-1 versions; however, the increase resulted in negligible differences for the constrained and hybrid-2 versions. For the A-CDM and the PWKL, increasing the block size from two to four resulted in negligible differences in the CVC rates regardless of the blocking version except for the constrained version. In that blocking version, the increase resulted in a moderate difference.

For the DINA model and the GDI, increasing the block size from one to two resulted in negligible differences in the CVC rates regardless of the blocking version. However, for the DINO model and the GDI, the same increase in the block size resulted in moderate differences in the CVC rates regardless of the blocking version.

For example, using the unconstrained version with the GDI, the differences were 0.01 and 0.06 for the DINA and DINO models, respectively. Moreover, for the DINA model, increasing the block size from two to four resulted in moderate differences regardless of the blocking version. For the DINO model, the differences were moderate for the constrained and hybrid-1 versions and negligible for the unconstrained and hybrid-2 versions. For the *A*-CDM and the GDI, increasing the block size from one to two resulted in moderate differences in the CVC rates for the constrained and hybrid-1 versions and negligible differences for the unconstrained and hybrid-2 versions. For example, the differences were 0.05 for the constrained and hybrid-1 versions and 0.00 for the unconstrained and hybrid-2 versions. Finally, increasing the block size from two to four resulted in moderate differences in the CVC rates for the unconstrained and constrained versions, and negligible differences for the hybrid-1 and hybrid-2 versions.

4.2.2.1.1.2 Medium-Length Tests with LQ Items

For the medium-length tests with LQ items, increasing the block size resulted in lower classification rates for the PWKL when the DINA and DINO models were used, and negligible differences when the *A*-CDM was used as the generating models. However, increasing the block size resulted in negligible to moderate differences in the classification rates for the GDI regardless of the blocking version and generating model except for several conditions. First, for the DINA and DINO models and the PWKL, increasing the block size resulted in substantial differences in the CVC rates for the unconstrained version, moderate differences for the hybrid-1 and hybrid-2 versions, and negligible differences for the constrained version. The differences were 0.17, 0.02, 0.05, and 0.10 for the unconstrained, constrained, hybrid-1, and hybrid-2 versions in the DINA model, respectively. For the *A*-CDM and the PWKL, increasing the block size from one to two resulted in negligible differences in the CVC rates for three

blocking versions. However, for the constrained version, the difference was moderate. Increasing the block size from two to four resulted in moderate differences regardless of the blocking version.

For the DINA model and the GDI, increasing the block size resulted in negligible differences in the CVC rates for the constrained and hybrid-1 versions and moderate differences for the unconstrained and hybrid-2 versions. For the DINO model and the GDI, increasing the block size resulted in negligible differences for the unconstrained and hybrid-2 versions and moderate differences for the constrained and hybrid-1 versions. For the *A*-CDM and the GDI, increasing the block size from one to two resulted in negligible differences in the CVC rates regardless of the blocking version; however, increasing the block size from two to four resulted in moderate differences regardless of the blocking version.

4.2.2.1.1.3 Long Tests with LQ Items

For long tests with LQ items, increasing the block size resulted in negligible differences in the CVC rates regardless of the blocking version, generating model, and item selection index except for several conditions. First, for the DINA and DINO models using the PWKL, the unconstrained version resulted in moderate differences when the block size was increased from one to two and substantial differences when the block size was increased from two to four. For the *A*-CDM and the PWKL, the constrained version yielded moderate differences when the block size was increased from two to four.

4.2.2.1.1.4 Short Tests with HQ Items

For short tests with HQ items, increasing the block size resulted in moderate to substantial differences in the classification rates when the PWKL was used; however, increasing the block size resulted in negligible differences when the GDI was used. Several additional findings should be noted. For the DINA model and the PWKL,

increasing the block size from one to two resulted in substantial differences for all four blocking versions. For the DINO model and the PWKL, increasing the block size resulted in substantial differences for the unconstrained, hybrid-1, and hybrid-2 versions; for the constrained version, the difference was moderate. For example, for the DINA and models and the unconstrained version, the differences were 0.31 and 0.24, respectively.

For the DINA model and the PWKL, increasing the block size from two to four resulted in substantial differences for the unconstrained and hybrid-2 versions, moderate differences for the hybrid-1 version, and negligible differences for the constrained version. For the DINO model and the PWKL, the same size increase resulted in substantial differences for the unconstrained version and moderate differences for the hybrid-2, hybrid-1, and constrained versions.

For the *A*-CDM and the PWKL, increasing the block size from one to two resulted in moderate differences in the CVC rates for the unconstrained, hybrid-2, and constrained versions, and in a negligible difference for the hybrid-1 version. In addition, increasing the block size from two to four yielded substantial differences for the unconstrained and hybrid-1 versions, moderate differences for the hybrid-2 version, and negligible differences for the constrained version.

4.2.2.1.1.5 Medium-Length and Long Tests with HQ Items

For medium-length and long tests with HQ items, increasing the block size resulted in negligible differences in the classification rates regardless of the blocking version, generating model, and item selection index, except for the 16-item test involving the PWKL with the unconstrained version. Increasing the block size from two to four resulted in substantial differences for the DINA and DINO models. The differences were 0.13 and 0.11 for the DINA and DINO models, respectively.

4.2.2.1.2 The Impact of the Test Length

4.2.2.1.2.1 LQ Items

As expected, increasing the test length resulted in substantial increases in the classification rates regardless of the blocking version, generating model, and block size. Moreover, the increases for the PWKL were greater than those for the GDI. For example, for the DINA model with the block size of one, increasing the test length from 8 to 16 resulted in 0.34 and 0.30 increases in the CVC rates for the PWKL and the GDI, respectively. Although the PWKL had higher augmentation in the CVC rates, the GDI still had higher classification accuracy when LQ items were used.

4.2.2.1.2.2 HQ Items

For a small block (i.e., J_s=1 and 2), increasing the test length resulted in negligible differences in the classification rates regardless of the blocking version, generating model, and item selection index except for the DINA and DINO models with the PWKL regardless of the blocking version. For the *A*-CDM with the PWKL, the differences were substantial for the unconstrained, hybrid-2, and constrained versions. In addition, for the *A*-CDM, the hybrid-2 and constrained versions resulted in moderate differences when small blocks were used. For the DINA and DINO models with the PWKL, increasing the test length from 8 to 16 resulted in substantial differences (i.e., 0.43) for the unconstrained version when the block size was two.

