LEEMING, MEGHAN K. Ph.D. A Comparison of Item Selection Methods for Multiple-Choice Options-Based CD-CAT. (2025)
Directed by Dr. Robert Henson. 119 pp.

This dissertation aimed to evaluate the performance of the multinomial extension of expected Shannon Entropy compared to two new item selection methods for CD-CAT based on theoretical correct classification rates, along with a random fixed form, and a well-designed fixed form using four different evaluation measures. These methods were evaluated using the attribute-level correct classification rates across all attributes and examinees and attribute profile CCRs, as well as calculating the proportion of items exposed in each item bank, the average absolute deviation of the posterior probabilities from 0.5, and the average time. This study was fully crossed across three conditions: number of attributes, test length, and the quality of the item bank. Consistent with prior research, a key finding was that each of the CD-CAT methods performed similarly, with no more than a 10% improvement when compared to a randomly constructed fixed form test, with a smaller gap in improvement compared to the well-designed fixed form. The multinomial extension of expected Shannon Entropy performed better than the single attribute and composite attribute theoretical CCR methods, as well as both fixed form methods across all measures. Due to the high computational demand of the multinomial extension of expected Shannon Entropy, the composite attribute theoretical CCR method is a more efficient choice for balancing performance and time. Recommendations were discussed for future research to vary more conditions to have broader applications to real data and assessments.

A COMPARISON OF ITEM SELECTION METHODS FOR MULTIPLE-CHOICE OPTIONS-BASED CD-CAT

by

Meghan K. Leeming

A Dissertation
Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro

2025

Approved by

Dr. Robert Henson
Committee Chair

APPROVAL PAGE

This dissertation written by Meghan K. Leeming has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair

Dr. Robert Henson

Committee Members

Dr. Kyung Yong Kim

Dr. William Stout

October 20, 2025

Date of Acceptance by Committee

October 17, 2025

Date of Final Oral Examination

ACKNOWLEDGEMENTS

I want to express my deepest gratitude to my advisor, Dr. Robert Henson, for his unwavering support and mentorship throughout this journey. Your guidance and insight were invaluable in shaping this research. To Dr. Kyung Yong Kim, thank you for your early guidance in the program and for your continued encouragement and support along the way. To Dr. William Stout, I am sincerely grateful for your expertise and thoughtful feedback, which helped strengthen and deepen my research.

I would also like to thank the ERM faculty and my fellow students for their support, collaboration, and encouragement over the years. Finally, to my family and friends, thank you for the constant motivation and belief in me—your support made this doctoral journey possible.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER I: INTRODUCTION 1

Diagnostic Classification Modeling 1

Overview of Multinomial Diagnostic Classification Models 2

Test Construction 4

Study Purpose 5

Research Questions 7

CHAPTER II: LITERATURE REVIEW 8

Diagnostic Classification Models 8

Key Components of DCMs 8

Types of DCMs 9

Conjunctive Models. 9

Disjunctive Models. 10

Dichotomous Models 10

DINA. 11

RRUM. 12

DINO. 13

More Examples of DCMs. 14

Multinomial Models 18

MC-DINA. 18

SICM. 19

ERUM-MC. 19

Computer Adaptive Testing 22

Item Selection 22

Early Item Selection Methods in DCM 23

Shannon Entropy. 23

Kullback Leibler Information. 24

Posterior Weighted Kullback Leibler Information. 25

Hybrid Kullback Leibler. 25
Mutual Information. 26
Modified Posterior Weighted Kullback Leibler. 27
Generalized Deterministic Inputs, Noisy "and" Gate Model Discrimination Index 27
Q-optimal design. 28
Posterior-Weighted CDM Discrimination Index and Posterior-Weighted Attribute-Level CDM Discrimination Index. 29
Modified Maximum Global Discrimination Index. 32
Standardized Weighted Deviation Global Discrimination Index. 32
Constrained Progressive Standardized Weighted Deviation Global Discrimination Index. 33
Progressive, Restrictive Progressive, and Restrictive Threshold Posterior Weighted Kullback Leibler. 34
Jensen-Shannon divergence index. 35

CHAPTER III: METHODS 38

Data Generation and Simulation Conditions 38
Item Selection Methods 41
Single Attribute CCR Method 42
Composite CCR Procedure 43
Shannon Entropy Procedure 43
Fixed Forms 44
Evaluation Criteria 45

CHAPTER IV: RESULTS 48

Average Correct Classification Rates 50
4 Attributes and 5 Items 51
4 Attributes and 10 Items 56
6 Attributes and 5 Items 61
6 Attributes and 10 Items 67
Average Absolute Deviations from 0.5 73
4 Attributes and 5 Items 73
4 Attributes and 10 Items 76
6 Attributes and 5 Items 78

6 Attributes and 10 Items 81
Average Item Exposure Rate 83
4 Attributes and 5 Items 84
4 Attributes and 10 Items 88
6 Attributes and 5 Items 92
6 Attributes and 10 Items 96
Average Time 100
CHAPTER V: DISCUSSION 102
Key Findings and Conclusion 102
Research Question 1: Multinomial extension of expected SHE 105
Research Question 2: CD-CAT method comparison 106
Research Question 3: Well-designed fixed form vs. CD-CAT 107
Limitations and Future Research 108
REFERENCES 112

LIST OF TABLES

Table 1. Taxonomy of Core DCMs 16

Table 2. Attribute Level Correct Classification Rates for 4 Attributes with a Test Length of 5 .. 53

Table 3. Attribute Profile Correct Classification Rates for 4 Attributes with a Test Length of 5 55

Table 4. Attribute Level Correct Classification Rates for 4 Attributes with a Test Length of 10 58

Table 5. Attribute Profile Correct Classification Rates for 4 Attributes with a Test Length of 10 60

Table 6. Attribute Level Correct Classification Rates for 6 Attributes with a Test Length of 5 .. 63

Table 7. Attribute Profile Correct Classification Rates for 6 Attributes with a Test Length of 5 66

Table 8. Attribute Level Correct Classification Rates for 6 Attributes with a Test Length of 10 69

Table 9. Attribute Profile Correct Classification Rates for 6 Attributes with a Test Length of 10 72

Table 10. Average Absolute Deviations from 0.5 for 4 Attributes with a Test Length of 5 75

Table 11. Average Absolute Deviations from 0.5 for 4 Attributes with a Test Length of 10 78

Table 12. Average Absolute Deviations from 0.5 for 6 Attributes with a Test Length of 5 80

Table 13. Average Absolute Deviations from 0.5 for 6 Attributes with a Test Length of 10 83

Table 14. Average Item Exposure Rate for 4 Attributes with a Test Length of 5 87

Table 15. Average Item Exposure Rate for 4 Attributes with a Test Length of 10 91

Table 16. Average Item Exposure Rate for 6 Attributes with a Test Length of 5 95

Table 17. Average Item Exposure Rate for 6 Attributes with a Test Length of 10 99

Table 18. Average Time Per Replication 101

LIST OF FIGURES

Figure 1. Attribute-Level Correct Classification Rates for 4 Attributes with a Test Length of 5 51

Figure 2. Attribute Profile Correct Classification Rates for 4 Attributes with a Test Length of 554

Figure 3. Attribute-Level Correct Classification Rates for 4 Attributes with a Test Length of 10 56

Figure 4. Attribute Profile Correct Classification Rates for 4 Attributes with a Test Length of 10 59

Figure 5. Attribute-Level Correct Classification Rates for 6 Attributes with a Test Length of 5 62

Figure 6. Attribute Profile Correct Classification Rates for 6 Attributes with a Test Length of 565

Figure 7. Attribute-Level Correct Classification Rates for 6 Attributes with a Test Length of 10 68

Figure 8. Attribute Profile Correct Classification Rates for 6 Attributes with a Test Length of 10 71

Figure 9. Average Absolute Deviations from 0.5 for 4 Attributes with a Test Length of 5..... 74

Figure 10. Average Absolute Deviations from 0.5 for 4 Attributes with a Test Length of 10..... 77

Figure 11. Average Absolute Deviations from 0.5 for 6 Attributes with a Test Length of 5..... 79

Figure 12. Average Absolute Deviations from 0.5 for 6 Attributes with a Test Length of 10..... 82

Figure 13. Average Item Exposure Rate for 4 Attributes with a Test Length of 5..... 85

Figure 14. Average Item Exposure Rate for 4 Attributes with a Test Length of 10..... 89

Figure 15. Average Item Exposure Rate for 6 Attributes with a Test Length of 5..... 92

Figure 16. Average Item Exposure Rate for 6 Attributes with a Test Length of 10..... 96

CHAPTER I: INTRODUCTION

Diagnostic Classification Modeling

In K-12 education, teachers want to know how their students are performing and whether the students are mastering the content. This information can be obtained through formal standardized testing and classroom assessments. Most of the time, test scores and grades are based on whether a student answers a question correctly or incorrectly, which often only provides information about a student's general knowledge. However, these tests are not usually designed or scored to drill down to the multiple skills associated with each question. As opposed to a total correct score, diagnostic classification modeling (DCM; Rupp et al., 2010), also referred to as cognitive diagnostic modeling (CDM; Nichols et al., 1995; Templin & Henson, 2006), can be used to obtain a more detailed profile of a student's, or examinee's, knowledge. Using DCM, it is possible to estimate a profile that describes the mastery of a set of attributes or skills for each examinee. Attributes will be the term used throughout this paper. This profile, or summary of the profile, for a class, can then be used to indicate which attributes an examinee should focus on learning or which attributes should be reviewed within a class.

It is important to note that, because the literature uses CDM and DCM synonymously, diagnostic classification modeling, or DCM, will be used throughout this document. Additionally, the term "examinee" will be used throughout this paper to refer to the individual or student taking the test. Finally, the DCM literature will often refer to attributes or skills that an examinee has or has not possessed. For clarity, the term "Attribute" will be used throughout this document.

Diagnostic classification models are constrained latent class models that place examinees into classes by building a latent profile based on the attributes they have mastered or not

mastered on a test. Examinees with the same mastery profile are assumed to belong to the same class. Knowing the profile (i.e., the examinee's latent class) of an examinee then specifically determines what attributes the examinee has mastered. In turn, the attributes mastered or not mastered are used to predict the probability of a response to an item. Diagnostic classification models differ in how mastered and non-mastered attributes interact to define the probability of a correct response for an item. Specifically, diagnostic classification models (DCMs; Rupp et al., 2010). can take two forms, compensatory and non-compensatory. Compensatory models, also known as disjunctive models, allow for nonmastery of an attribute to be compensated by mastery of another attribute, typically through an additive model; however, the model can also contain products (Rupp et al., 2010). Non-compensatory, or conjunctive models, do not allow for mastery of one attribute to compensate for nonmastery of another attribute (Rupp et al., 2010). For a conjunctive model, given an item and all the attributes that are measured by it, an examinee would need to have mastered all those attributes to have a high probability of getting the item correct.

Overview of Multinomial Diagnostic Classification Models

Traditionally, literature in DCM focuses on dichotomous item scoring (e.g., DiBello et al., 1995; Hartz, 2002; Junker & Sijtsma, 2001; Macready & Dayton,1977; Templin & Henson, 2006; Templin; 2006) that is used to assess an examinee's strength and weaknesses, which results in an attribute profile that indicates whether each attribute has been mastered or not (Cheng, 2009). In the classroom, small tests or assessments are more common, and determining the probability of mastery of multiple attributes would prove to be difficult with a typical dichotomous model. As an alternative to dichotomous models, research has been conducted on models that not only assess whether an item is answered correctly or not, but also incorporate

information on how an item is missed. Common situations in research involve multiple-choice items, where the specific response, or option, chosen is used to determine mastery or nonmastery. Multinomial models for DCM can be helpful for short tests, as they consider both the correct answer option and information from the incorrect answer options of a multiple-choice item. For most multinomial DCMs, each option in the multiple-choice item is coded nominally. In addition, when using a multinomial model, the latent space might also be intentionally expanded to include misconceptions, along with attributes (Stout et al., 2022). The misconceptions might be added such that "possession" (a more general term to replace mastery) of a misconception might make specific options more likely. Examples of multinomial models for DCM in the literature are the Multiple-Choice Deterministic, Input, Noisy, and Gate model(MC-DINA: de la Torre, 2009), the Scaling Individuals and Classifying Misconceptions model (SICM; Bradshaw & Templin, 2014), and the Generalized Diagnostic Classification Model Multiple Choice (GDCM-MC; DiBello et al., 2015).

This study focused on a specific multinomial model called the Generalized Diagnostic Classification Model Multiple Choice (GDCM-MC) Option-Based Scoring family of models (DiBello et al., 2015) that can be adapted to use the DINA (Junker & Sijtsma, 2001), the reparameterized unified model (RUM; DiBello et al., 1995; Hartz, 2002; Templin, 2006), the loglinear cognitive diagnosis model (LCDM; Henson et al., 2009), and the general diagnostic model (GDM; von Davier, 2005). The specific GDCM-MC model that was used in this study is the extended reparametrized unified model-multiple choice (ERUM-MC; DiBello et al., 2015). The ERUM-MC is a multinomial extension of the Reparametrized Unified Model (RUM) that allows each option of a multiple-choice test to be more or less attractive depending on the

mastery (or possession) of various attributes, where some are skills and others are misconceptions.

Test Construction

Test assembly and design can take many forms, ranging from pencil-and-paper tests to computer-based tests, which can be delivered using either fixed form or computer-adaptive test (CAT) approaches. Examinees taking a fixed form test will all receive the same items, but the order of the items may vary. Fixed forms can be administered as both paper-and-pencil tests and computer-based tests. Computer adaptive tests adjust to an examinee's estimated latent trait level after each item (or set of items) is administered. The different items are chosen for each examinee based on their estimated latent trait at that point in the test. These items are ultimately selected in a way that measures the examinee's latent trait "best" (Yu et al., 2019). Yu et al. (2019) identify five main components of a computer-adaptive test: a calibrated item bank, a starting rule, an item selection strategy, a scoring method, and a stopping rule. Because items are selected in a customized manner for that examinee, it has been shown that CAT can be a more efficient way to assess an examinee's ability. For example, if enough items have been given to an examinee to determine that they are above average with respect to some construct (e.g., English ability), it is not informative to give that examinee extremely easy items. In addition, there is evidence that Cognitive Diagnostic Computer Adaptive Testing (CD-CAT) can be a powerful tool to determine an examinee's attribute profile in a shorter test (Li et al., 2023; Stout et al., 2003; Yigit et al., 2019).

Since diagnostic classification models contain discrete latent classes, compared to the continuous nature of Item Response Theory (IRT) models, the methods for item selection and various parts of CD-CAT need to differ from those used in CAT in IRT (Cheng, 2009). Common

CD-CAT item selection strategies use Shannon Entropy (SHE; Xu et al., 2003) and Kullback-Leibler Information (KL; Cheng, 2009). For every item selected in CD-CAT, the goal is to select items that are informative about attribute mastery for a candidate until the stopping rule is obtained.

In CD-CAT, several different stopping rules can be employed. One such rule is based on the number of items. In that case, the adaptive stops when a fixed number of items is given. As an alternative, the number of items might be variable. When the number of items is variable, the test stops when some other criteria is satisfied. Having a fixed-length test means that every examinee answered a predetermined number of items (Yu et al., 2019). A few different variable-length stopping rules have been proposed (Tatsuoka, 2002; Hsu et al., 2013). Tatsuoka (2002) proposed a method that would have a test stop assigning new items when the posterior probability of a specific attribute mastery pattern reached 0.8. Hsu et al. (2013) proposed two new stopping rules, in which the first included having an exam stop when the largest posterior probability of all the possible latent classes (i.e., attribute profiles in DCM) is no smaller than a specified value, which was suggested to be 0.7 in the paper. Hsu et al. (2013) proposed a second-stopping rule that included the constraints from the first-stopping rule, along with a second constraint. This second constraint was on the second largest value, such that it must be less than or equal to a specific value, which Hsu et al. (2013) suggested could be 0.1.