For a large block (i.e., J_s=4) and the PWKL, increasing the test length from 8 to 16 resulted in substantial differences in the classification rates regardless of the blocking version and generating model, except for the constrained version using the *A*-CDM−the difference was moderate. However, for a large block with the GDI, the same increase in the test length resulted in negligible to moderate increases in the classification rates. For the DINA model and the GDI, the hybrid-1, hybrid-2, and constrained versions resulted in moderate differences, and the unconstrained

version resulted in negligible differences. For the DINO model and the same index, the unconstrained and hybrid-2 versions resulted in moderate differences, and the constrained and hybrid-1 versions resulted in negligible differences. For the *A*-CDM and the same index, the unconstrained, hybrid-2, and constrained versions resulted in moderate differences, whereas the hybrid-1 version resulted in negligible differences.

For a large block, increasing the test length from 16 to 32 resulted in negligible differences in the classification rates regardless of the blocking version, generating model, and item selection index, except for the DINA and DINO models and using the PWKL with the unconstrained version, where the differences were substantial.

4.2.2.1.3 The Impact of the Item Quality

As expected, using HQ items instead of LQ items resulted in substantial differences in the classification rates when the test length was shorter (i.e., 8- and 16-item tests) regardless of the blocking version, generating model, and item selection index. However, for long tests (i.e., 32-item tests), varying results were observed.

For a small block (i.e., J_s=1), using HQ items instead of LQ items resulted in negligible differences in the classification rates regardless of the blocking version, generating model, and item selection index. For large blocks (i.e., J_s=2 and 4) and the DINA and DINO models, using HQ items resulted in moderate differences for the PWKL, except for the unconstrained version, where the difference was substantial when the block size was four. Moreover, for the same block size and models, the GDI yielded negligible differences in the CVC rates regardless of the blocking version. For the *A*-CDM, when the block size was two, using HQ items resulted in moderate differences for the unconstrained and hybrid-1 versions and negligible differences for the hybrid-2 and constrained versions regardless of the item selection index. Last, for the *A*-CDM, using HQ items yielded moderate differences regardless of the blocking version and item selection index when the block size was two.

4.2.2.2 Item Usage

To get a deeper understanding of the differences in item usage across the different blocking versions, items were grouped based on their required attributes. An additional simulation study was carried out using the same factors except for one: item quality. For this study, the lowest and highest success probabilities were fixed across all of the items, specifically, $P(\mathbf{0})$=0.1 and $P(\mathbf{1})$=0.9. This design aimed to eliminate the effect of item quality on item usage. The test administration was divided into periods that each compared four items. The item usage was recorded in each period. Only the results for the GDI, 8-item tests, and $\boldsymbol{\alpha}_3$ using the unconstrained and hybrid-1 versions are shown in Figures 4.2, 4.4, and 4.6, and using the hybrid-2 and constrained versions are shown in Figures 4.3, 4.5, and 4.7 for the DINA model, the DINO model, and the A-CDM, respectively.

In the first period, which includes the first four items, single attribute items were mostly used regardless of the blocking version, generating model, and block size. For a small block (i.e., J_s=1), single attribute items, whose q-vectors were different, were mostly administered in the first period. Because the uniform distribution was used as before for each blocking version and item selection index at the beginning of the test, the four single attribute items were the same regardless of the blocking version and generating model when the block size was one. For example, items with the q-vectors of (0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0), and (0,0,0,0,1) were used in the first period for each blocking version and generating model when the block size was one. However, for large blocks (i.e., J_s=2 and J_s=4), the blocking versions resulted in different item types. For example, the unconstrained and hybrid-2 versions used two types of single attribute items (e.g., items whose q-vectors were (0,0,1,0,0) and (0,0,0,1,0)) when the block size was two, and only one type of single attribute item (e.g., items whose q-vector was (0,0,1,0,0)) when the block size was four regardless of the generating

model. Moreover, because of the constraint, the hybrid-1 and constrained versions used four single attribute items, whose q-vectors were different, in the first period regardless of the generating model.

In the second period, item usage differed based on the blocking versions, generating model, and block size. When the block size was one, the item usage patterns were similar to those observed in the first part of the study. For example, the DINA model showed the following pattern for item usage: The model used items that required single attributes which were not mastered by the examinee (e.g., items whose q-vectors were (0,0,0,1,0) with 10% and (0,0,0,0,1) with 8% usage) and items that required the same attributes as the examinee's true attribute mastery vector (e.g., items whose q-vectors were (1,1,1,0,0) with 8% usage).

The DINO model showed the following pattern of item usage: The model used items that required single attributes which were mastered by the examinee (e.g., items whose q-vectors were (1,0,0,0,0) with 13%, (0,1,0,0,0) with 8%, and (0,0,1,0,0) with 10% usage) and items that required the same attributes as the examinee's true attribute nonmastery vector (e.g., items whose q-vectors were (0,0,0,1,1) with 8% usage). The *A*-CDM used items that required single attributes regardless of the true attribute vector. In addition to the item usage in each model, the single attribute item with the q-vector of (1,0,0,0,0) was used 13% of the time regardless of the blocking version and generating model in the second period.

When the block size was two and four, the blocking versions resulted in different item usage patterns. The unconstrained version used only single attribute items for the large block regardless of the generating model. For example, the DINA model mostly used items whose q-vectors were (0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0), and (0,0,0,0,1) when the block size was two, and items with (0,0,1,0,0) and (0,0,0,1,0) when the block size was four. The hybrid-2 version mostly used single attribute items in addition to the two-attribute items when the block size was larger for the

DINA and DINO models. For example, the DINA and DINO models used all single attribute items and items with the q-vector of (1,0,1,0,0) when the block size was two. The hybrid-1 and constrained versions yielded the same item usage patterns for the generating model when the block size was two. However, it used only one type of single attribute items when the block size was four. Again, the *A*-CDM used only single attribute items regardless of the blocking version and block size.

In addition, the unconstrained version used certain item types for a certain block size regardless of the generating model. For example, when the block size was two, the most commonly used items were (0,0,1,0,0) and (0,0,0,1,0) in the first period, and (0,1,0,0,0) and (0,0,0,0,1) in the second period; when the block size was four, the items were (0,0,0,1,0) in the first period and (0,0,1,0,0) in the second period regardless of the generating model. In other words, as expected, different types of one-attribute items were used in different periods because a block of items was administered at a time, and the item selection index tended to administer only one-attribute items until it can obtain enough information to proceed to the other item types.