Study Purpose

Previous research studies have focused on item selection strategies for DCMs, primarily using dichotomous scoring models (e.g., Cheng, 2009; Henson & Douglas, 2005; Kaplan et al., 2015; Tatsuoka, 2002; Wang, 2013; Xu et al., 2003; Zheng & Chang, 2016). While there are a few multinomial models for DCM (e.g., Bradshaw & Templin, 2014; de la Torre, 2009; DiBello

et al., 2015), only one has been incorporated into CD-CAT, the MC-DINA (de la Torre, 2009). Stout et al. (2022) discussed how item discrimination indices have been studied widely in DCM (e.g., Henson et al., 2008; Henson et al., 2018; Kuo et al., 2016; Shear & Roussos, 2017; Stout et al., 2019; Wang et al., 2015) with some studies linking item discrimination to correct classification rates (CCRs). Theoretical CCRs, which are based on the assumption that an examinee is given a single item, can be computed at the item level before the exam is administered. These theoretical CCRs can then be used as a measure of item quality (Stout et al., 2022). If items are selected in this way, it is believed that an exam will be of high quality and, as a result, will correctly classify examinees. It should be noted that while test CCRs have been widely used as an evaluation measure for DCM studies indicating the quality of a test, theoretical CCRs at the item level have yet to be incorporated into an item selection algorithm for CD-CAT.

This study aimed to explore the effectiveness of using theoretical CCRs of multiple-choice items for developing item selection algorithms for CD-CAT. Specifically, two different approaches are defined using the theoretical CCRs. The first algorithm selects the items such that the theoretical CCRs are the highest for an attribute that had a posterior probability of mastery, PPM_k, closest to 0.5. The second algorithm selected the next items by calculating a weighted composite using the posterior probability of mastery from the previous item as the weights for the attribute-level theoretical CCRs. These two new methods based on theoretical CCRs were compared to a multinomial extension of the expected Shannon Entropy procedure (SHE; Tatsuoka, 2002). In addition to comparing these three methods to one another, the three methods are then compared to two different fixed forms. Specifically, all three item selection algorithms for CD-CAT were compared to a random fixed form, which is what has typically been done in the literature. However, beyond that, these three methods are also compared to a fixed form that

is created from strategically picked items that have higher theoretical correct classification rates from the item bank. This last approach is used to create a ”good” fixed form exam, which should be a more accurate comparison to what would be done in practice if an adaptive approach were not used.

Since previous studies have only shown that CD-CAT outperforms a randomly created fixed form across many methods, this study's goal was to determine if using a multinomial diagnostic classification model along with a theoretical CCR-based item selection algorithm will outperform a well-designed fixed form or whether, in some cases, a well-designed fixed form is the better option to use. For this simulation study, multiple factors were varied. Factors that were selected were in line with previous studies (e.g., DiBello et al., 2015; Cheng, 2010; Lin & Cheng, 2019; Zheng & Cheng, 2016). This study examines variations in levels related to the number of attributes measured by the test, the test length, and the quality of the item bank. Attribute-level CCRs, attribute profile CCRs, item exposure, and the average absolute deviations of the posterior probabilities from 0.5 were evaluated throughout the study to answer the following research questions:

Research Questions

1. How does the multinomial extension of expected Shannon Entropy perform?
2. How do CD-CAT methods using theoretical correct classification rates compare to CD-CAT based on expected Shannon Entropy?
3. Does CD-CAT perform better than using a well-designed fixed form test under the GDCM ERUM-MC model?

CHAPTER II: LITERATURE REVIEW

Diagnostic Classification Models

Measuring examinees' level of mastery can be challenging with traditional methods in educational measurement, such as Classical Test Theory (CTT) or Item Response Theory (IRT). In contrast, identifying the fine-grain level of examinee mastery can be accomplished with DCMs. DCMs are multidimensional models that allow for the measurement of a set of attributes that map out the process of cognitive response for a specific construct (Rupp et al., 2010). For example, in mathematics under the construct fraction subtraction, some of the attributes to be measured could be "Convert a whole number to a fraction" and "Separate a whole number from a fraction" (de la Torre & Douglas, 2004, as cited in Rupp et al., 2010). Defining attributes can vary depending on how fine-grained they need to be for a judgment to be made about an examinee.

Key Components of DCMs

Three components that are common to most DCMs are attributes, the Q-matrix, and an examinee's mastery profile. When building a test, the skills being measured by the items that an examinee needs to possess to correctly answer an item are called attributes (Templin & Henson, 2006). In addition to attributes, DCMs require the specification of a Q-matrix. The Q-matrix is an indicator matrix that specifies which attributes are measured by each item. Specifically, a Q-matrix is a matrix where each row represents an item, while each column represents an attribute measured on the test. For each item, a 1 or 0 in the row represents whether an attribute is measured by that item or not, respectively (Rupp et al., 2010). Examinee profiles ($\boldsymbol{\alpha}$) define whether or not an examinee has mastered or not mastered a set of attributes (Rupp et al., 2010).

Types of DCMs

In addition to these main components of most DCMs, two main types of models have been developed that differ based on other components using a condensation rule. Specifically, models differ with respect to how the measured attributes of an item combine to predict the probability of a correct response. The rule that describes how these attributes combine is often described as the condensation rule (Rupp et al., 2010). There are two general types of condensation rules, which are described as either a conjunctive condensation rule or a disjunctive condensation rule. Because of this, models using these two rules are referred to as conjunctive or disjunctive models.

Conjunctive Models. Conjunctive, or noncompensatory, models apply a conjunctive condensation rule, as seen in Equation 2.1 (Rupp et al., 2010). In Equation 2.1, the response processes for all the attributes measured by an item are multiplied together, resulting in a value of 0 or 1. The result is only 1 if all response processes of the measured attributes of an item are 1, which is consistent with the idea of a conjunctive model, where it is assumed, for dichotomous items, that all attributes measured by an item should be mastered in order to have a high probability of correctly answering an item (Rupp et al., 2010). If even one of the measured attributes is not mastered, then the probability of correctly responding to an item is predicted to be much lower. This rule is suitable for situations where one cannot compensate for nonmastery of one attribute by mastery of another attribute. A conjunctive rule was described by Rupp et al. (2010) as the product of indicators representing mastery of each attribute measured by an item. Generally speaking, Equation 2.1 provides a general mathematical form of the conjunctive rule using a total of A attributes measured by an item.

$$Result = Attribute\ 1 \times Attribute\ 2 \times Attribute\ 3 \times \ldots \times Attribute\ A \tag{2.1}$$

Note that if even a single attribute is not mastered, indicated with a 0, then $Result = 0$. Some common examples of conjunctive models are the deterministic, input, noisy and-gate (DINA; Junker & Sijtsma, 2001; Macready & Dayton,1977), the noisy-input, deterministic-and-gate (NIDA; Junker & Sijtsma, 2001), and the noncompensatory reparametrized unified model (NC-RUM; DiBello et al., 1995; Hartz, 2002), which has full and reduced model variations.

Disjunctive Models. In contrast to conjunctive models, disjunctive, or compensatory, models apply a disjunctive compensation rule, as seen in Equation 2.2 (Rupp et al., 2010), where the response process for all the attributes measured by an item are subtracted from 1 and then multiplied together. To obtain an overall result of 1, which indicates that an individual will have the highest chance of getting the item correct, at least of the attribute multiplication portion of Equation 2.2 would have to equal 0. This equation demonstrates the assumption for dichotomous items where at least one attribute needs to be mastered for the examinee to correctly answer an item and receive a score of 1 (Rupp et al., 2010). Thus, in the most extreme models, mastery of a single attribute can "compensate" for nonmastery of all other attributes measured by the item.

$$Result = 1 - [(1 - Attribute\ 1) \times (1 - Attribute\ 2) \times (1 - Attribute\ 3) \times \ldots \times (1 - Attribute\ A)] \quad (2.2)$$

Some common examples of disjunctive models are the deterministic input, noisy-or-gate (DINO; Templin & Henson, 2006; Templin, 2006), the noisy input, deterministic-or-gate (NIDO; Templin, 2006), and the compensatory reparameterized unified model (C-RUM; Hartz, 2002; Templin, 2006)

Dichotomous Models

The most commonly cited dichotomous DCMs include both conjunctive and disjunctive models. These models include the deterministic, input, noisy, and-gate (DINA; Junker &

Sijtsma, 2001; Macready & Dayton,1977), the reduced reparametrized unified model (RRUM; DiBello et al., 1995; Hartz, 2002), and the deterministic, input, noisy or-gate model (DINO, Templin & Henson, 2006).

DINA. The deterministic, input, noisy, and-gate (DINA; Macready & Dayton, 1977; Junker & Sijtsma, 2001) model is a noncompensatory model that categorizes examinees into two groups for each item. One group contains examinees who have mastered all attributes measured by that item (i.e., the item masters). The other group of examinees are nonmasters of at least one attribute that is being measured by that item (i.e., item nonmasters). The assumption is that if an examinee lacks knowledge of at least one attribute that is measured by that item, then the examinee will not be able to answer the item correctly. The DINA model has two parameters, slipping, s_i, and guessing, g_i. Even if an examinee has all the knowledge and skills to answer an item correctly, the examinee may still answer incorrectly, or slip, on the item. In contrast, it is also possible for an examinee who lacks at least one attribute to still correctly answer the item. When an examinee lacks one or more attributes needed to answer an item but answers the item correctly, that is considered a correct "guess" for the item. With these two parameters, the DINA model, shown in Equation 2.3, defines the probability of correctly responding to item *i*.

$$P(X_{i\boldsymbol{\alpha}} = 1|\xi_{i\boldsymbol{\alpha}}) = (1 - s_i)^{\xi_{i\boldsymbol{\alpha}}} g_i^{1-\xi_{i\boldsymbol{\alpha}}} \tag{2.3}$$

The probability of correctly responding to item *i,* for an individual belonging to attribute profile, α, is represented by $P(X_{i\boldsymbol{\alpha}} = 1|\xi_{i\boldsymbol{\alpha}})$. The value $X_{i\boldsymbol{\alpha}}$ is the dichotomous response that is observed for item *i* for a given $\boldsymbol{\alpha}$. An indicator of mastery is $\xi_{i\boldsymbol{\alpha}}$, where $\xi_{i\boldsymbol{\alpha}} = 1$ indicates the examinee mastered all attributes for a given item *i*, or $\xi_{i\boldsymbol{\alpha}} = 0$, where the examinee lacks the knowledge of at least one attribute for a given item *i*. The probability of not slipping on item *i*, is represented by $(1 - s_i)$ and g_i represents the probability of guessing on item *i* (Rupp et al.,

2010). While the DINA is one of the most popular models used in studies due to its computational ease (Cheng, 2009), it has been criticized for assuming that all attributes are equally critical to answering an item, which may not always be useful in real-world situations.

RRUM. Another conjunctive model commonly discussed in the literature is the reduced reparametrized unified model (RRUM; DiBello et al., 1995; Hartz, 2002). The reduced reparametrized unified model (RRUM), also known as the reduced noncompensatory reparametrized unified model or the fusion model (DiBello et al., 1995; Hartz, 2002), has two parameters, $\pi_i^*$and r_{ia}^*. The parameter π_i^* represents the probability of correctly responding to item *i* when all the attributes measured by that item have been mastered. The parameter r_{ia}^* represents a penalty term to the probability of correctly responding to item *i* when attribute a has not been mastered (Rupp et al., 2010). That is, r_{ia}^* represents the factor that the probability of a correct response is reduced when the a^{th} attribute has not been mastered. Thus, the closer this penalty term is to zero for an attribute, the more important that attribute is to a particular item. Using these two parameters, the RRUM can be formulated by

$$P(X_i = 1|\boldsymbol{\alpha}) = \pi_i^* \prod_{k=1}^{K} r_{ik}^{*(1-\alpha_k)q_{ik}} \tag{2.4}$$

In Equation 2.4, $P(X_i = 1|\boldsymbol{\alpha})$ represents the probability of correctly responding to item *i,* given the attribute profile $\boldsymbol{\alpha}$ for a given examinee. The dichotomous response that is observed for item *i* is represented by X_i . The value q_{ik} is the Q-matrix element indicating whether the i^{th} item measures attribute k. The vector $\boldsymbol{\alpha}$ is a vector that includes information about attribute mastery for a given examinee, which includes α_k, that represents whether attribute k is mastered for an individual $\boldsymbol{\alpha}$ (Rupp et al., 2010).

DINO. In contrast to conjunctive models, a disjunctive, or compensatory, model that is commonly used is the deterministic, input, noisy or-gate model (DINO; Templin & Henson, 2006). The DINO model is the disjunctive version of the DINA. Because the DINO model is compensatory, it employs the disjunctive condensation rule, also known as the "or"-gate, which indicates that if at least one attribute measured by the item is present, it can compensate for all the other attributes (Templin & Henson, 2006). Similar to the DINA, the DINO model breaks examinees into two groups: those who should get the item right and those who are most likely to miss the item. Specifically, one group of examinees is those who have mastered at least one of the attributes measured by the item. This group is expected to have a high probability of correctly responding to the item. The second group consists of examinees who are not masters of any of the attributes measured by an item. Because none of the attributes have been mastered, they are expected to incorrectly respond to the item. Once the two groups are identified, the DINO model is similar to the DINA model, shown in Equation 2.5, having the same two item parameters, slipping, s_i, and guessing, g_i, but they are interpreted slightly differently because the groups have been defined differently. A slip, s_i, is the probability of a score of 0 when at least one of the attributes being measured is present and a guess, g_i, is the probability of a score of 1 on an item when no attributes that are measured are present(Rupp et al., 2010). The DINO model equation for the probability of correctly responding to item *i* is:

$$P(X_{i\boldsymbol{\alpha}} = 1|\omega_{i\boldsymbol{\alpha}}) = (1 - s_i)^{\omega_{i\boldsymbol{\alpha}}} g_i^{1-\omega_{i\boldsymbol{\alpha}}} \tag{2.5}$$

In Equation 2.5, $P(X_{i\boldsymbol{\alpha}} = 1|\omega_{i\boldsymbol{\alpha}})$ represents the probability of correctly responding to item *i,* which belongs to attribute profile $\boldsymbol{\alpha}$. The value $X_{i\boldsymbol{\alpha}}$ is the response that is observed for item *i* for the attribute profile $\boldsymbol{\alpha}$. The value $\omega_{i\boldsymbol{\alpha}}$ is an indicator of whether at least one attribute has been mastered or not for item i where $\omega_{i\boldsymbol{\alpha}} = 1$ means at least one of the attributes being

measured is present, and $\omega_{i\boldsymbol{\alpha}} = 0$, means that all measured attributes are absent. The probability of not slipping on item *i* is represented by $(1 - s_i)$, and g_i represents the probability of guessing on item *i* (Rupp et al., 2010).

More Examples of DCMs. While the previous models are examples of DCMs, it is important to note that the literature contains many other DCMs. Some of the first DCM methods that were developed include the rule space method (RSM; Tatsuoka, 1983, 1995, 2009) and the attribute hierarchy method (AHM; Gierl et al., 2007; Leighton et al., 2004). The noncompensatory NIDA model (Junker & Sijtsma, 2001) and the compensatory NIDO model (Templin, 2006) are counterparts in that both model response behavior at the attribute level but have equality constraints across items (Rupp et al., 2010). Additive models have been developed, including the additive CDM (A-CDM; de la Torre, 2011) and the linear logistic model (LLM; Maris, 1999). The log-linear CDM (LCDM; Henson et al., 2009) and the generalized DINA (G-DINA; de la Torre, 2011) are less restrictive diagnostic classification models with more flexible frameworks compared to simpler, less constrained models (Rupp et al., 2010).

Building on the DINO model, the following section focuses on modifications of the model that were developed to extend its applications. A few models that are modifications of the DINO model include the Bug-DINO (Kuo et al., 2017) and the simultaneously identifying skills and misconceptions model (SISM; Kuo et al., 2017), which use skills and/or misconceptions to determine the probability of a correct response. Other work has extended the way in which the attribute space was modeled. For example, a more complex structure was modeled using a higher order structure (de la Torre & Douglas, 2004), while other models allow for polytomous attributes (e.g., Chen & de la Torre, 2013; Haberman et al., 2008; Karelitz, 2004; Templin, 2004; von Davier, 2008; von Davier, 2005).