Longer test lengths (i.e., 16- and 32-item tests) yielded similar item usage patterns in the first period as on the 8-item test. Moreover, in the last periods, the blocking versions yielded similar item usage patterns for the generating models, except for the block size of four in which different types of items were used because of the constraint.

4.3 Discussion and Conclusion

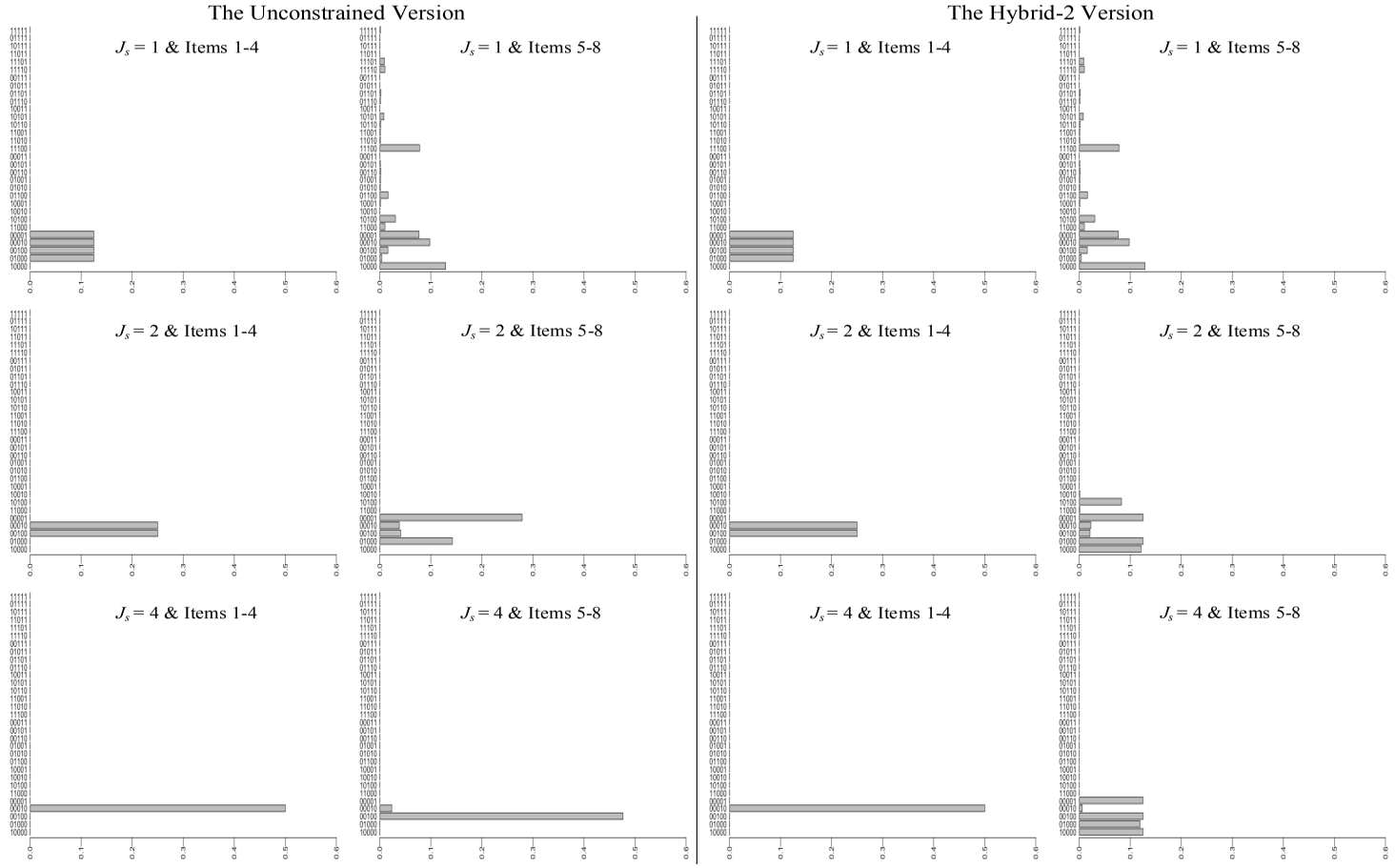
Item review and answer change have several benefits for test takers such as reduced test anxiety, the opportunity to correct careless errors, and, most importantly, increased testing validity. However, these options have several drawbacks, including decreased testing efficiency and demand of more complicated item selection algorithms. In a blocked-design CAT, item review was allowed within a block of items,

and several studies showed that there was no significant difference in the accuracy of the ability estimated with limited review and no review procedures. Another procedure that allows item review and answer change is MST in which test adaption occurs at the sets of item level instead of the item level. In this paper, a new CD-CAT procedure was proposed to allow item review and answer change during test administration. In this procedure, a block of items was administered with and without a constraint on the q-vectors of the items. Different from MST, content balancing and item difficulty were not applicable in the new procedure. Based on the factors in the simulation study, using the new procedure with the GDI is promising for item review especially with HQ items and long tests without too large decrease in the classification accuracy. In addition, the different blocking versions yielded similar classification rates. However, the constrained version with the PWKL had the best classification accuracy, whereas the unconstrained version with the PWKL had the worst classification accuracy regardless of the block size, test length, and item quality except on long tests with HQ items. The results of this study suggest several findings that are of practical value. First, it is not advisable to use the PWKL with the blocked-design CD-CAT especially with larger block sizes because of the substantial decrease in the classification rates across many conditions. Second, from this study, the practitioners, so as to allow students to review and change their answers, can determine the tolerable level of loss in classification accuracy in deciding the appropriate block size to be used. Last, item usage patterns revealed in this study can be helpful in test construction strategies in the context of cognitive diagnosis.

Although this study showed promise with respect to item review for CD-CAT, more research must be conducted to determine the viability of the blocked-design CD-CAT. First, only a single constraint on the q-vectors was considered in the current study; however, it would be interesting to examine different possible constraints (e.g., hierarchical structures) on items. Second, further research needs to be done in

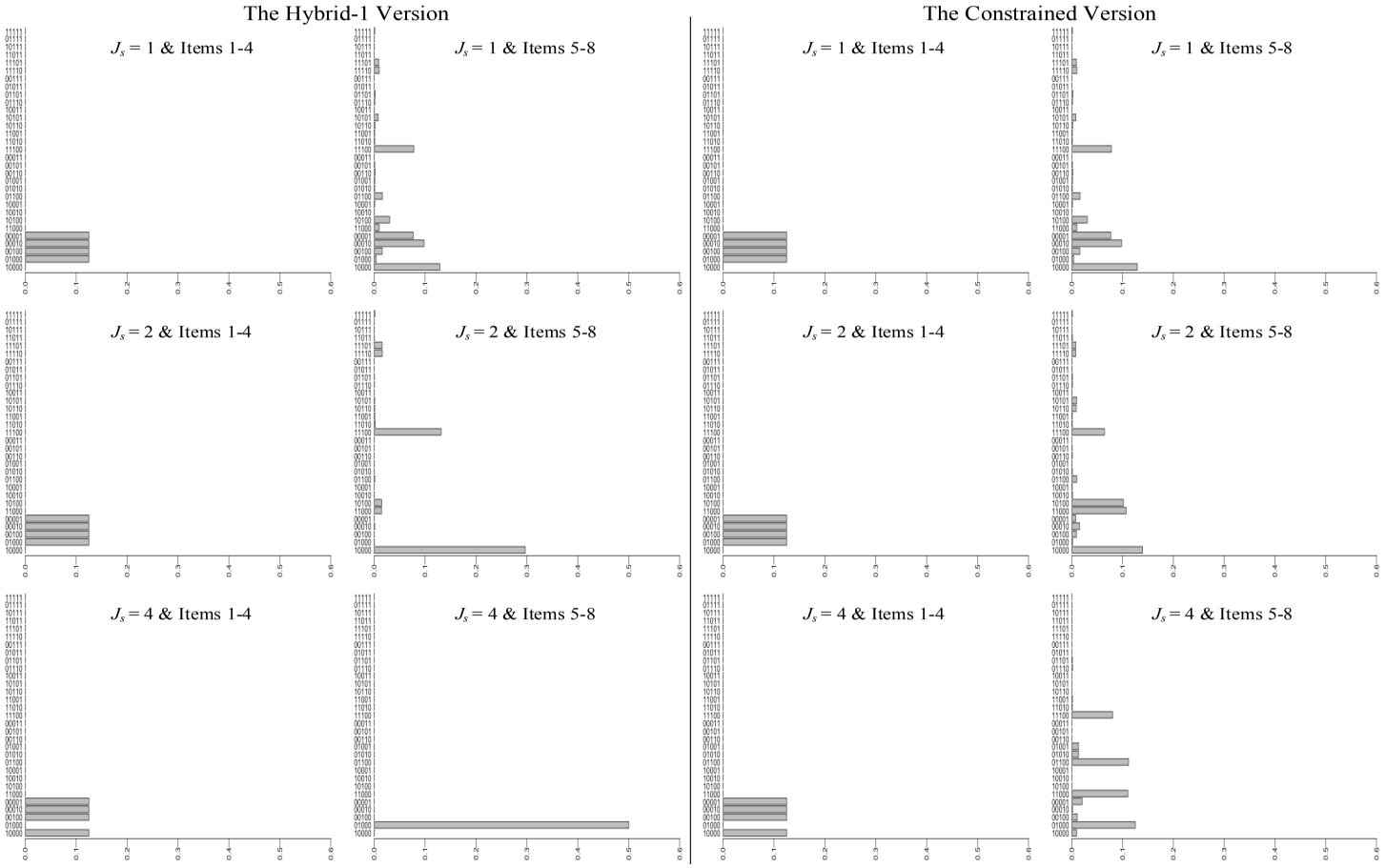
the multistage applications for cognitive diagnosis. For example, CDMs are multidimensional models, and there is no difficulty parameter for every relevant dimension. Therefore, it is still challenging to construct the blocks in MST for cognitive diagnosis. Third, the impact of the number of attributes and item pool size was not considered; these factors also affect the performance of the indices in real CAT applications. Last, the data sets were generated using a single reduced CDM. It would be more practical to examine the use of a more general model, which allows the item pool to be made up of various CDMs.

Figure 4.2: The Proportion of Item Usage for the Unconstrained and Hybrid-2, DINA, $\boldsymbol{\alpha}_3$, GDI, and $J = 8$



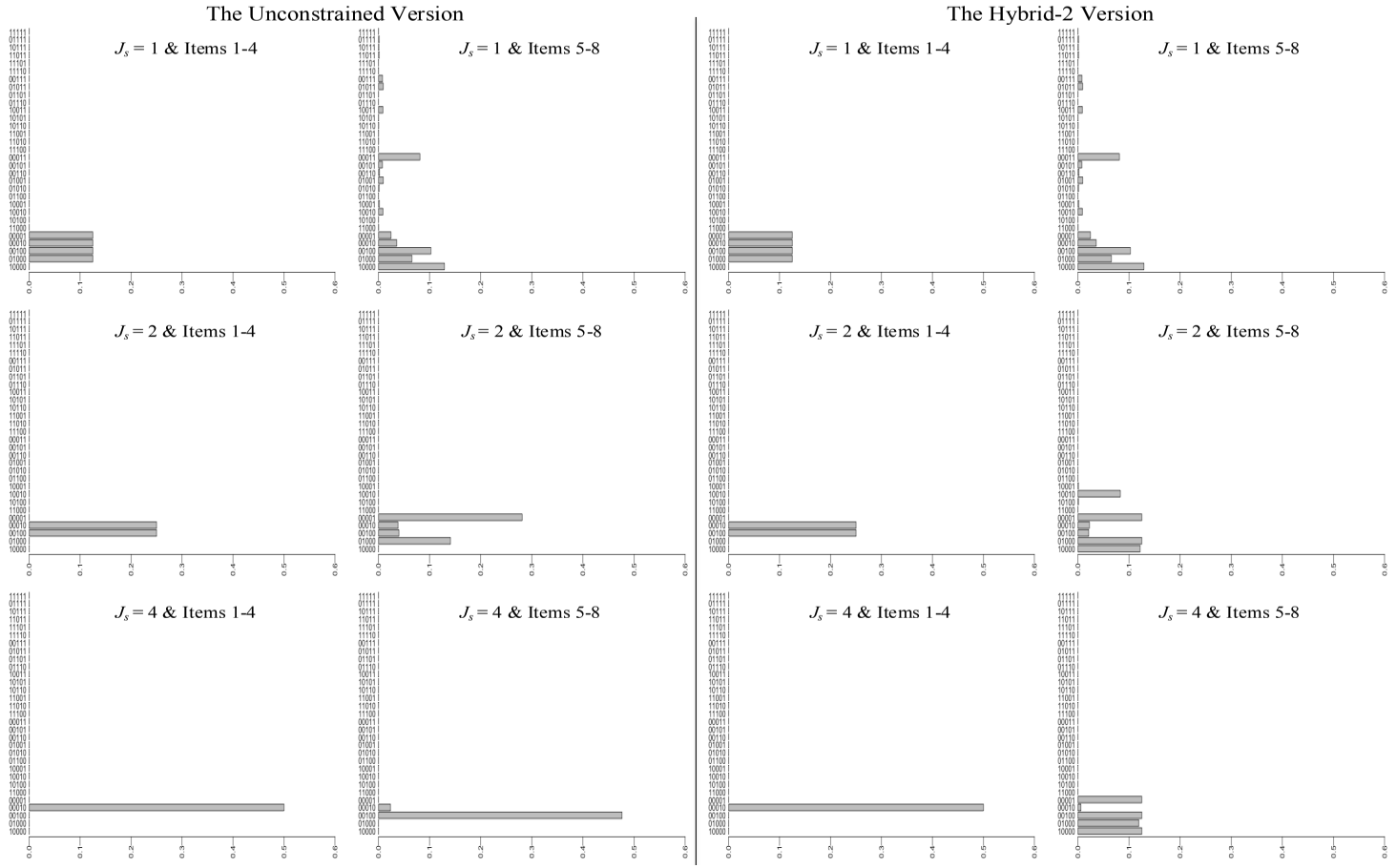
Note: DINA = deterministic inputs, noisy "and" gate; GDI = G-DINA model discrimination index; G-DINA = generalized DINA.