Beyond the attribute space, some work has been done to extend the basic assumptions of DCM. Examples of these models include the model for multiple strategies (de la Torre & Douglas, 2008), the assign partial credit model (de la Torre, 2010), and a model with nominal responses (e.g., Chen & Zhou, 2017; Templin et al., 2008). In addition to these models, two diagnostic classification modeling frameworks do not represent a specific model. These include the Bayesian inference networks (BIN; Almond et al., 2015), a more general modeling framework representing a class of statistical models, and multiple classification latent class models (MCLCMs; Maris, 1999), a more flexible modeling framework (Rupp et al., 2010). Table 1 is adapted from Rupp et al. (2010) and summarizes these models and more core DCMs based on properties such as dichotomous or polytomous variables and the type of model they are, whether conjunctive or disjunctive.

Table 1. Taxonomy of Core DCMs

Observed Response Variable	Latent Predictor Variables: Dichotomous	Latent Predictor Variables: Polytomous	Model Type
Dichotomous	RSM		Conjunctive
	AHM		
	DINA		
	HO-DINA		
	MS-DINA		
	NIDA		
	RERUM		
	BIN	BIN	
	MCLCM	MCLCM	
	Full NC-RUM	Full NC-RUM	
	Reduced NC-RUM	Reduced NC-RUM	
	DINO		Disjunctive
	NIDO		
	BIN	BIN	
	MCLCM	MCLCM	
	C-RUM	C-RUM	
	GDM	GDM	
	H-GDM	H-GDM	
	LCDM	LCDM	

G-DINA	G-DINA	
RSM		
AHM		
BIN	BIN	Conjunctive
MCLCM	MCLCM	
Full NC-RUM	Full NC-RUM	
Reduced NC-RUM	Reduced NC-RUM	
BIN	BIN	
MCLCM	MCLCM	
C-RUM	C-RUM	
GDM	GDM	Disjunctive
H-GDM	H-GDM	
LCDM	LCDM	
G-DINA	G-DINA	

Note. RSM: rule-space method; AHM: attribute hierarchy method; BIN: Bayesian inference network; DINA: deterministic inputs, noisy "and" gate; HO-DINA- higher-order DINA; MS-DINA: multistrategy DINA; G-DINA: generalized DINA; DINO: deterministic inputs, noisy "or" gate; NIDA: noisy, inputs, deterministic "and" gate; NIDO: noisy inputs, deterministic "or" gate; GDM: general diagnostic model; HGDM: hierarchical GDM; MCLCM: multiple classification latent class models; RUM: reparametrized unified model/fusion model; C-RUM: compensatory RUM; NC-RUM: noncompensatory RUM; full NC-RUM: NC-RUM with continuous latent interaction term; reduced NC-RUM: NC-RUM without latent interaction term; LCDM: log-linear cognitive diagnosis model.

Multinomial Models

Sometimes a dichotomous scoring model may not be the best option, as it overlooks valuable information, especially with multiple-choice items that can provide more diagnostic detail at the option level. Measuring the level of mastery of multiple attributes on a short test can be challenging. However, in some instances, it might be possible to obtain additional information from an item based on the specific response, as opposed to whether it has only been answered correctly. For example, a well written multiple-choice item may contain formation about what has been mastered and not mastered based on which of the incorrect answers is selected. Thus, multinomial models may allow for fewer items to be used to still measure the same number of attributes with a test.

Multiple-choice options-based scoring is an alternative to dichotomous scoring, where more diagnostic information can be gained from each option, or distractor, of the item (de la Torre, 2009a). Response options in multiple-choice items reveal the type of thinking that an examinee may exhibit, whether it is problematic or desirable. Problematic thinking can manifest as misconceptions about an item or partial correct thinking about an item, whereas desirable thinking focuses on the skills and conceptual understanding of an item (DiBello et al., 2015). Some examples of diagnostic classification multinomial models include the Deterministic, Input, Noisy, and Gate for Multiple Choice items (MC-DINA: de la Torre, 2009a), the Scaling Individuals and Classifying Misconceptions model (SICM; Bradshaw & Templin, 2014), and Generalized Diagnostic Classification Models for Multiple Choice Option-Based Scoring family of models (GDCM-MC; DiBello et al., 2015).

MC-DINA. de la Torre (2009a) proposed one of the first multinomial DCMs called the Deterministic Noisy and Gate for Multiple Choice items (MC-DINA). The MC-DINA uses

multiple-choice items, where the correct response specifies the attributes that must be mastered, and the distractors are assumed to require only a subset of those attributes. That is, at least one of the necessary attributes for that item has not been mastered (de la Torre, 2009a). Compared to the DINA model, which only has two latent groups for a specific item, the MC-DINA has the number of latent groups equal to the number of options for the item plus one. Each latent group for the different options possesses a varying number of attributes that match the vector of the Q-matrix for that option. The plus one is for the group that does not possess any of the attributes or only possesses a combination of attributes that do not match any of the vectors from the Q-matrix for that item. For each group, there is a probability of selecting each option for a given item.

SICM. Using principles from IRT and DCM, Bradshaw & Templin (2014) proposed the Scaling Individuals and Classifying Misconceptions (SICM) model for nominal responses. Unlike other DCMs, both dichotomous and multinomial, the SICM model classifies examinees based on misconceptions while also measuring each examinee's ability. The SICM model assumes that a continuous trait, θ_e, primarily accounts for the covariance in the item responses. An assumption is that item responses that measure the same misconceptions are not independent when considering θ_e, which makes it different from unidimensional IRT. A second assumption about the SICM model is that there is a set of categorical misconceptions. It is assumed that each misconception can either be possessed or not be an examinee, which accounts for the variation in the selection of incorrect options (Bradshaw & Templin, 2014).

ERUM-MC. One suitable model used in this study is the extended reparameterized unified model-multiple choice (ERUM-MC; DiBello et al., 2015), which belongs to the Generalized Diagnostic Classification Models for Multiple Choice Option-Based Scoring family

of models (GDCM-MC; DiBello et al., 2015). DiBello et al. (2015) proposed the GDCM-MC family of models that includes four unique features compared to other models. One feature is an expanded latent space that includes both problematic and desired thinking. Problematic thinking occurs when an examinee possesses misconceptions or partially correct thinking about a concept, whereas desirable thinking is when an examinee possesses the skills and has a conceptual understanding of the topics presented in an item (DiBello et al., 2015). The second feature is a Q-matrix that is expanded, where each item has the same number of rows as the number of options it contains. The expanded Q-matrix for the GDCM-MC shows that possessing certain attributes and lacking certain misconceptions help make the correct option more attractive. In the Q-matrix, a coding scheme with three values is used, including 0, 1, and N. A third feature includes guessing in the model, where an examinee either randomly guesses an option or uses a combination of cognition to eliminate a few choices and then randomly guesses from the remaining choices. The last general feature of the GDCM is that it provides a modeling framework that can incorporate any diagnostic classification model functionality with respect to how the attributes interact to make specific options more or less attractive. Equation 2.6 defines the general GDCM item response function (IRF).

$$P(h|\boldsymbol{\alpha}) = \begin{cases} F(h|\boldsymbol{\alpha}) + \big(1 - S(\boldsymbol{\alpha})\big)\dfrac{1}{H} \; if \; S(\boldsymbol{\alpha}) < 1 \\ \dfrac{F(h|\boldsymbol{\alpha})}{S(\boldsymbol{\alpha})} \; if \; S(\boldsymbol{\alpha}) \geq 1 \end{cases} \tag{2.6}$$

In Equation 2.6, h represents the number of options for an item. The cognitive diagnostic function is represented by $F(h|\boldsymbol{\alpha})$, which is the probability of an examinee with an attribute profile, $\boldsymbol{\alpha}$, endorses option h assuming option h is shown in isolation. In Equation 2.6, the function $S(\boldsymbol{\alpha}) = \sum_h F(h|\boldsymbol{\alpha})$ (DiBello et al., 2015). The top equation of $P(h|\boldsymbol{\alpha})$ represents when an examinee is guessing. The bottom equation represents the competition among options that

occurs when an examinee is shown all options and has to cognitively choose the correct response option. A specific example of this modeling framework is the ERUM-MC, where a slight modification of the RRUM is used to determine the appeal of each option.

The ERUM-MC (DiBello et al., 2015) has two item parameters, π and r, similar to the RUM diagnostic classification model. The ERUM-MC model defines the probability of correctly responding to each option of a multiple-choice item using the cognitive kernel based on the RRUM. The equation for the cognitive kernel of the ERUM-MC for option h of item i, using the GDCM IRF is displayed in Equation 2.7 (DiBello et al., 2015)

$$F_{ERUM,i}(h|\boldsymbol{\alpha}) = \pi_{ih} \prod_{k\ such\ that\ q_{ihk} \neq N} r_{ihk}^{|q_{ihk} - \alpha_k|} \tag{2.7}$$

Within the ERUM-MC, there is a penalty for a mismatch between mastery status and the Q-matrix element related to that option for any given item. For example, if there is a 1 in the Q-matrix for a given option and attribute, but the examinee has not mastered that particular attribute, then this would be considered a mismatch. However, a mismatch is also possible if the Q-matrix element is specified to be a 0 when an examinee has mastered that specific attribute. Note that this represents a significant difference between a 0 entry in a traditional Q-matrix for dichotomous DCMs and an option in a Q-matrix for the ERUM-MC. For dichotomous DCMs, a 0 Q-matrix entry means that item does not measure the attribute. In contrast, in the ERUM-MC, a 0 entry for an option in a Q-matrix indicates that the option is more attractive when an examinee does not possess that attribute. An entry of N in the Q-matrix indicates that the attribute has no effect on how attractive the option is. When the Q-matrix does not equal N and α do not match the Q-matrix entry, r is multiplied as a penalty to reduce the probability of that response.

Multinomial models for DCM have been developed over the past 15 years. Most models not only address mastery of attributes, similar to dichotomous models, but also introduce skills and misconceptions into the model to examine items at the option level. Although multinomial models in DCM are not new to the field, they are relatively new to their application in computer adaptive testing.

Computer Adaptive Testing

Computer-adaptive testing (CAT) has been used in Item Response Theory (IRT) for many years but is relatively new to diagnostic classification modeling (DCM). Computer adaptive tests modify the difficulty of questions that are presented to an examinee based on the examinee's estimated ability level. As a result, different items are selected for each test-taker depending on their ability estimate at that moment in the test, allowing for a more accurate measurement of their latent traits. Using a scoring approach such as Item Response Theory, the usefulness of an item for estimating an examinee's ability depends on the examinee's ability. For example, it would not be nearly as useful to give an easy item to an examinee who had a high ability because it would be expected that the examinee would correctly respond to that item. In contrast, giving a high-ability examinee a difficult item would help refine an estimate of just how much that examinee knew.

Item Selection

Item selection strategies in the context of IRT have been discussed in the literature for many years, with the maximum Fisher information (MFI; Lord, 1980; Thissen & Mislevy, 2000) being a widely used approach. Fisher information quantifies how much information an observable random variable provides about an unknown ability estimate, θ, of an examinee (Cheng, 2009). The item selection algorithm based on MFI picks the next item which yields the

largest Fisher information at the ability estimate obtained after the previous item. Because diagnostic classification models involve discrete latent classes, unlike the continuous nature of IRT models, the approaches for item selection and the components of CD-CAT must differ from those used in CAT within IRT, and as a result, Fisher information is not a viable option (Cheng, 2009).

Although it has not been around as long as CAT in IRT, numerous research studies on various item selection algorithms, including those that consider attribute balancing and item exposure control, have been conducted in cognitive diagnostic computer adaptive testing (CD-CAT). Many research studies on various item selection algorithms for DCMs have been completed (e.g., Bao and Bradshaw, 2018; Cheng, 2009; Cheng, 2010; Kaplan et al., 2015; Lin & Chang, 2019; Tatsuoka, 2002; Wang et al., 2011; Wang, 2013; Xu et al., 2016; Xu et al., 2003; Zheng & Chang, 2016; Zheng & Wang, 2017), but there is only one algorithm for item selection using a multinomial model the has been studied (Yigit et al., 2019).

Early Item Selection Methods in DCM

Two of the early algorithms for item selection were Shannon Entropy (Tatsuoka, 2002) and Kullback-Liebler information (Xu et al., 2003). Many methods have been developed since then that are new or extensions of these algorithms.

Shannon Entropy. Shannon entropy (SHE) measures the association of uncertainty in a probability distribution (Cheng, 2009). In CD-CAT, the goal is to select items, using the calibrated item parameters, which minimize the expected Shannon entropy for the posterior distribution of the attribute vector. The goal is to choose items in an order that makes the posterior probability of the attribute pattern of an examinee approach one as fast as possible. Equation 2.8 shows the expected Shannon entropy

$$Sh(\pi_n, X_j) = \sum_{0}^{1} \left\{ E_n(\pi_n | X_j = x) \left(\sum_{c=1}^{2^M} P_j^x(\alpha_c) \left[1 - P_j(\alpha_c)\right]^{1-x} \pi_{n-1}(\alpha_c) \right) \right\} \tag{2.8}$$

where n is the number of items, π is a vector of probabilities, X_j is the j^{th} item, α_c is the given attribute pattern, M is the number of attributes, and $P_j(\alpha_c)$is the probability of correct response of X_j given α_c (Xu et al., 2003). Based on the expected Shannon entropy, the next item is chosen to minimize the uncertainty of the posterior distribution for a given examinee's estimate attribute pattern (Cheng, 2009).

Kullback Leibler Information. Another item selection method similar to Shannon entropy is Kullback-Leibler (KL) information. Cheng (2009) describes KL information as measuring the distance between one probability distribution that represents the true distribution of data, while the second distribution is a model, theory, or approximation of the true distribution. Equation 2.9 shows K-L information

$$D[f, g] = E_f \left[log \frac{f(x)}{g(x)} \right] \tag{2.9}$$

where $f(x)$ is the true distribution while $g(x)$ model, theory, or approximation of $f(x)$. The KL algorithm that is used for CD-CAT constructs a discrimination index that is based on the KL distance using the distribution of a person's responses, U_{ih}, given the person's current latent cognitive profile, or attribute pattern, $\hat{\alpha}_i^{(t)}$, and the distribution of a person's responses given other attribute patterns, α_c. Based on that KL distance, the sum is calculated across all attribute patterns, and the resulting formula is

$$KL_h\left(\hat{\alpha}_i^{(t)}\right) = \sum_{c=1}^{2^K} \sum_{q=0}^{1} log \left(\frac{P\left(U_{ih}=q|\hat{\alpha}_i^{(t)}\right)}{\left(U_{ih}=q|\alpha_c\right)} \right) P\left(U_{ih} = q|\hat{\alpha}_i^{(t)}\right). \tag{2.10}$$

The KL algorithm then selects the next item that maximizes the KL information function (Cheng, 2009). Xu et al. (2003) examined Shannon entropy, the KL algorithm, and random

selection using the fusion model to see which performed better. Shannon entropy was more accurate in correctly classifying individuals into latent classes than the KL algorithm for both 30-item and 50-item tests with five and eight attributes by about 5 to 10% per attribute. The KL algorithm performed about 10% better than a randomly selected form for each attribute for both test lengths and number of attributes.

Posterior Weighted Kullback Leibler Information. An extension of the KL algorithm is the posterior weighted KL (PWKL) algorithm. The PWKL algorithm expands on the KL algorithm by including an informative prior linking the current latent state, $\boldsymbol{\alpha}$, to the previous data's latent state (Cheng, 2009). By imposing priors on the KL algorithm, the latent states are weighted more when they are more likely based on previous response data. The prior imposed represents the probability of a given latent state when *t* items have been administered. The resulting Equation 2.11 for the PWKL index is

$$PWKL_h\left(\hat{\alpha}_i^{(t)}\right) = \sum_{c=1}^{2^K} KL_h\left(\hat{\alpha}_i^{(t)} || \alpha_c\right)\pi_t(\alpha_c) \tag{2.11}$$

where $\pi_t(\alpha_c)$ represents the probability of a given latent state, α, when t items have been administered (Yu et al., 2019). In this item selection method, the next item is chosen based on the item that maximizes the PWKL index. A modification to the PWKL index to consider latent states that are closer to the current α estimate is the Hybrid KL (HKL) index.