Figure 4.3: The Proportion of Item Usage for the Hybrid-1 and Constrained, DINA, $\boldsymbol{\alpha}_3$, GDI, and $J = 8$



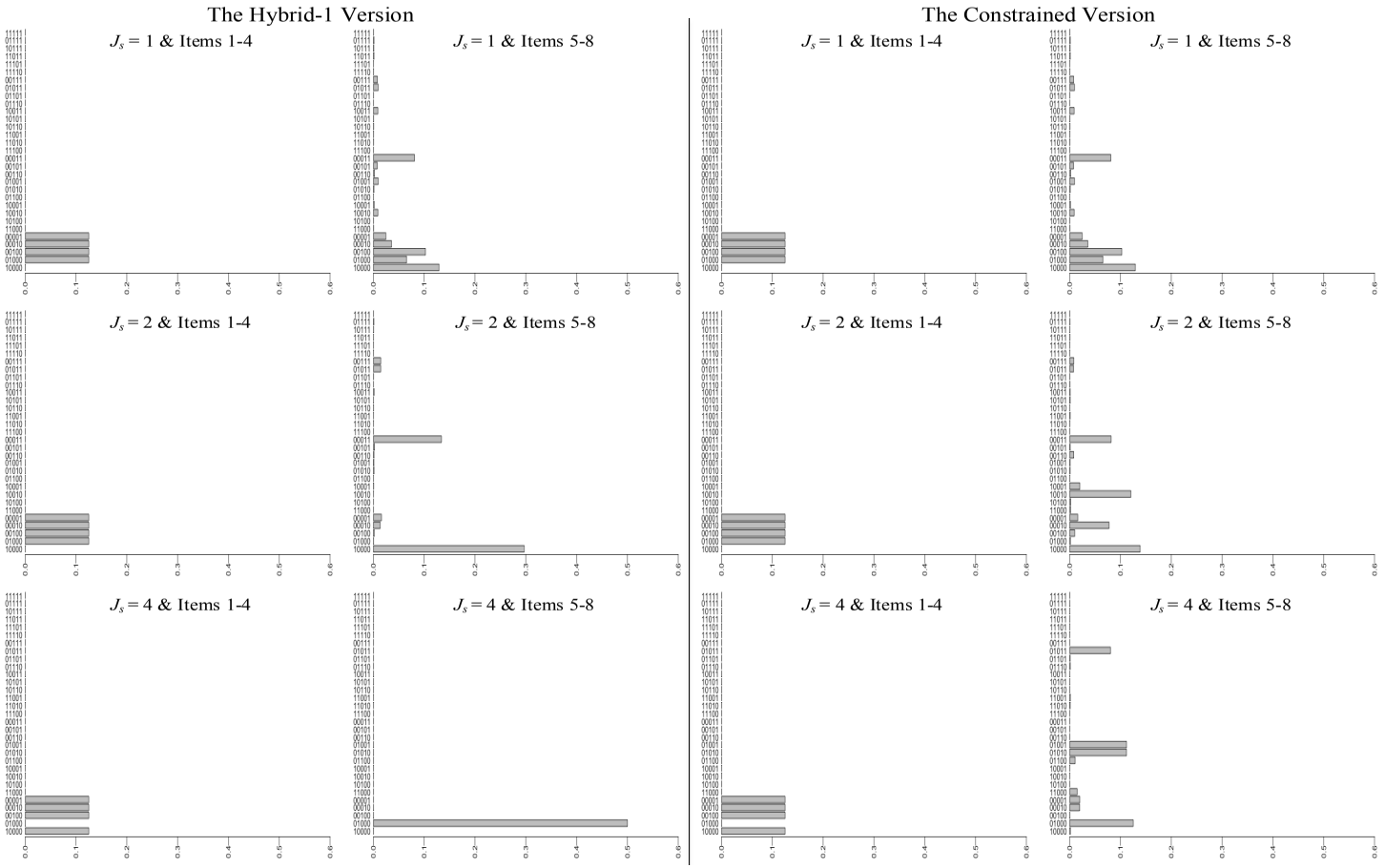
Note: DINA = deterministic inputs, noisy "and" gate; GDI = G-DINA model discrimination index; G-DINA = generalized DINA.