Hybrid Kullback Leibler. The HKL index gives a weighting to latent states that are close to the current estimate of the latent state, $\hat{\alpha}_i^{(t)}$ using the inverse of the distance between the latent states (Cheng, 2009). The HKL index can be calculated using Equation 2.12.

$$HKL_h\left(\hat{\alpha}_i^{(t)}\right) = \sum_{c=1}^{2^K} KL_h\left(\hat{\alpha}_i^{(t)} || \alpha_c\right) \pi_t(\alpha_c) \frac{1}{\sqrt{\sum_{k=1}^{K}\left(\alpha_{ck} - \hat{\alpha}_{ik}^{(t)}\right)^2}} \tag{2.12}$$

Cheng (2009) ran a simulation study to compare the HKL index to PWKL, KL, SHE algorithms, and random selection. The SHE algorithm still outperforms the KL algorithm, as found in Xu et al. (2003). The PWKL and HKL methods perform slightly better, within 1-4%, of the SHE algorithm, but have similar outcomes to each other. Random selection of a test form performed about 10-15% less than the other methods' correct classification rates for individual attributes. In terms of attribute pattern recovery rates, the PWKL and HKL methods had similar overall rates. Overall, the SHE, PWKL, and HKL were twice as efficient as the KL algorithm and random selection. Efficiency was calculated by comparing each method to the random selection method by dividing each whole pattern recovery rate by the whole pattern recovery rate of random selection. Cheng (2009) also found comparable results using a real data set.

Mutual Information. Wang (2013) proposed the Mutual Information (MI) item selection method to use in CD-CAT. MI measures how dependent the X and Y distributions are on each other using their joint distributions and the product of their marginal distribution. The MI is larger when the dependence X has on Y is larger. In the context of CD-CAT, mutual information means the information gained about a particular attribute pattern, $\boldsymbol{\alpha}$, after another item is chosen for the test (Wang, 2013). The goal is to maximize Equation 2.13:

$$\sum_{y=0}^{1} p(Y_n = y|y_{n-1}) \left[\sum_{c=1}^{2^K} \pi(\alpha_c|y_{n-1}, Y_n = y) log \frac{\pi(\alpha_c|y_{n-1}, Y_n = y)}{\pi(\alpha_c|y_{n-1})}\right] \tag{2.13}$$

where $p(Y_n = y|y_{n-1})$ is the binomial distribution of the next response based on all prior items, and $\pi(\alpha_c|y_{n-1})$is the posterior distribution of a given attribute pattern, α_c, based on the $n-1$ items given so far (Wang, 2013). A simulation study compared MI to SHE, KL, PWKL,

and random selection. Results showed that the MI method was most effective for recovering full attribute patterns in a 5-item test, using both simple and complex Q-matrices, as well as high- and low-quality items. The PWKL method recovered the full attribute pattern best for the 10-item test for both the simple and complex Q-matrix, along with tests with low- and high-quality items. Similar to previous studies, random selection performed about 10% worse than the other methods across all conditions. After the MI was developed, Kaplan et al. (2015) developed the modified posterior-weighted Kullback-Leibler index (MPWKL) and the generalized deterministic inputs, noisy "and" gate (G-DINA) model discrimination index (GDI).

Modified Posterior Weighted Kullback Leibler. Sometimes, with short tests, using the current attribute pattern estimate may not lead to a good summary of the posterior distribution of the attribute vectors. The MPWKL method can help alleviate this issue in shorter tests by looking at the whole posterior distribution of the 2^K attribute patterns. By using the MPWKL method, it does not require a point estimate of the attribute vector; instead, it assigns a weighting to all attribute vectors (Kaplan et al., 2015). By modifying the PWKL method based on the new weighting, the MPWKL method is defined as follows in Equation 2.14.

$$MPWKL_{ij}^{(t)} = \sum_{d=1}^{2^K}\left[\sum_{c=1}^{2^K}\left[\sum_{x=0}^{1} log\left(\frac{P(X_j = x|\alpha_d)}{P(X_j = x|\alpha_c)}\right)P(X_j = x|\alpha_c)\pi_i^{(t)}(\alpha_c)\right]\pi_i^{(t)}(\alpha_d)\right] \tag{2.14}$$

Generalized Deterministic Inputs, Noisy "and" Gate Model Discrimination Index. The GDI was first introduced as a method of Q-matrix validation but is now used as an item selection method using a reduced attribute vector (de la Torre and Chiu, 2010, as cited in Kaplan et al., 2015). The reduced attribute vector, α_{cj}^*, contains only the attributes that are measured by

an item (i.e., those attributes with 1's in the Q-matrix for an item). The GDI is a measure of how well an item can distinguish between all the reduced attribute vectors and their probabilities of success (Kaplan et al., 2015). Given the reduced attribute vector, the probability of success for item j is $P\left(X_{ij}=1|\alpha_{cj}^{*}\right)$ and the probability of obtaining a reduced attribute vector, α_{cj}^{*}, is $\pi(\alpha_{cj}^{*})$. The mean probability of success, or $\bar{P}_j$, is the probability, summed across all 2^{K_j} attributes, of obtaining the reduced attribute vector times the probability of success for item j. Here K_j is the number of required attributes for item j. Based on this knowledge, the formula for the GDI index is given by Equation 2.15.

$$\varsigma_j^2 = \sum_{c=1}^{2^{K_j^*}} \pi\left(\alpha_{cj}^{*}\right)\left[P\left(X_{ij}=1|\alpha_{cj}^{*}\right)-\bar{P}_j\right]. \tag{2.15}$$

Kaplan et al. (2015) ran a simulation to compare the MPWKL and GDI methods to the PWKL method using multiple DCMs. The MPWKL and GDI methods both outperformed the PWKL method, for a fixed length test, when it comes to correct attribute vector classification for the deterministic, input, noisy, "and" gate (DINA; de la Torre, 2009; Haertel, 1989; Junker & Sijtsma, 2001) and deterministic, input, noisy, "or" gate (DINO; Templin & Henson, 2006) models but have mixed results for the additive cognitive diagnostic model (A-CDM model; de la Torre, 2011). In terms of a variable-length test, the MPWKL and GDI methods both produced shorter tests than the PWKL method under both the DINA and DINO models. Comparing the GDI, SHE, and MPWKL methods, a new method for initial item selection within these item selection strategies was developed.

Q-optimal design. Xu et al. (2016) introduced an initial item selection method based on a Q-optimal design for CD-CAT. Four theorems were introduced by Xu et al. (2016) to demonstrate the best selection methods for initial items. One theorem shows the optimal Q-matrix for linear tests, two theorems deal with the optimal Q-matrix for sequential tests and then

more specifically for the DINA model, and the last theorem generalizes the findings from the other theorems to other conjunctive models or models that do not allow for one attribute to compensate for another attribute (Rupp et al., 2010). Under the third theorem, using the DINA model, a Q-matrix is optimal when, for item *k*, the k^{th} column has a 0 when an examinee answers the item incorrectly, and the entry is either a 0 or a 1 when the *k*th item is correct. These theorems help with narrowing the number of items needed to satisfy the optimal Q-matrix condition. Xu et al. (2016) conducted a simulation study to compare the MPWKL, GDI, and SHE methods, which satisfy an optimal Q-matrix, with a complete Q-matrix, and the PWKL method. All three methods, MPWKL, GDI, and SHE, exhibit similar attribute pattern recovery rates and outperform the PWKL and complete Q-matrix methods. Other methods that aim to attain a person's attribute pattern with a short test are the posterior-weighted CDM discrimination index, or PWCDI, and the posterior-weighted attribute-level CDM discrimination index, or PWACDI.

Posterior-Weighted CDM Discrimination Index and Posterior-Weighted Attribute-Level CDM Discrimination Index. PWCDI and PWACDI methods are derived from the CDM discrimination index (CDI, Henson & Douglas, 2005) and the attribute-level CDI (ACDI; Henson et al., 2008; Rupp et al., 2010), respectively. The PWCDI starts with the CDI index that is built using a D matrix, shown in Equation 2.16, for the j^{th} item that consists of the expected KL distance between the distributions of responses for any two attribute profiles.

$$D_{juv} = E_{\alpha_u}\left[log\frac{P_{\alpha_u}(X_j)}{P_{\alpha_v}(X_j)}\right]. \tag{2.16}$$

Building on the D matrix, the CDI, shown in Equation 2.17, is calculated by taking the weighted mean of the off-diagonal elements of D_j (Henson & Douglas, 2005).

$$CDI_j = \frac{1}{\sum_{u \neq v}[\sum_{k=1}^{K}|\alpha_{uk}-\alpha_{vk}|]^{-1}} \sum_{u \neq v}[\sum_{k=1}^{K}|\alpha_{uk} - \alpha_{vk}|]^{-1} D_{juv}. \quad (2.17)$$

The ACDI (Henson et al., 2008; Rupp et al., 2010) was developed for attribute k and uses the relevant cells of the D matrix where the sum of the absolute difference of the two attribute profiles across all attributes, or the Hamming distance, equals 1 and is calculated for item j using Equation 2.18.

$$ACDI_j = \frac{1}{2^K} \sum_{all\ relevant\ cells} D_{juv}. \quad (2.18)$$

Similar to how the PWKL method utilizes a posterior weighting of the KL algorithm, PWCDI and PWACDI are constructed based on the weighting of the CDI and ACDI methods (Zheng & Chang, 2016). The difference in the D matrix for these two new methods is that the weights need to be applied across both rows and columns, not just columns, as shown in the KL algorithm. Using the new posterior weighted D matrix, the PWCDI and PWACDI can be represented similarly to the CDI and ACDI methods by

$$PWCDI_j = \frac{1}{\sum_{u \neq v}[\sum_{k=1}^{K}|\alpha_{uk}-\alpha_{vk}|]^{-1}} \sum_{u \neq v}[\sum_{k=1}^{K}|\alpha_{uk} - \quad (2.19)$$

$$\alpha_{vk}|]^{-1} E_{\alpha_u}\left[\pi(\alpha_u) \times \pi(\alpha_v) \times log\left(\frac{P(X_j|\alpha_u)}{P(X_j|\alpha_v)}\right)\right] \text{ and}$$

$$PWACDI_j = \frac{1}{2^K} \sum_{all\ relevant\ cells} E_{\alpha_u}\left[\pi(\alpha_u) \times \pi(\alpha_v) \times \quad (2.20)$$

$$log\left(\frac{P(X_j|\alpha_u)}{P(X_j|\alpha_v)}\right)\right].$$

Zheng and Chang (2016) used a simulation study to compare the PWCDI and PWACDI with the methods previously mentioned using both fixed-length and variable-length CAT tests. Based on the fixed-length test study, PWCDI and PWACDI perform similarly to the MI method in terms of the correct classification rate of the whole attribute pattern. All three methods perform about 10% better than the PWKL method for the correct classification rate for the whole

attribute pattern, and about 50% better than the CDI and ACDI methods. For the variable-length test, all three methods —MI, PWACDI, and PWCDI — have a mean test length smaller than PWKL.

Bao and Bradshaw (2018) developed an extension of the ACDI sometimes referred to as CDI_A (Henson et al., 2008 as cited in Bao and Bradshaw, 2018), where, based on the response to the previous item, the marginal posterior probability is calculated for each attribute and the one that is closest to 0.5 will be the target for the next item. After deciding which attribute will be the target, the discrimination of that attribute is calculated for each item. The item with the largest discrimination for that attribute is chosen as the next item. The item discrimination for an attribute is calculated by taking the sum of the discrimination for the masters of that attribute and the discrimination for the nonmasters of that attribute. Based on this method, the goal is to look for the attribute whose estimate is the least certain and choose an item that gives the most information about that item (Bao & Bradshaw, 2018).

Bao and Bradshaw (2018) ran two simulation studies comparing the CDI_A to the KL, PWKL, and HKL item selection methods. In the first study, each item measured only one attribute, and items that measured the first attribute provided more statistical information than those that measured the second attribute. The correct classification rate for attribute one was the lowest for the new CDI_A method, while it was the highest for attribute two compared to the other three methods. For study two, three attributes were measured, and each item only measured one attribute. In addition, attribute three had the least information compared to attributes one and two. The CDI_A method had a correct classification rate of above 94% for all conditions for attribute one but had lower rates than both the PWKL and HKL methods. For attribute three, CDI_A performed better than all three methods, KL, PWKL, and HKL (Bao & Bradshaw, 2018).

Coupled with item selection, content constraints, and exposure control can improve classification rates.

Modified Maximum Global Discrimination Index. The modified maximum global discrimination index, or MMGDI, method was proposed by Cheng (2010) to help with attribute balancing that looks at the minimum number of items required to measure an attribute, B_k, and the number of items that have been selected so far that measure that attribute, b_k. An attribute balancing index is added to the global discrimination index (Xu et al., 2003) to form the MMGDI. The attribute balancing index is the product across all attributes of the proportion of the difference of B_k and b_k out of the minimum number of items required to measure an attribute, raised to the 0 or 1 power, the entry for that item from the Q-matrix. Combing the attribute balancing index with the GDI, the resulting formula, developed by Cheng (2010), is

$$MGDI_j(\widehat{\underset{\sim}{\alpha}}_l) = \prod_{k=1}^{K} \left(\frac{B_k - b_k}{B_k}\right)^{g_{jk}} \cdot GDI_j(\widehat{\underset{\sim}{\alpha}}_l). \tag{2.21}$$

When using the MGDI, if an entry in the Q-matrix is 0 for a particular item, then the next item is chosen based solely on the GDI. Cheng (2010) conducted a simulation study comparing a short (24-item) test and a long (30-item) test, using the GDI and MMGDI item selection methods. For both test lengths, the MMGDI recovered the entire attribute profile better than the GDI method.

Standardized Weighted Deviation Global Discrimination Index. Two additional examples of attribute balancing methods are the standardized weighted deviation global discrimination index (SWDGDI) and the constrained progressive index (CP_SWDGDI). Lin and Chang (2019) proposed a new attribute-balancing item selection method, the SWDGDI, and subsequently proposed the CP_SWDGDI based on this method. The authors' goals were to create an attribute-balancing item selection method that included a weighting term that accounted for

scenarios where some attributes are more important than others. This approach would also allow for other constraints based on an item's content. To obtain the SWDGDI, the weighted deviation GDI (WDGDI) was introduced using a weighted deviation index that is multiplied by the GDI, similar to how the MGDI index is constructed (Lin & Chang, 2019). The weighted deviation index, or WD, is constructed using the weighting given to an item, and then positive deviations from the lower and upper boundaries that are set by the stakeholders. Test length, for example, is the constraint given for the lower, D_{jL_k}, and upper boundary, D_{jU_k}, in the simulation study run by Lin & Chang (2019). WD and GDI were then standardized so that they would be on the same scale, and resulted in the SWDGDI, which is calculated by

$$SWDGDI_j(\hat{\alpha}_i) = \left(\frac{Max(WD_j) - WD_j}{Max(WD_j) - Min(WD_j)}\right) \times \left(\frac{GDI_j(\hat{\alpha}_i) - MinGDI_j(\hat{\alpha}_i)}{GDI_j(\hat{\alpha}_i) - MinGDI_j(\hat{\alpha}_i)}\right) \text{ where} \tag{2.22}$$

$$WD_j = \sum_{k=1}^{K}\left(W_k D_{jL_k}\right) + \sum_{k=1}^{K}\left(W_k D_{jU_k}\right). \tag{2.23}$$

Constrained Progressive Standardized Weighted Deviation Global Discrimination Index. The CP_SWDGDI (Lin & Chang, 2019) included a progressive term that introduces randomization into the selection method to allow for the selection of items not just based on the item giving the maximum information along with an exposure parameter, $\frac{r\_max}{r_j}$. The exposure parameter changes as the test progresses since the exposure rate, r, changes with each new item administered. CP_SWDGDI is calculated by

$$CP\_SWDGDI_j(\hat{\alpha}_i) = \frac{r\_max}{r_j} \times \left[\left(1 - \frac{X}{L}\right) R_j\right] + \frac{X}{L} \times R_{jl} \tag{2.24}$$

where $R_j$is the randomization term generated by a uniform distribution bounded by $LB_j = SWDGDI_j(\hat{\alpha}_i) - \frac{\left(SWDGDI_j(\hat{\alpha}_i) - Min\right)}{s}$ and $UB_j = SWDGDI_j(\hat{\alpha}_i) + \frac{\left(Max - SWDGDI_j(\hat{\alpha}_i)\right)}{s}$. *Min* and *Max* in the lower and upper bounds represent the minimum and maximum information in the item

pool, and *s* is a factor for adjusting the width of the information interval (Lin & Chang, 2019). A simulation study was conducted that showed that the SWDGDI method performed better than the GDI method in terms of overall attribute pattern recovery. In contrast, the CP_SWDGDI method performed poorly for pattern recovery due to the fact that it controlled for item exposure. In terms of item exposure rates, Lin & Chang (2019) found that the CP_SWDGDI had the lowest rate at 0.25, which is the maximum exposure rate in the simulation study, while GDI and SWDGDI both had high rates of exposure over 0.89, and many unused items. Item exposure was also researched by Wang et al. (2011) and Zheng & Wang (2017).