Figure 4.4: The Proportion of Item Usage for the Unconstrained and Hybrid-2, DINO, $\boldsymbol{\alpha}_3$, GDI, and $J = 8$



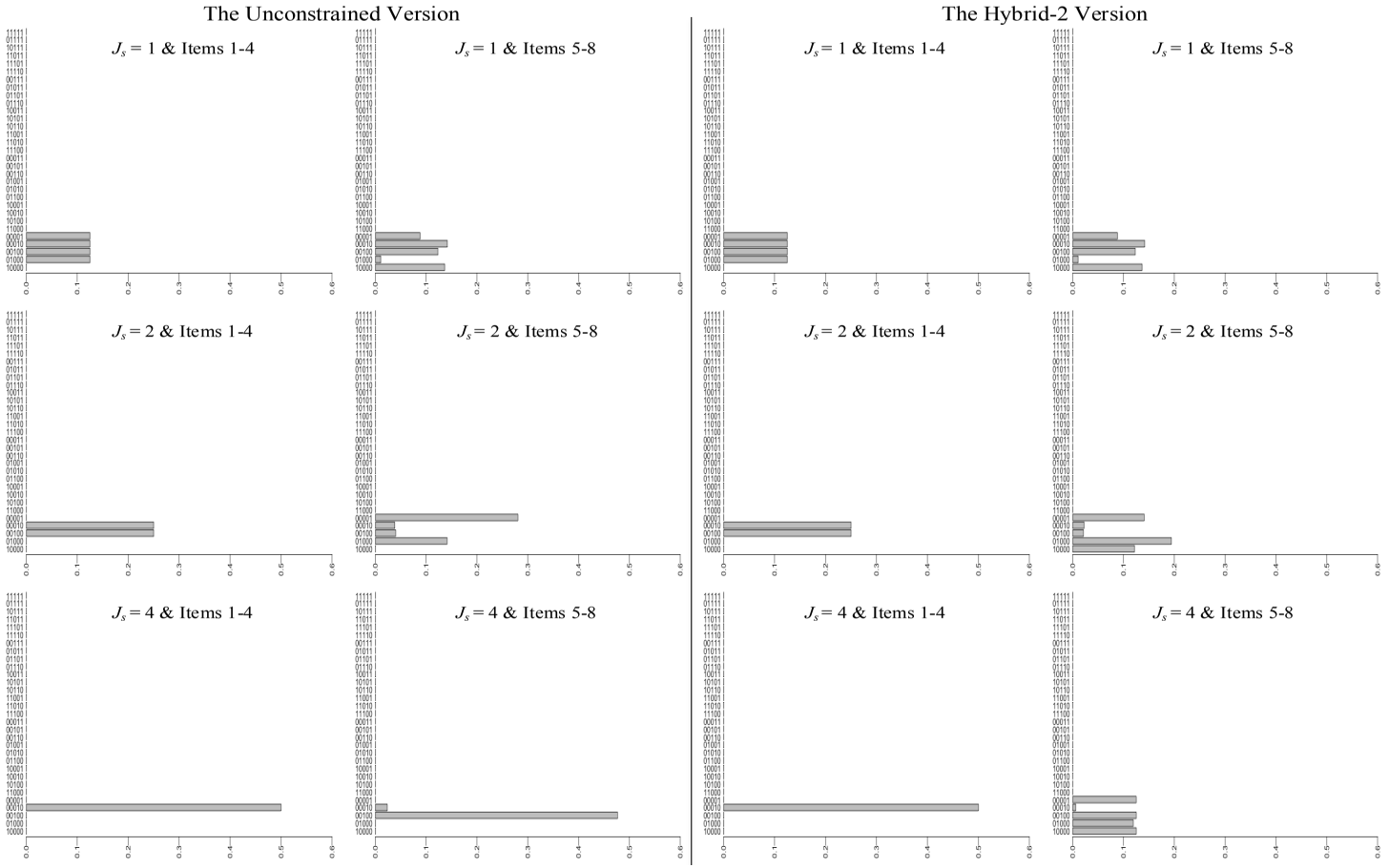
Note: DINO = deterministic input, noisy "or" gate; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; DINA = deterministic inputs, noisy "and" gate.

Figure 4.5: The Proportion of Item Usage for the Hybrid-1 and Constrained, DINO, $\boldsymbol{\alpha}_3$, GDI, and $J = 8$



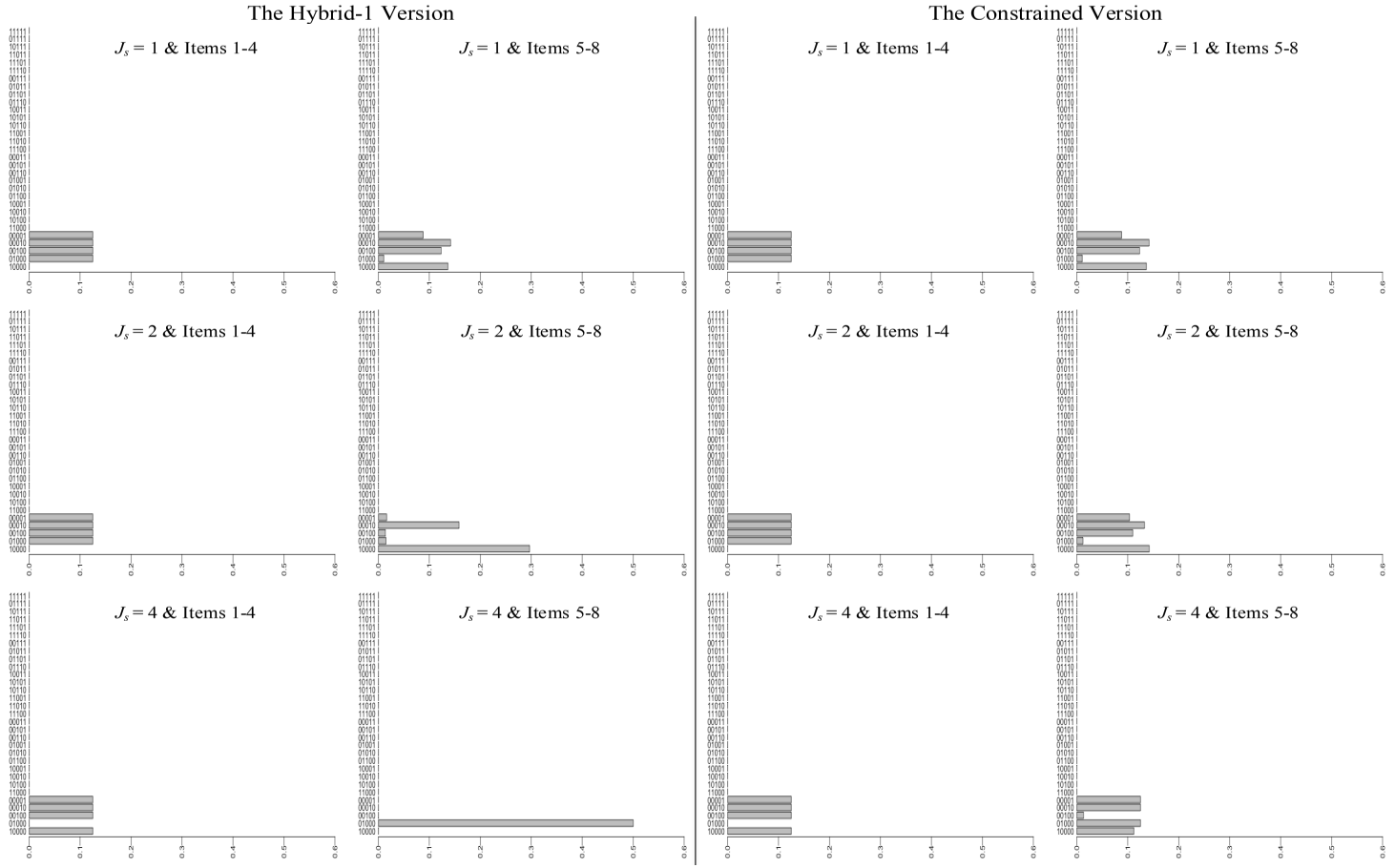
Note: DINO = deterministic input, noisy "or" gate; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; DINA = deterministic inputs, noisy "and" gate.