Progressive, Restrictive Progressive, and Restrictive Threshold Posterior Weighted Kullback Leibler. Three methods were proposed by Wang et al. (2011): the progressive method (P_PWKL), the restrictive progressive method (RP_PWKL), and the restrictive threshold method (RT_PWKL). P_PWKL is based on adding an importance parameter, β, to the PWKL method, and is represented by

$$P\_PWKL_j = (1 - x/L)R_j + PWKL_j * \frac{\beta x}{L};\ \beta > 0 \tag{2.25}$$

where $(1 - x/L)$ is the proportion of items left to choose from, L is the test length. The value β is the weighting added to the model to control item exposure, such that the lower the β value the more secure a test is valued. In addition, R_j is a random number generated from a uniform distribution from 0 to the maximum PWKL for item *j* (Wang et al., 2011). A further constrained method, RP_PWKL, incorporates a control based on the maximum exposure rate, ensuring that items are only used up to a certain threshold, r, as given by Equation 2.26.

$$RP\_PWKL_j = \left(1 - \frac{exp_j}{r}\right)\left[(1 - x/L)R_j + PWKL_j * \frac{\beta x}{L}\right]. \tag{2.26}$$

The restrictive threshold method uses an information interval based on the maximum PWKL value given by

$$[max(PWKL_i) - \delta, max(PWKL_i)] \tag{2.27}$$

where δ is a constant, and the items that fall in this interval form a set and the next item chosen comes from this set (Wang et al., 2011). A simulation study was run comparing the PWKL to these new methods with constraints. Overall, the PWKL method, without any exposure control, yielded the largest recovery rate for the attribute profile. The restrictive threshold and restrictive progressive methods have lower exposure balance chi-squared values than the PWKL method, indicating that they are less skewed than the other methods and have less over- or under-exposure of items for both test lengths of 25 and 40 items. The restrictive progressive method has the second lowest chi-squared value, with the only method lower being the random selection method.

Jensen-Shannon divergence index. Despite many item selection algorithms being used with dichotomous DCMs, only one algorithm has used a multinomial diagnostic classification model. Yigit et al. (2019) proposed the Jensen-Shannon divergence (JSD) index using the MC-DINA model with a small test length. The JSD index is equivalent to MI in the context of multinomial distributions, where it measures the relative entropy between the product of the marginal distributions of two random variables and the joint distribution of those random variables. A measure of item discrimination, the JSD (Yigit et al., 2019) for item j is shown in Equation 2.28.

$$JSD_j = S\left(P_j \times \pi'\right) - \sum_{\alpha}^{2^K} \pi_\alpha S\left(P_{j\alpha}\right) \tag{2.28}$$

In Equation 2.28, P_j represents a matrix that is the number of options, H, by the number of possible attribute patterns, 2^K, where column $\boldsymbol{\alpha}$ is the probability of selecting each of the H item options for examinees that belong to attribute profile, α, $S\left(P_j \times \pi'\right)$ and $S\left(P_{j\alpha}\right)$ represent

calculating SHE for $P_j \times \pi'$ and $P_{j\alpha}$, respectively, and π_α is the posterior distribution weights for each $\boldsymbol{\alpha}$. Through simulation, the MC-DINA was compared to random selection using the JSD index. The JSD performed better by 15-25% for items with high discrimination and by 6-26% for items with low discrimination for the MC-DINA compared to random selection for whole pattern classification accuracy.

Even though many methods have been proposed and tested out in CD-CAT, not all are created equal. Depending on the goal of the test and the client, some methods may be better than others for item selection. Early methods consisted of Shannon Entropy (SHE; Tatsuoka, 2002) and Kullback-Liebler information (KL; Xu et al., 2003), which included extensions of KL like posterior weighted KL (PWKL; Cheng, 2009) and the Hybrid KL (HKL; Cheng, 2009). Attribute-balancing procedures like the modified maximum global discrimination index (MMGDI; Cheng, 2010) and item exposure control procedures, for example, the restrictive progressive method (RP_PWKL; Wang et al., 2011) and the restrictive threshold method (RT_PWKL; Wang et al., 2011) were also proposed. For multinomial models, the Jensen-Shannon divergence (JSD: Yigit et al., 2019) was the only method used. Every method has its strengths and weaknesses, and some are best suited for specific testing situations.

Out of the many studies discussed here, most item selection methods for CD-CAT outperform a random fixed form by about 10-15%. In addition to some methods performing better than others, most comparisons are always made to a randomly constructed fixed test. As a result, very little is known about how much better these methods perform compared to a well-designed fixed-length test. For this study, the goal was to first extend expected Shannon Entropy to a multinomial model. Next, the goal was to explore the effectiveness of two item selection methods for CD-CAT using theoretical correct classification rates of multiple-choice items

compared to the multinomial extension of the expected Shannon Entropy. Along with these CD-CAT methods, the last goal was to determine how much better they perform compared to a random fixed form test and a well-designed fixed-length test. Overall, the aim was to see whether CAT is a better solution for diagnostic classification models compared to a well-designed fixed form test.

CHAPTER III: METHODS

This chapter is divided into three sections and presents a simulation study to demonstrate the effectiveness of two newly proposed CAT methods, as well as a comparison with three other assessment approaches using a multinomial model. In the first section, data generation and simulation conditions will be described, including the Q-Matrix, item bank, number of replications, item parameters, number of attributes, and number of examinees. The second section focuses on item selection methods, including three methods for CAT, and two ways to generate fixed forms. The last section includes evaluation measures that demonstrate how effectively each item selection method correctly classifies each attribute, as well as how they control for item exposure.

Data Generation and Simulation Conditions

This study was a simulation that fully crossed three factors. The three factors were test length, number of attributes, and simulated item parameters, resulting in 16 simulation conditions. Each condition had a total of 50 replications, with the exception of the multinomial extension of SHE, which was only run for 10 replications for the condition with six attributes and 10 items across item parameter conditions. Using GDCM Simulation Software (Sessoms et al., 2019), built in *R* (R Core Team, 2022), the Q-matrix, item parameters, and ultimately response data for each examinee were simulated. The Q-matrix, within each replication, was randomly simulated for 300 four-option multiple-choice items. The study factors included two different attribute conditions, two test length conditions, and four simulated item parameter conditions.

When using GDCM Simulation Software (Sessoms et al., 2019), many options can be manipulated for a study. The key inputs for this study for the ERUM-MC model included the

number of examinees, the number of replications, the number of attributes, the number of items for the CAT item bank, the number of response options, and the distribution of the item parameters. Using a fixed-length stopping rule for CAT, this study used both 5- and 10-items to calculate the probability of mastery for two attribute conditions, four and six attributes. Having two test lengths allows for the ability to determine whether more items are needed to determine attribute mastery in CD-CAT or if it can be accomplished with a shorter test.

When selecting the number of attributes to be simulated, the simulation software provides the option to choose the number of skills and misconceptions for each attribute case, which, when added together, equals the total number of attributes. Through the software (Sessoms et al., 2019), the maximum number of skills and misconceptions that can influence a correct or incorrect option can be chosen when setting up the simulation. These values should be less than the number of skills and misconceptions, respectively, for each attribute case. For the 4-attribute case, two skills and two misconceptions were modeled, where the maximum number of skills per correct option was 2, as an item can be difficult to build if it measures too many skills. Additionally, one skill was assigned per incorrect option, as incorrect options should not be influenced by too many skills. For misconceptions, the maximum number influencing a correct option was 1, as there should be more skills than misconceptions associated with a correct option. There was a maximum of one misconception for incorrect options, as they usually focus on a specific misconception.

For the 6-attribute case, three skills and three misconceptions were modeled, similar to what a real test would resemble (DiBello et al., 2015), where the max number of skills per correct option was 2, as an item can be difficult to build if it measures too many skills, and was 1 per incorrect option as incorrect options should not be influenced by too many skills. For

misconceptions, the maximum number influencing a correct option was 1, while there was a maximum of 1 misconception for incorrect options for the 6-attribute case. Since sometimes attributes can represent not only skills but also misconceptions, the terminology can be framed in terms of whether an attribute is possessed or not. In this dissertation, though, the term attribute will encompass both skills and misconceptions and will be indicated as mastered or not mastered.

Based on the entries in the Q-matrix, the item parameters were simulated for the 300 items. Using the ERUM-MC model, π and r were simulated at the option level to represent a high and low profile with $\pi \sim U(0.5,0.75)$ and $\pi \sim U(0.75,0.95)$, respectively (Dibello et al., 2015). The values for the r parameters were simulated using the distributions $r \sim U(0.1,0.3)$ and $r \sim U(0.3,0.5)$ to represent high and low cognitive structure (DiBello et al., 2015). Having a π closer to 1 for one of the options of an item would mean an easier detection of attribute mastery, while having π of 0.5 would be less likely to detect attribute mastery, so having two uniform distributions for π allows for more common conditions with real items. When simulating r, a combination of options with weak discrimination, $U(0.3,0.5)$, and high discrimination, $U(0.1,0.3)$, give conditions of real items in practice (DiBello et al., 2015).

With respect to simulating the examinee distribution of ability, a uniform latent distribution $P(\alpha) = \frac{1}{2^k}$ for every α, where k is the number of attributes, was assumed where each α is equally likely to happen and was randomly chosen for each examinee. In CAT, the focus is on attribute profile estimation, which depends on the number of items given to each examinee, with the goal being to estimate the attribute profile in fewer items than in a fixed form test. Additionally, item parameters are treated as known, so the sample size does not need to vary. Therefore, the sample size was selected to ensure a reasonable estimation of correct classification

rates for examinees. The latent attribute profiles for 1,000 examinees α, are simulated using the uniform latent distribution. An item bank of 300 items was available to use for CAT, and examinee response data will be simulated for all items based on the examinee's true attribute profile, the simulated item parameters, and the Q-matrix. By simulating the responses for all items, the difference in CAT approaches will be specifically with respect to item selection, as opposed to small random differences in the simulated responses. In CD-CAT, every examinee received the same first well-constructed item based on the criteria being used for item selection. However, subsequent items were chosen for the examinee based on the information an item provides in helping estimate the examinee's attribute profile.

Item Selection Methods

Given the simulated responses to all 300 items, the goal is to compare two new CAT approaches that use theoretical CCRs to a multinomial extension of the expected Shannon Entropy procedure (SHE; Tatsuoka, 2002). In addition, these three methods for CD-CAT were compared to fixed forms that have been constructed in two different ways. The two new methods are item selection based on theoretical CCRs for each attribute of items and item selection based on a composite computed using the posterior probability of mastery from the previous item as the weights for the attribute-level theoretical CCRs.

For the two CAT item selection methods based on theoretical CCRs, as described in Chapter 1, within each replication, each examinee began with the same item, which is the item that has the highest CCR for a specific attribute based on the simulated item bank. For the multinomial extension of Shannon entropy, each examinee will begin with the same item, which is the exact item that minimizes the expected Shannon Entropy when the posterior probability of

each attribute pattern is equally likely, or $P(\alpha) = \frac{1}{2^k}$. The fixed-length stopping rules that were used were test lengths of 5 and 10.

Single Attribute CCR Method

The first new method proposed examined the use of theoretical CCRs (Stout et al., 2023) for each attribute for the items. For each item bank, the theoretical correct classification rates of the items need to be calculated. The theoretical CCRs, which can be thought of as an item discrimination index, show how effective each item, when administered in isolation, is in diagnosing an attribute *k*. The theoretical CCRs (Stout et al., 2022) are computed by Equation 3.1

$$CCR(k) = \left(\frac{1}{2}\right) \sum_h Max_{a=0,1} P(h|A_k = a) \text{ for each k} \tag{3.1}$$

where the sum across *h* represents the sum across the number of options for an item (in this study it was four), *h* is the simulated examinee's test response based on random simulated examinee latent attribute profile, *A*. The item CCRs were stored for each condition replication and used to choose the next item. Once everyone received the first item, the next item was chosen based on the posterior probability of mastery for each examinee. The goal is to identify the attribute with the smallest absolute distance from the posterior probability of mastery, PPM_k, to 0.5, which provides the attribute with the least amount of certainty with respect to mastery or nonmastery. Given the attribute, the next item was selected such that the $CCR(k)$ is the highest. This process continued to choose the next items for CD-CAT until the expected test length was reached, whether it was 5 or 10.

Composite CCR Procedure

When determining the best approach to choose the next item in CAT, it is often difficult when trying to estimate the attribute profile efficiently. The first newly proposed method uses the marginal posterior probability of mastery. Sometimes, using a joint classification method may be more efficient, as it gathers information from all attributes, not just one. A second method proposed and explored in this study expands the previous method. Rather than selecting items based on the attribute with the least information (i.e., the posterior probability of mastery is close to 0.50), we calculate a composite value that uses the posterior probability of mastery, PPM_k, as weights for the theoretical CCRs for each item. The item with the highest composite value was then selected. Specifically, after gathering responses to the items, the PPM_k values are computed, and for each item, a composite score was calculated using the item's CCR_k using Equation 3.2.

$$Z_i = \sum_k \min(PPM_k, 1 - PPM_k) * CCR_k. \tag{3.2}$$

The minimum value between PPM_k and 1-PPM_k was found and then multiplied across the theoretical CCRs for all items in the item bank. This approach resulted in a composite score, Z_i, for every item available in the item bank to be selected. Once each examinee received the first item, the next item was chosen based on the item where Z_i was the largest. This process continued to select the next items for CAT until the expected test length was reached whether it was 5 or 10.

Shannon Entropy Procedure

The expected Shannon Entropy (SHE; Tatsuoka, 2002) has been established and studied in the literature for dichotomous models, but it has not been extended to multinomial models.

SHE was extended to multinomial models in this study, calculating the sum across all four options of the items given by the following equation

$$Sh(\pi_n, X_j) = \sum_{0}^{3} \left\{ E_n(\pi_n | X_j = x) \left(\sum_{c=1}^{2^M} P_j^x(\alpha_c) \left[1 - P_j(\alpha_c)\right]^{1-x} \pi_{n-1}(\alpha_c) \right) \right\} \tag{3.3}$$

SHE was calculated by multiplying the posterior probability of each attribute pattern after the previous item was administered by the probability of a correct response for each option of each item given an attribute pattern and then summing across all attribute patterns. That value was then multiplied by the Shannon entropy of the estimated posterior probability of each attribute pattern given that an examinee responds to an item with a given option. This result was then summed across all options, resulting in the expected SHE for each item. Note that SHE depends on the estimated posterior probabilities, and as a result, the expected posterior probabilities will change after each item is given. Therefore, for each step, the SHE must be recomputed for each item using the process defined above. Given these values, the next item was selected to minimize SHE (Chen, 2009; Wang, 2013; Xu et al., 2016). SHE was recalculated after each item was administered for the items remaining in the item bank, and a new item was chosen based on the item that minimized SHE. This process continued to choose the next items for CAT until the expected test length was reached, whether it was five or 10 items.

Fixed Forms

Most of the time in the literature, when comparing methods of item selection for CD-CAT, there is also a comparison to a random fixed form. While this may be a useful and valid

comparison, it may oversimplify what would happen in an actual application. Specifically, it is believed that a fixed form would be created in a strategic way using well-selected items to ensure the measurement of all attributes. In this study, two methods for choosing fixed forms were used to compare with the three CAT item selection strategies. The first method involved selecting a random set of items from the item bank for each test length within each condition, allowing direct comparisons with results obtained in previous studies of CAT. For each simulation replication, five fixed forms will be constructed.

However, traditionally, better-performing items are chosen when constructing a well-designed, fixed form test; therefore, for the second method, "good" items were selected. The second method involved constructing a fixed form based on items chosen for their favorable item properties, specifically, high theoretical CCR values for attributes. Looking at the Q-matrix, see which items measured attribute 1, and then determine which one of the items had the largest theoretical CCR value, and that was item one. This pattern was followed for item two with attribute two, item three with attribute three, and so on. Once each attribute has been used once, pick an item that has the next largest theoretical CCR for any of the attributes and follow that pattern until the projected test length was reached (e.g., five or 10 items). One form was built for each replication within each condition. For the specific case of a five-item test that measured six attributes, the five attributes with the largest theoretical CCRs were identified, and the Q-matrix was used to determine which items were measured by those attributes. The five items that corresponded to those five attributes were then selected.

Evaluation Criteria

Four evaluation criteria were used for this study to answer the research questions. These included the attribute level CCRs across all attributes and all examinees along with attribute

profile CCRs, item exposure, the average absolute deviations of the posterior probabilities from 0.5, and average time.

When the response data for each examinee was simulated, a true attribute profile showed whether each examinee was a master of each attribute or not. After running each item selection method, the posterior probability of mastery for each examinee was calculated. Since these are probabilities, they were rounded to the closest value, either 0 or 1, then compared to the true value that was simulated. If the values are equal, then that examinee received a 1 for that attribute; otherwise, it was a 0. The values for each attribute were averaged across examinees, resulting in the attribute-level CCRs across all attributes and all examinees. Another measure for looking at classification is examining the attribute profile CCRs. Similar to how the correct classification rate was calculated at the attribute level, the attribute profile CCRs compared the true simulated attribute patterns for each examinee with the posterior probabilities of mastery across the entire attribute profile.

Item exposure can be critical to look at in computer adaptive tests, especially when high-stakes testing is involved, and is informative about how the items in an item bank are performing. When items are over-exposed, examinees may have prior knowledge about the items, as many have seen them before, and the validity of the items may be compromised, along with the examinees' ability being inflated (Wang et al., 2011). Underexposure of items can also be critical, as companies develop large item banks for use in CAT. However, if some items are rarely used, they waste the money spent on developing the item banks (Wang et al., 2011). Item exposure rates are calculated for each simulated item bank that corresponds with each condition to determine the proportion of items used 0%, between 0 and 25%, between 25 and 50%, between 50 and 90%, and between 90 and 100% of the time.

The third evaluation criterion was calculating the average absolute deviations of the posterior probabilities from 0.5. When the posterior probability of mastery is 0.5 or close to 0.5, there is little to no information about whether the examinee is a master or not of that attribute. The farther the posterior probability is from 0.5, the more information about mastery there is. Using this as an evaluation measure allowed for another way to determine how well mastery was achieved through each CD-CAT item selection method.

The average time for each replication will be the last evaluation criterion. While comparing the performance of each CD-CAT method, the efficiency is also a good metric to look at to determine which method is best to use. The average time will be calculated using the average number of days it takes for 1 replication.

Overall, this study aimed to evaluate the performance of two new item selection methods for CD-CAT based on theoretical correct classification rates compared to the multinomial extension of expected Shannon Entropy, a random fixed form, and a well-designed fixed form using three different evaluation measures. The research questions will be evaluated by calculating the attribute-level correct classification rates across all attributes and examinees and attribute profile CCRs, as well as calculating the proportion of items exposed in each item bank and the average absolute deviation of the posterior probabilities from 0.5. This analysis will be done by fully crossing three factors, which include test length, number of attributes, and simulated item parameters.

CHAPTER IV: RESULTS

This study was designed to evaluate whether using a multinomial diagnostic classification model the ERUM-MC, with CCR-based selection outperforms well-designed fixed forms. A simulation approach was employed in this study, considering factors such as the number of attributes measured, test length, and variation in simulated item parameters. The results are presented in three sections, summarizing the three evaluation criteria used in the simulation study. First, attribute-level CCRs across all attributes and all examinees were calculated for all methods, including the multinomial extension of SHE, both theoretical CCR methods and both fixed-form methods, along with whole-pattern recovery rates. Second, to gain a deeper understanding of how well each method determines mastery, the average absolute deviations from 0.5 were calculated for all five methods. Lastly, average item exposure rates were calculated for the three CD-CAT methods at five different intervals.

In each figure, the colored lines represent CD-CAT or fixed form methods that were evaluated and compared in the simulation. “Single attribute” represents the single attribute theoretical CCR method, while “Composite” represents the composite theoretical CCR method, and “SHE” represents the multinomial extension of the Shannon Entropy procedure.

Recall that theoretical CCRs serve as an item-discrimination index, showing how well a single item can diagnose an attribute, and were computed for each condition and stored for item selection. For the single attribute theoretical CCR method, after all examinees answered the first item, the next item was chosen by finding the examinee’s attribute whose posterior probability of mastery was closest to 0.5—that is, the attribute with the greatest uncertainty—and then selecting the item with the highest theoretical CCR for that attribute. A second proposed method, the composite attribute theoretical CCR method, selects items using a weighted composite score

rather than focusing only on the least-certain attribute. After collecting responses, posterior probabilities of mastery were calculated. For each item, a composite score was computed by finding the minimum of the posterior probability of mastery and one minus the posterior probability of mastery and multiplying it by the item's theoretical CCRs across all attributes. The next item administered was the one with the highest composite score. The multinomial extension of SHE was computed by first multiplying the posterior probability of each attribute pattern (after the previous item) by the probability of a correct response for each option of each item and summing across patterns. This result was then multiplied by the Shannon entropy of the posterior probability of each attribute pattern given a specific response option and summed across all options to give the item's expected SHE. The next item chosen was one that minimized SHE. "Fixed" represents a fixed form test built with items chosen at random. "Good Fixed" represents a well-designed fixed form test. The well-designed fixed form was built by selecting items with the highest theoretical CCRs for each attribute. Using the Q-matrix, the item measuring attribute 1 with the largest CCR became item 1; the same process was repeated for attribute 2, attribute 3, and so on.

In Figures 1-8, the x-axis represents each of the item parameter conditions. The y-axis for Figures 1,3, 5, and 7 represents the attribute-level CCRs across all attributes and all examinees, and for Figures 2, 4, 6, and 8, it represents the attribute profile CCRs, ranging from 0 to 1 or 0 to 100% (based on whether results are reported as proportions or percentages, respectively). In Figures 9-12, the y-axis represents the absolute deviation from 0.5, which ranges from 0 to 0.5. For Figures 13-16, the x-axis represents the item exposure rates to the examinees for the items in the item bank, including items exposed 0%, between 0 and 25%, between 25 and 50%, between 50 and 90%, and between 90 and 100% of the time. The y-axis of Figures 9-12 represents the

proportion of items exposed from the item bank for each corresponding exposure rate for each CD-CAT method.

For Figures 1-6, 9-11, and 13-15, all methods had 50 replications completed. For Figures 7, 8, 12, and 16, all methods, except SHE, had 50 replications completed. SHE only had 10 replications completed. In each figure, four item parameter conditions are represented: HH, HL, LH, and LL. HH represents when $\pi \sim U(0.75,0.95)$, where items give more information about attribute mastery and $r \sim (0.3,0.5)$, where the options have weak discrimination. HL represents when $\pi \sim U(0.75,0.95)$, where items give more information about attribute mastery, and $r \sim (0.1,0.3)$, where the options have a high discrimination. LH represents when $\pi \sim U(0.50,0.75)$, where items give less information about attribute mastery, and $r \sim (0.3,0.5)$ meaning the options have weak discrimination. Finally, LL represents when $\pi \sim U(0.50,0.75)$, where items give less information about attribute mastery, and $r \sim (0.1,0.3)$, meaning the options have high discrimination.

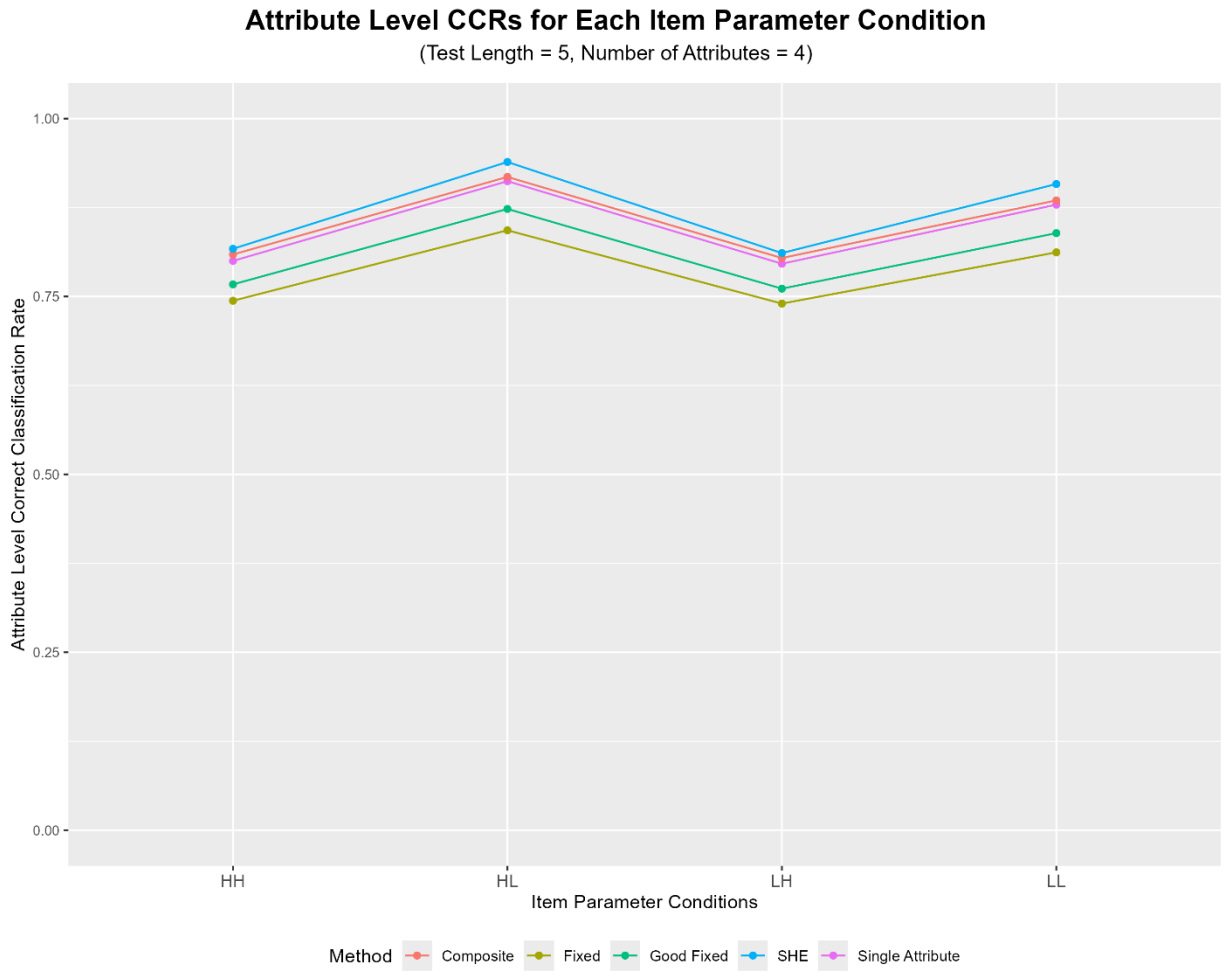
Average Correct Classification Rates

Attribute-level correct classification rates (CCRs) and attribute profile CCRs for each of the simulation conditions are presented in Figures 1 to 8. Attribute-level CCRs were calculated by comparing the mastery value for each attribute in the true attribute profile to the posterior probability of mastery of each attribute for each examinee, rounded to 0 or 1, representing the proportion of times that an examinee was correctly classified as a master or nonmaster. Attribute profile CCRs were calculated similarly, but the complete true attribute profile was compared to the posterior probabilities of mastery across the whole attribute profile.

4 Attributes and 5 Items

Figure 1 displays the attribute-level CCRs for all item parameter conditions for the condition with four attributes and a test length of five items. Across all item parameter conditions, the multinomial extension of SHE correctly classified each attribute most often. The single-attribute and composite theoretical CCR methods yielded similar attribute-level CCRs across all attributes. Fixed form performance differed slightly between random fixed forms and well-designed fixed forms.

Figure 1. Attribute-Level Correct Classification Rates for 4 Attributes with a Test Length of 5



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

Table 2 displays the attribute-level CCRs across all attributes. For the multinomial extension of SHE, the attribute-level CCRs ranged from 0.811 to 0.939 for the LH and HL item parameter conditions, respectively. The composite method performed slightly better than the single attribute method, about 1% on average, across all attributes. Attribute-level CCRs for the composite attribute theoretical CCR method ranged from 0.804 to 0.918 for the LH and HL item parameter conditions, respectively. For the single attribute theoretical CCR method, attribute-level CCRs ranged from 0.796 to 0.912 for the LH and HL item parameter conditions, respectively.

Well-designed fixed forms had attribute-level CCRs approximately 0.026 higher than those of random fixed forms. Attribute-level CCRs, as displayed in Table 2, for the well-designed fixed form method ranged from 0.761 to 0.873 for the LH and HL item parameter conditions, respectively. For the random fixed form method, attribute-level CCRs ranged from 0.740 to 0.843 for the LH and HL item parameter conditions, respectively. For Table 2, comparing fixed forms to the CD-CAT methods, the multinomial extension of Shannon Entropy performs better than random fixed forms by an average of 0.085, and on average about 0.060 better than a well-designed fixed form, with the biggest gap in performance for the HL and LL item parameter conditions, respectively.

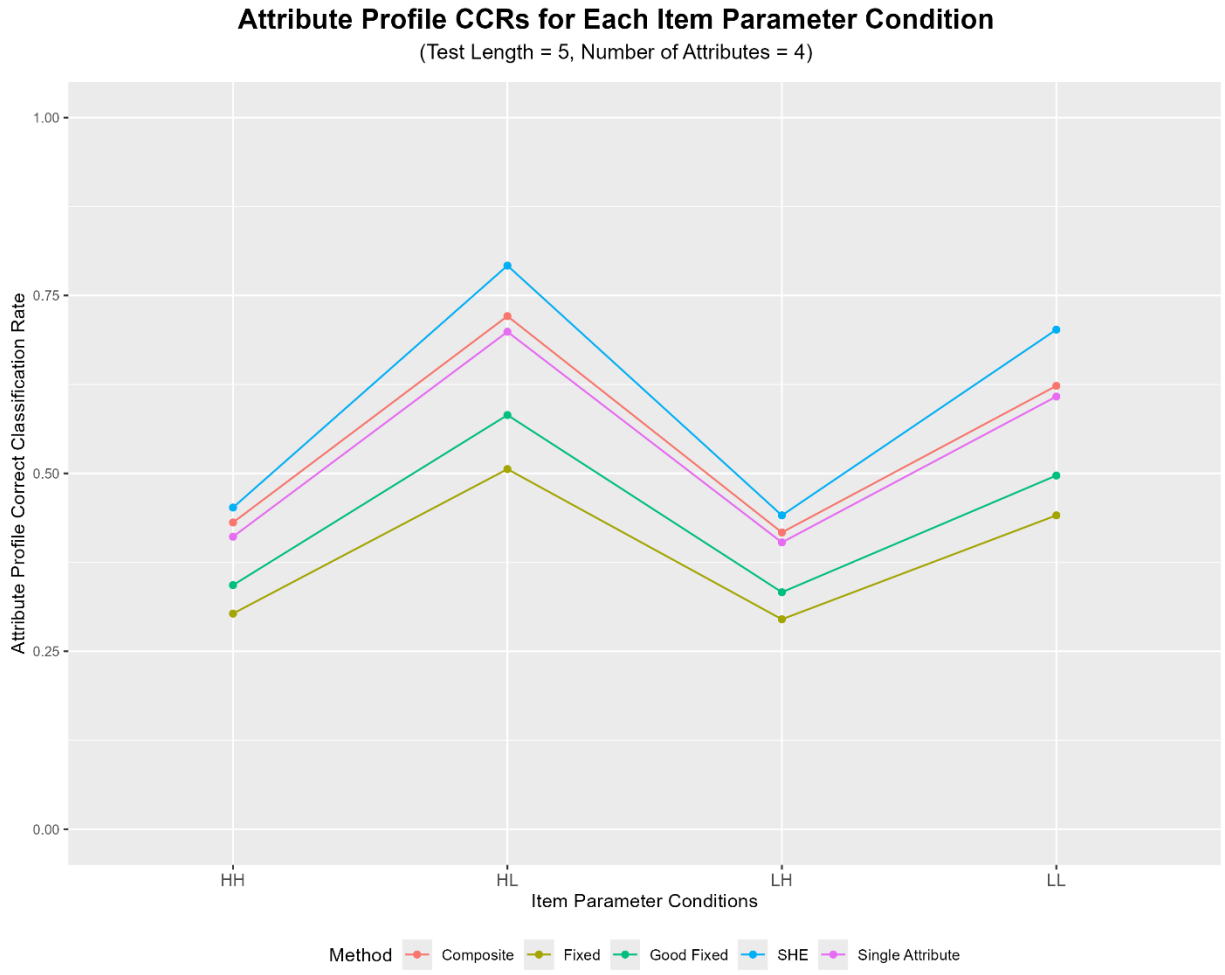
Table 2. Attribute-Level Correct Classification Rates for 4 Attributes with a Test Length of 5

Method	Attribute-Level CCRs Across All Attributes			
	HH	**HL**	**LH**	**LL**
Single Attribute	0.800	0.912	0.796	0.879
Composite	0.809	0.918	0.804	0.885
SHE	0.817	0.939	0.811	0.908
Fixed	0.744	0.843	0.740	0.812
Good Fixed	0.767	0.873	0.761	0.839

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

Figure 2 displays the attribute profile CCRs for all item parameter conditions for the condition with four attributes and a test length of five items. The multinomial extension of SHE had the highest attribute profile CCRs across all item parameter conditions, with the largest difference between the multinomial extension of SHE and the other methods being for the HL and LL item parameter conditions.