Figure 4.6: The Proportion of Item Usage for the Unconstrained and Hybrid-2, A-CDM, $\boldsymbol{\alpha}_3$, GDI, and $J = 8$



Note: A-CDM = additive CDM; CDM = cognitive diagnosis model; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; DINA = deterministic inputs, noisy "and" gate.

Figure 4.7: The Proportion of Item Usage for the Hybrid-1 and Constrained, A-CDM, $\boldsymbol{\alpha}_3$, GDI, and $J = 8$



Note: A-CDM = additive CDM; CDM = cognitive diagnosis model; GDI = G-DINA model discrimination index; G-DINA = generalized DINA; DINA = deterministic inputs, noisy "and" gate.

References

Benjamin, L. T., Cavell, T. A., & Schallenberger, W. R. I. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology, 11*, 133-141.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619-632.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York, NY: John Wiley.

Crocker, L., & Benson, J. (1980). Does answer-changing affect test quality? *Measurement and Evaluation in Guidance, 12*, 233-239.

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

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

de la Torre, J., & Chiu, C.-Y. (2015). A general method of empirical Q-matrix validation. *Psychometrika*. Advance online publication. doi:10.1007/s11336-015-9467-8

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

Gershon, R. C., & Bergstrom, B. (1995). *Does cheating on CAT pay?* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

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

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.

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

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*, 44-52.

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

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

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*, 167-188.

Kingsbury, G. (1996). *Item review and adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Legg, S., & Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice, 11*, 23-27.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York: Springer.

Liu, O. L., Bridgeman, B., Lixiong, G., Xu, J., & Kong, N. (2015). Investigation of response changes in the GRE revised general test. *Educational and Psychological Measurement*, Advance online publication. doi:10.1177/0013164415573988.

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

Mathews, C. O. (1929). Erroneous first impressions on objective tests. *Journal of Educational Psychology, 20*, 280-286.

McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40*, 808-821.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*, 187-194.

Mueller, D. J., & Wasser, V. (1977). Implications of changing answers on objective test items. *Journal of Educational Measurement, 14*, 9-13.

Olea, J., Revuelta, J., Ximenez, M. C., & Abad, F. J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. *Psicologica, 21*, 157-173.

Papanastasiou, E. C., & Reckase, M. D. (2007). A rearrangement procedure for scoring adaptive test with review options. *International Journal of Testing, 7*, 387-407.

Revuelta, J., Ximenez, M. C., & Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. *Educational and Psychological Measurement, 63*, 791-808.

Robin, F., Steffen, M., & Liang, L. (2014). The multistage test implementation of the GRE revised general test. In Y. Duanli, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (p. 325-342). Boca Raton: CRC Press.

Smith, M., White, K., & Coop, R. (1979). The effect of item type on the consequences of changing answers on multiple choice tests. *Journal of Educational Measurement, 16*, 203-208.

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

Stone, G. E., & Lunz, M. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. *Applied Measurement in Education, 7*, 211-222.

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

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

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer et al. (Ed.). *Computerized adaptive testing: A primer* (pp. 101-133). Hillsdale: Erlbaum.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3-30). Boston, MA: Kluwer.

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self adaptive vocabulary tests. *Journal of Educational Measurement, 35*, 328-345.

Vispoel, W. P. (2000). Reviewing and changing answers on computerized fixed-item vocabulary tests. *Educational and Psychological Measurement, 60*, 371-384.

Vispoel, W. P., Clough, S. J., & Bleiler, T. (2005). A closer look at using judgments of item difficulty to change answers on computerized adaptive tests. *Journal of Educational Measurement, 42*, 331-350.

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

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

Vispoel, W. P., Rocklin, T., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adaptive testing. *Applied Measurement in Education, 7*, 53-79.

Vispoel, W. P., Rocklin, T., Wang, T., & 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*, 141-157.

Vispoel, W. P., Wang, T., de la Torre, R., Bleiler, T., & Dings, J. (1992). *How review options, administration mode, and test anxiety influence scores on computerized vocabulary tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

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

von Davier, M., & Cheng, Y. (2014). Multistage testing using diagnostic models. In Y. Duanli, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (p. 219-227). Boca Raton: CRC Press.

Waddell, D. L., & Blankenship, J. C. (1995). Answer changing: A meta-analysis of the prevalence and patterns. *Journal of Continuing Education in Nursing, 25*, 155-158.

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

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wise, S. L. (1996). *A critical analysis of the arguments for and against item review in computerized adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Wise, S. L., Finney, S., Enders, C., Freeman, S., & Severance, D. (1999). Examinee judgments of changes in item difficulty: Implications for item review in computerized adaptive testing. *Applied Measurement in Education, 12*, 185-198.

Xu, X., Chang, H.-H., & Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Xu, X., & Douglas, J. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika, 71*, 121-137.

Yen, Y.-C., Ho, R.-G., Liao, W.-W., & Chen, L.-J. (2012). Reducing the impact of inappropriate items on reviewable computerized adaptive testing. *Educational Technology and Society, 15*, 231-243.