Figure 2. Attribute Profile Correct Classification Rates for 4 Attributes with a Test Length of 5



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

Table 3 displays the attribute profile CCRs for the condition with four attributes and a test length of five items. For the multinomial extension of SHE, attribute profile CCRs range from 0.441 to 0.792 for the LH and HL item parameter conditions, respectively. The single attribute and composite theoretical CCR methods had similar attribute profile CCRs across all

item parameter conditions. The composite method performed slightly better, 0.018 on average, for overall pattern recovery. Attribute profile CCRs for the composite attribute theoretical CCR method ranged from 0.417 to 0.721 for the LH and HL item parameter conditions, respectively. For the single attribute theoretical CCR method, attribute profile CCRs ranged from 0.403 to 0.699 for the LH and HL item parameter conditions, respectively. Attribute profile CCRs for the well-designed fixed form method ranged from 0.333 to 0.582 for the LH and HL item parameter conditions, respectively. Attribute profile CCRs for random fixed forms ranged from 0.295 to 0.506 for the LH and HL item parameter conditions, respectively. For Table 3, comparing fixed forms to the CD-CAT methods, the multinomial extension of Shannon Entropy performs, on average, 0.210 better than random fixed forms and 0.158 better than a well-designed fixed form, with the biggest gap in performance observed for the HL item parameter conditions.

Table 3. Attribute Profile Correct Classification Rates for 4 Attributes with a Test Length of 5

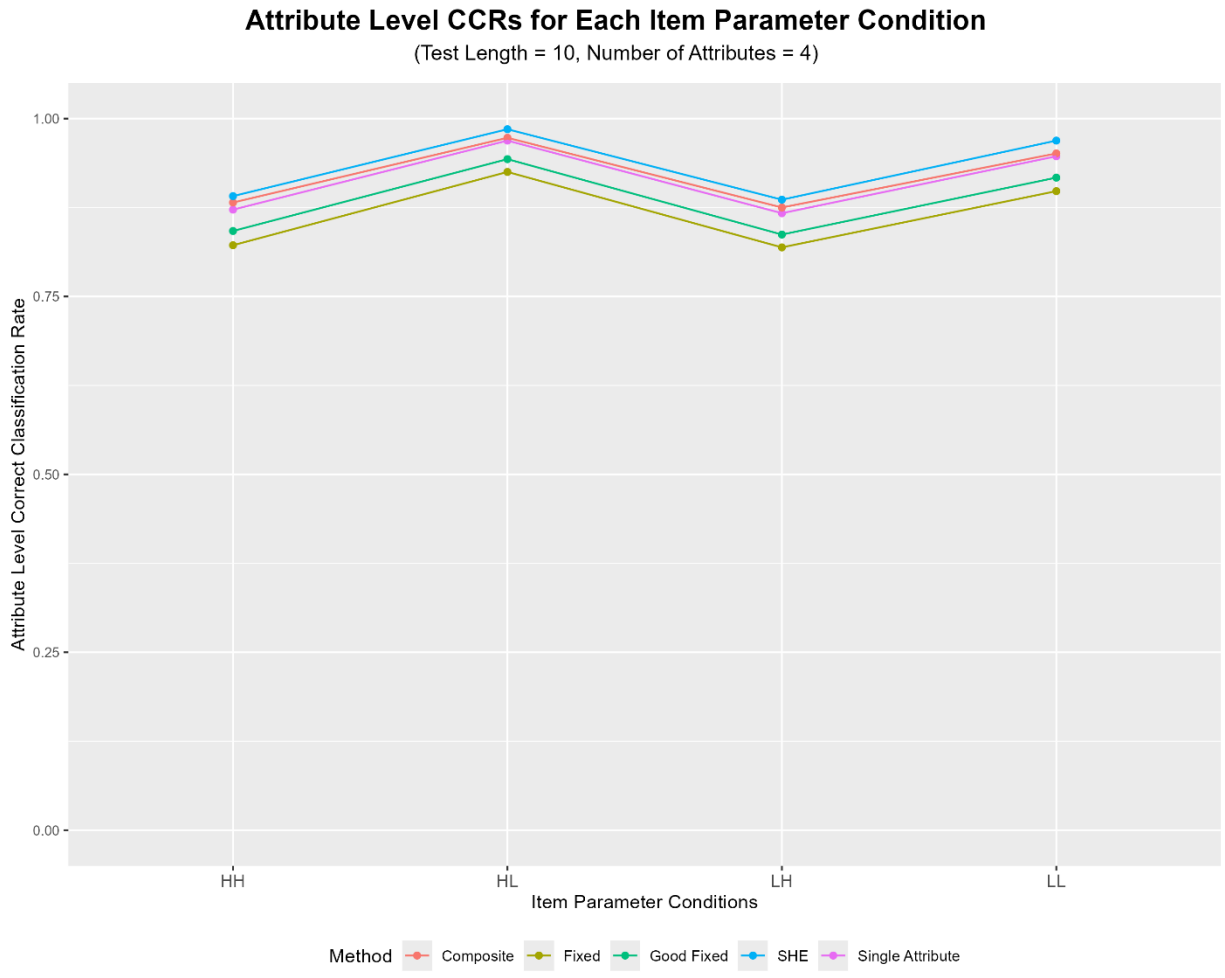
Method	**Attribute Profile CCRs**			
	HH	**HL**	**LH**	**LL**
Single Attribute	0.411	0.699	0.403	0.608
Composite	0.431	0.721	0.417	0.623
SHE	0.452	0.792	0.441	0.702
Fixed	0.303	0.506	0.295	0.441
Good Fixed	0.343	0.582	0.333	0.497

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

4 Attributes and 10 Items

Attribute-level CCRs for all item parameter conditions for the condition with four attributes and a length of 10 items are displayed in Figure 3. Similar to the condition with four attributes and a test length of five, the multinomial extension of SHE outperforms the other two CD-CAT methods, single attribute, and composite, along with both types of fixed form assembly, by achieving the highest CCRs across all attributes and the overall pattern recovery rate.

Figure 3. Attribute-Level Correct Classification Rates for 4 Attributes with a Test Length of 10



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

Attribute-level CCRs for the multinomial extension of SHE, as displayed in Table 4, ranged from 0.886 to 0.985 for the LH and HL item parameter conditions, respectively. The single attribute and composite theoretical CCR methods had similar attribute-level CCRs, with the composite method performing slightly better, 0.007% on average, than the single attribute method for attributes. For the composite attribute theoretical CCR method, attribute-level CCRs ranged from 0.875 to 0.973, for the LH and HL item parameter conditions, respectively. For the single attribute theoretical CCR method, attribute-level CCRs ranged from 0.867 to 0.969 for the LH and HL item parameter conditions, respectively.

Fixed form performance for the condition with four attributes and a test length of 10 differed slightly between random fixed forms and well-designed fixed forms, as displayed in Table 4. Well-designed fixed forms had attribute-level CCRs 0.019 higher than those of random fixed forms. Attribute-level CCRs for the well-designed fixed form method ranged from 0.837 to 0.973 for the LH and HL item parameter conditions, respectively. For the random fixed form method, attribute-level CCRs ranged from 0.819 to 0.925 for the LH and HL item parameter conditions, respectively. For Table 4, comparing fixed forms to the CD-CAT methods, the multinomial extension of Shannon Entropy performs 0.066 better than random fixed forms and 0.048 better than a well-designed fixed form, on average, with the biggest gap in performance for the LL item parameter condition.

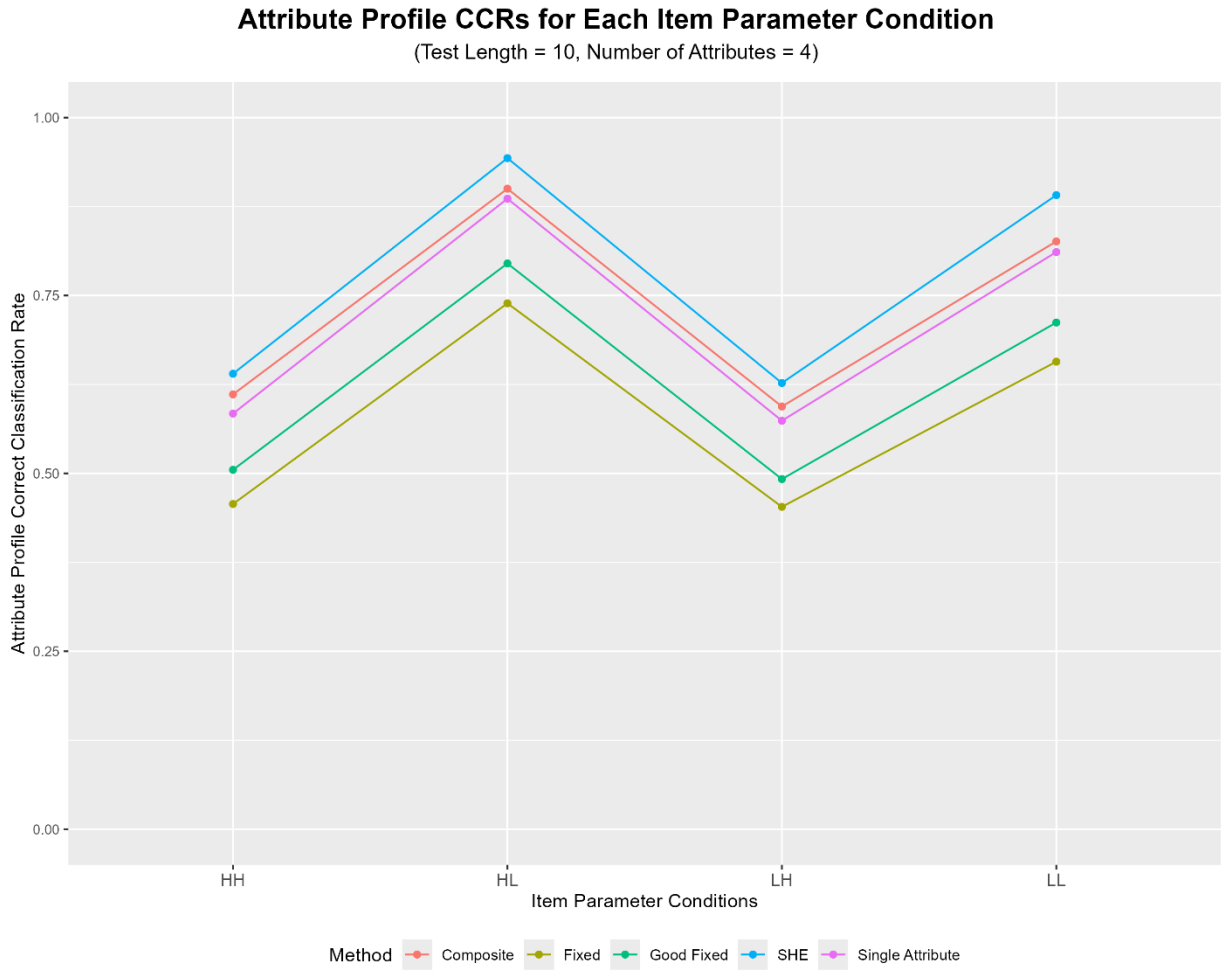
Table 4. Attribute-Level Correct Classification Rates for 4 Attributes with a Test Length of 10

Method	Attribute-Level CCRs Across All Attributes			
	HH	**HL**	**LH**	**LL**
Single Attribute	0.872	0.969	0.867	0.947
Composite	0.882	0.973	0.875	0.951
SHE	0.891	0.985	0.886	0.969
Fixed	0.822	0.925	0.819	0.898
Good Fixed	0.842	0.943	0.837	0.917

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

Figure 4 displays the attribute profile CCRs for all item parameter conditions for the condition with four attributes and a test length of 10 items. The multinomial extension of SHE achieved the highest attribute profile CCRs across all item parameter conditions, with the largest difference between the multinomial extension of SHE and the other methods observed for the HL and LL item parameter conditions.

Figure 4. Attribute Profile Correct Classification Rates for 4 Attributes with a Test Length of 10



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

Attribute profile CCRs , as displayed in Table 5, for the multinomial extension of SHE ranged from 0.627 to 0.943 for the LH and HL item parameter conditions, respectively. The single attribute and composite theoretical CCR methods had similar attribute profile CCRs , with the composite method performing slightly better, 0.019 on average. for the LH and HL item

parameter conditions, respectively. For the composite attribute theoretical CCR method, attribute profile CCRs ranged from 0.594 to 0.900, for the LH and HL item parameter conditions, respectively. For the single attribute theoretical CCR method, attribute profile CCRs ranged from 0.574 to 0.886 for the LH and HL item parameter conditions, respectively.