Chapter 5

Summary

Compared to unidimensional item response theory (IRT) models, cognitive diagnosis models (CDMs) provide more detailed evaluations of students' strengths and weaknesses in a particular content area and, therefore, provide more information that can be used to inform instruction and learning (de la Torre, 2009). Computerized adaptive testing (CAT) has been developed as an alternative tool to paper-and-pencil tests and can be used to create tests tailored to each examinee (Meijer & Nering, 1999; van der Linden & Glas, 2002). CAT procedures are generally built on IRT models; however, different psychometric models (i.e., CDMs) can also be used in CAT procedures. Considering the advantages of CAT, the use of CDMs in CAT can provide better diagnostic evaluations with more accurate estimates of examinees' attribute vectors.

At present, most of the research in CAT has been performed in the context of IRT; however, a small number of studies have recently been conducted in CD-CAT. One reason the research on CD-CAT is limited is that some of the concepts in traditional CAT (i.e., Fisher information) cannot be applied in CD-CAT because of the discrete nature of attributes. With a general aim to address needs in formative assessments, this dissertation introduced new item selection indices that can be used in CD-CAT, showed the use of item exposure control methods with one of the new indices, proposed an alternative CD-CAT administration procedure in which examinees have the benefit of item review and answer change options, and introduced a more efficient simulation design that can be generalized to different distributions of attribute vectors, despite

involving a smaller sample size.

In the first study, two new item selection indices, the modified posterior-weighted Kullback-Leibler index (MPWKL) and the generalized deterministic inputs, noisy "and" gate (G-DINA) model discrimination index (GDI), were introduced for CD-CAT. The efficiency of the indices was compared with the posterior-weighted Kullback-Leibler index (PWKL). The results showed that compared to the PWKL, the MPWKL and the GDI performed very similarly and had higher attribute classification rates or shorter mean test lengths depending on the test termination rule. Moreover, item quality had an obvious impact on the classification rates: Higher discrimination and higher variance resulted in higher classification accuracy. Thus, the combination of higher-discriminating items with higher variance had the best classification accuracy and/or shortest test lengths, whereas low-discriminating items with lower variance had the worst classification accuracy and/or longest test lengths regardless of the item selection index and the generating model. Moreover, generating models can affect the efficiency of the indices: For the DINA and DINO models, the results were more distinguishable; however, the efficiency of the indices was essentially the same for the *A*-CDM, except in a few conditions.

To get a deeper understanding of the differences in item usage among the models, the items were grouped based on their required attributes and item usage in terms of the number of required attributes recorded for each condition. Overall, the DINA model showed the following pattern of item usage: The model used items that required the same attributes as the examinee's true attribute mastery vector and items that required single attributes that were not mastered by the examinee. In contrast, the DINO model showed a different pattern of item usage: This model used items that required the same attributes as the examinee's true nonmastery vector and items that required single attributes that were mastered by the examinee. The *A*-CDM used items that required single attributes regardless of the true attribute vector. The

GDI had the shortest implementation time among the three indices.

In the second study, the use of two item exposure control methods, restrictive progressive (RP) and restrictive threshold (RT), in conjunction with the GDI was introduced. When new item selection indices are proposed in CAT, the measurement accuracy and the test security the indices provide are commonly investigated (Barrada, Olea, Ponsoda, & Abad, 2008). Typically, high item exposure rates accompany efficient item selection indices, and it is crucial to decrease the use of overexposed items and increase the use of underexposed items. In this study, the efficiency of the GDI was investigated in terms of the classification accuracy and the item exposure using the RP and RT methods. Based on the factors manipulated in the simulation study, as expected, the RP method resulted in more uniform item exposure rates and higher classification rates compared to the RT method. Moreover, the factors, including the item quality, test length, pool size, prespecified desired exposure rate, and β, generally had a substantial impact on the exposure rates when the RP method was used; however, fewer factors, such as the pool size, prespecified desired exposure rate, and β, generally had a substantial impact on the exposure rates when the RT method was used. The other factors had moderate or negligible effects on the item exposure rates with some exceptions.

In the third study, a new CD-CAT administration procedure, where blocks of items are administered, was introduced. Using the new procedure, examinees would be able to review their responses within a block of items. Originally, Stocking (1997) proposed a blocked-design CAT in which item review was allowed within a block of items, and the results showed that there was no significant difference in the accuracy of the ability estimated with limited review and no review procedures. In this study, a block of items was administered with and without a constraint on the q-vectors of the items. Four blocking versions of the new procedure (i.e., unconstrained, constrained, hybrid-1, and hybrid-2) were proposed. Based on the factors in the simulation study,

the constrained version with the PWKL had the best classification accuracy, whereas the unconstrained version with the PWKL had the worst classification accuracy regardless of the block size, test length, and item quality except on long tests with HQ items. However, the differences between the blocking versions were negligible when the GDI was used. Using the new procedure with the GDI is promising for item review especially with HQ items and long tests without too large a decrease in classification accuracy.

In this dissertation, new item selection indices were proposed for CD-CAT that can be used instead of traditional CAT procedures when more detailed evaluations of examinees' strengths and weaknesses are needed. The dissertation's first study was important in understanding how different information statistics can be used as item selection methods in the CAT administration. The second study was useful in examining how to implement item exposure control methods with a new item selection index and what factors should be taken into account when controlling high item exposure rates. The third study was essential in obtaining more accurate validity of tests by providing an adequate opportunity for item review and answer change to examinees. Finally, this dissertation helped deepen our understanding of how different item selection indices behaved in terms of item usage with respect to different CDMs and examinee true attribute vectors using a more efficient simulation design.

A successful realization of these objectives led to a deeper understanding of the CDMs and CAT, and increased the joint applicability of these procedures. Nonetheless, there are still questions that need to be investigated in the context of CD-CAT. For example, in simulation studies, the response data are mostly generated based on a model, and therefore, it provides a perfect model fit. However, it would be interesting to analyze the efficiency of the new indices using real data, especially when the response data do not fit any existing model. In addition, one of the most difficult parts of traditional CAT procedures is the item pool development. This also applies

to CD-CAT procedures. With respect to this point, a successful implementation of CD-CAT depends on several factors, including a well-developed item pool, accurately estimated item parameters, and a well-constructed Q-matrix.

References

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *The British Journal of Mathematical and Statistical Psychology, 61*, 493-513.

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

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*, 187-194.

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

van der Linden, W. J., & Glas, C. A. W. (2002). Preface. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. Vii-Xii). Boston, MA: Kluwer.