Fixed form performance for the condition with four attributes and a test length of 10, as displayed in Table 5, differed slightly between random fixed forms and well-designed fixed forms. Attribute profile CCRs for the well-designed fixed form method ranged from 0.492 to 0.795 for the LH and HL item parameter conditions, respectively. For the random fixed form method, attribute profile CCRs ranged from 0.453 to 0.739 for the LH and HL item parameter conditions, respectively. For Table 5, comparing fixed forms to the CD-CAT methods, the multinomial extension of Shannon Entropy performs 0.199 better than random fixed forms and 0.150 better than a well-designed fixed form, on average, with the largest gap in performance observed for the LL item parameter condition for attribute profile CCRs

Table 5. Attribute Profile Correct Classification Rates for 4 Attributes with a Test Length of 10

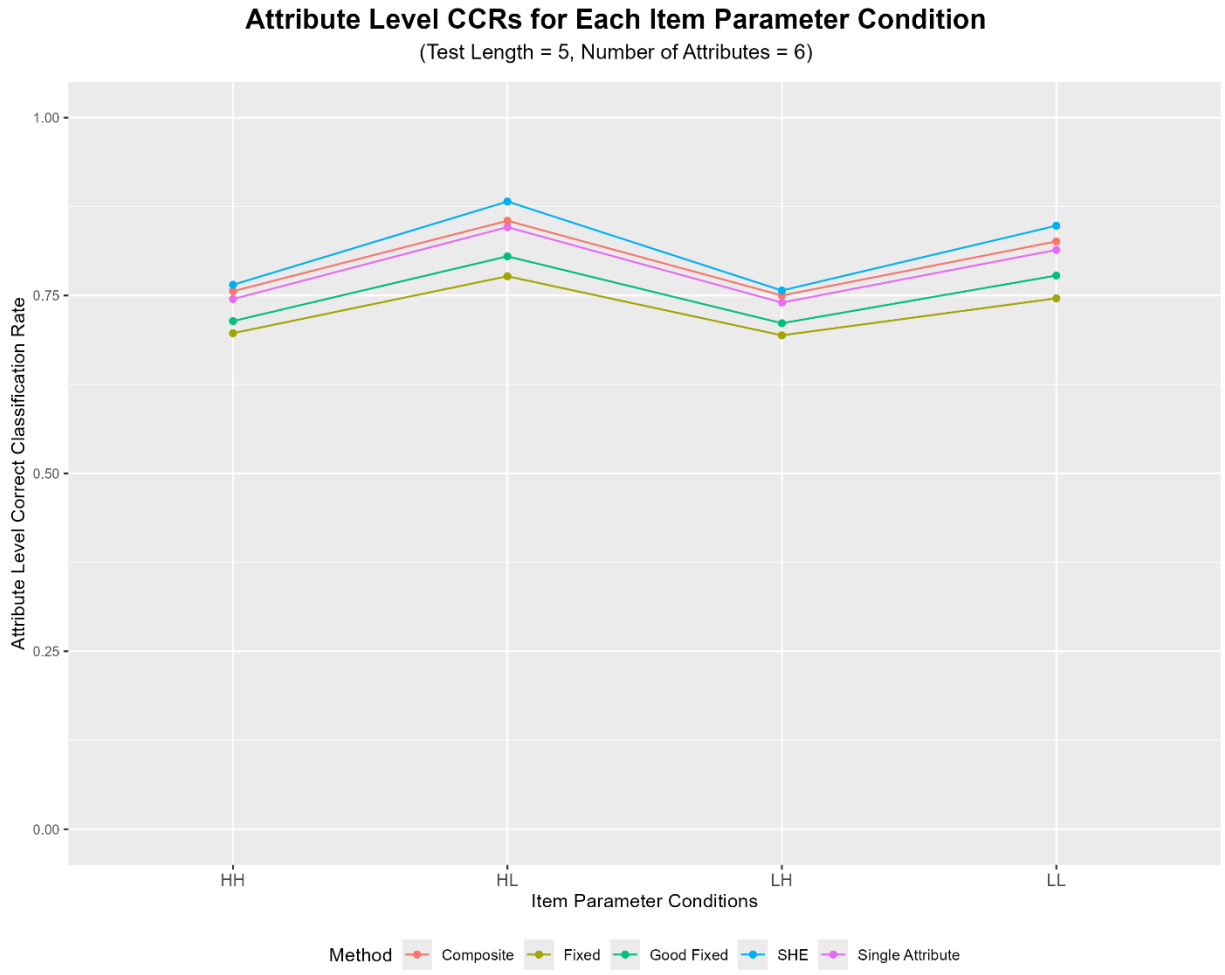
Method	**Attribute Profile CCRs**			
	HH	**HL**	**LH**	**LL**
Single Attribute	0.584	0.886	0.574	0.811
Composite	0.611	0.900	0.594	0.826
SHE	0.640	0.943	0.627	0.891
Fixed	0.457	0.739	0.453	0.657
Good Fixed	0.505	0.795	0.492	0.712

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

6 Attributes and 5 Items

Attribute-level CCRs for all item parameter conditions for the test condition with six attributes and a length of five items are displayed in Figure 5. Unlike the conditions with four attributes, the multinomial extension of SHE had only slightly higher attribute-level CCRs for all item parameter conditions. Fixed form performance for the condition with six attributes and a test length of five items, as displayed in Figure 3, differed slightly between random fixed forms and well-designed fixed forms.

Figure 5. Attribute-Level Correct Classification Rates for 6 Attributes with a Test Length of 5



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

The attribute-level CCRs, displayed in Table 6, for all attributes for the multinomial extension of SHE had ranged from 0.757 to 0.882 for the LH and HL item parameter conditions, respectively. The single attribute and composite theoretical CCR methods had similar attribute-level CCRs, with the composite method performing slightly better, 0.011 on average, than the

single attribute method for all attributes. For the composite attribute theoretical CCR method, attribute-level CCRs ranged from 0.750 to 0.855 for the LH and HL item parameter conditions, respectively. For the single attribute theoretical CCR method, attribute-level CCRs ranged from 0.740 to 0.846 for the LH and HL item parameter conditions, respectively. Well-designed fixed forms had attribute-level CCRs that were, on average, 0.024 higher than those of random fixed forms. Attribute-level CCRs for the well-designed fixed form method ranged from 0.711 to 0.805 for the LH and HL item parameter conditions, respectively. For the random fixed form method, individual CCRs ranged from 0.694 to 0.77 for the LH and HL item parameter conditions, respectively. For Table 6, which compares fixed forms to the CD-CAT methods, the multinomial extension of Shannon Entropy performs 0.085 better than random fixed forms and 0.061 better than a well-designed fixed form, on average, with the biggest gap in performance observed for the HL and LL item parameter conditions.

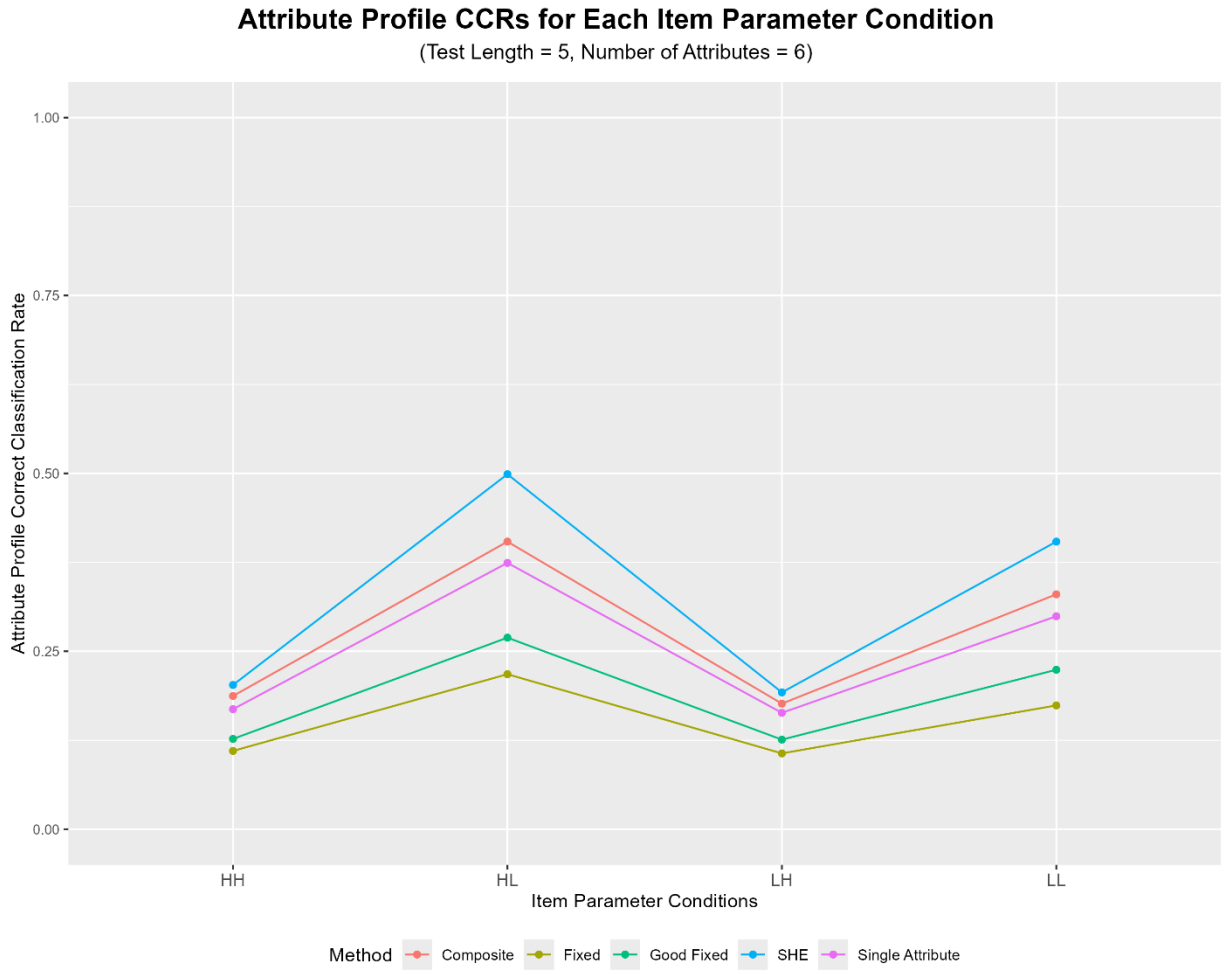
Table 6. Attribute-Level Correct Classification Rates for 6 Attributes with a Test Length of 5

Method	**Attribute-Level CCRs Across All Attributes**			
	HH	**HL**	**LH**	**LL**
Single Attribute	0.745	0.846	0.740	0.814
Composite	0.756	0.855	0.750	0.826
SHE	0.765	0.882	0.757	0.848
Fixed	0.697	0.777	0.694	0.746
Good Fixed	0.714	0.805	0.711	0.778

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

Figure 6 displays the attribute profile CCRs for all item parameter conditions for the condition with six attributes and a test length of five items. The multinomial extension of SHE achieved the highest attribute profile CCRs across all item parameter conditions, with the largest difference between the multinomial extension of SHE and the other methods being observed for the HL and LL item parameter conditions.

Figure 6. Attribute Profile Correct Classification Rates for 6 Attributes with a Test Length of 5



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

Table 7 displays the attribute profile CCRs for the condition with six attributes and five items. The attribute profile CCRs for the multinomial extension of SHE ranged from 0.196 to 0.491 for the LH and HL item parameter conditions, respectively. The single attribute and composite theoretical CCR methods had similar attribute profile CCRs , with the composite

method performing slightly better, 0.023 on average, than the single attribute method for attribute profile CCRs . Attribute profile CCRs for the composite method ranged from 0.176 to 0.404 for the LH and HL item parameter conditions, respectively. For the single attribute theoretical CCR method, attribute profile CCRs ranged from 0.164 to 0.374 for the LH and HL item parameter conditions, respectively.

Attribute profile CCRs for the well-designed method ranged from 0.126 to 0.269 for the LH and HL item parameter conditions, respectively. For the random fixed form method, attribute profile CCRs ranged from 0.107 to 0.218 for the LH and HL item parameter conditions, respectively. For Table 7, comparing fixed forms to the CD-CAT methods, the multinomial extension of Shannon Entropy performs 0.173 better than random fixed forms and 0.138 better than a well-designed fixed form, on average, with the largest gap in performance for the HL item parameter condition and the smallest gap for the LH condition.

Table 7. Attribute Profile Correct Classification Rates for 6 Attributes with a Test Length of 5

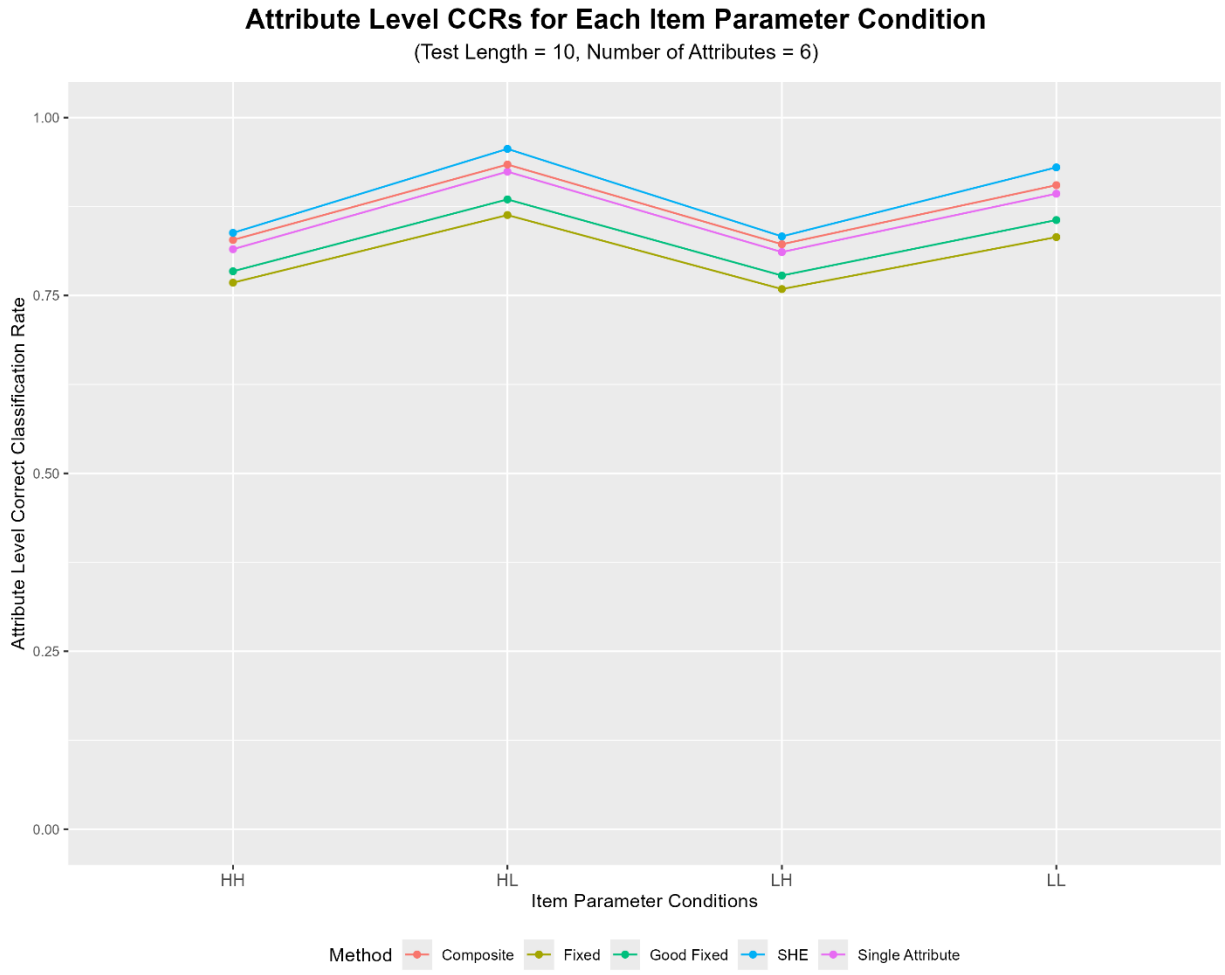
Method	**Attribute Profile CCRs**			
	HH	**HL**	**LH**	**LL**
Single Attribute	0.169	0.374	0.164	0.299
Composite	0.187	0.404	0.176	0.330
SHE	0.203	0.499	0.192	0.404
Fixed	0.110	0.218	0.107	0.174
Good Fixed	0.127	0.269	0.126	0.224

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

6 Attributes and 10 Items

Attribute-level CCRs for all item parameter conditions for the test condition with six attributes and a length of 10 items are displayed in Figure 7. Similar to the condition with 6 attributes and five items, the multinomial extension of SHE had only slightly higher attribute-level CCRs for all item parameter conditions. Fixed form performance for the condition with six attributes and a test length of 10 items, as displayed in Figure 3, differed slightly between random fixed forms and well-designed fixed forms.

Figure 7. Attribute-Level Correct Classification Rates for 6 Attributes with a Test Length of 10



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods were run for 50 replications, except for Shannon Entropy, which was run for only 10 replications.

Attribute-level CCRs, displayed in Table 8, for the multinomial extension SHE ranged from 0.833 to 0.956 for the LH and HL item parameter conditions, respectively. The single attribute and composite theoretical CCR methods had similar attribute-level CCRs, with the composite method performing slightly better, about 1% on average, than the single attribute

method for all attributes. For the composite attribute theoretical CCR method, attribute-level CCRs ranged from 0.822to 0.934 for the LH and HL item parameter conditions, respectively. For the single attribute theoretical CCR method, attribute-level CCRs ranged from 0.811 to 0.924 for the LH and HL item parameter conditions, respectively.

Well-designed fixed forms had attribute-level CCRs that were, on average, 0.020 higher than those of random fixed forms. Attribute-level CCRs for the well-designed fixed form method ranged from 0.778 to 0.885 for the LH and HL item parameter conditions, respectively. For the random fixed form method, individual CCRs ranged from 0.759 to 0.863. For Table 8, which compares fixed forms to the CD-CAT methods, the multinomial extension of Shannon Entropy performs, on average, 0.084 better than random fixed forms and 0.063, on average, better than a well-designed fixed form, with the largest gap in performance observed for the LL item parameter condition.

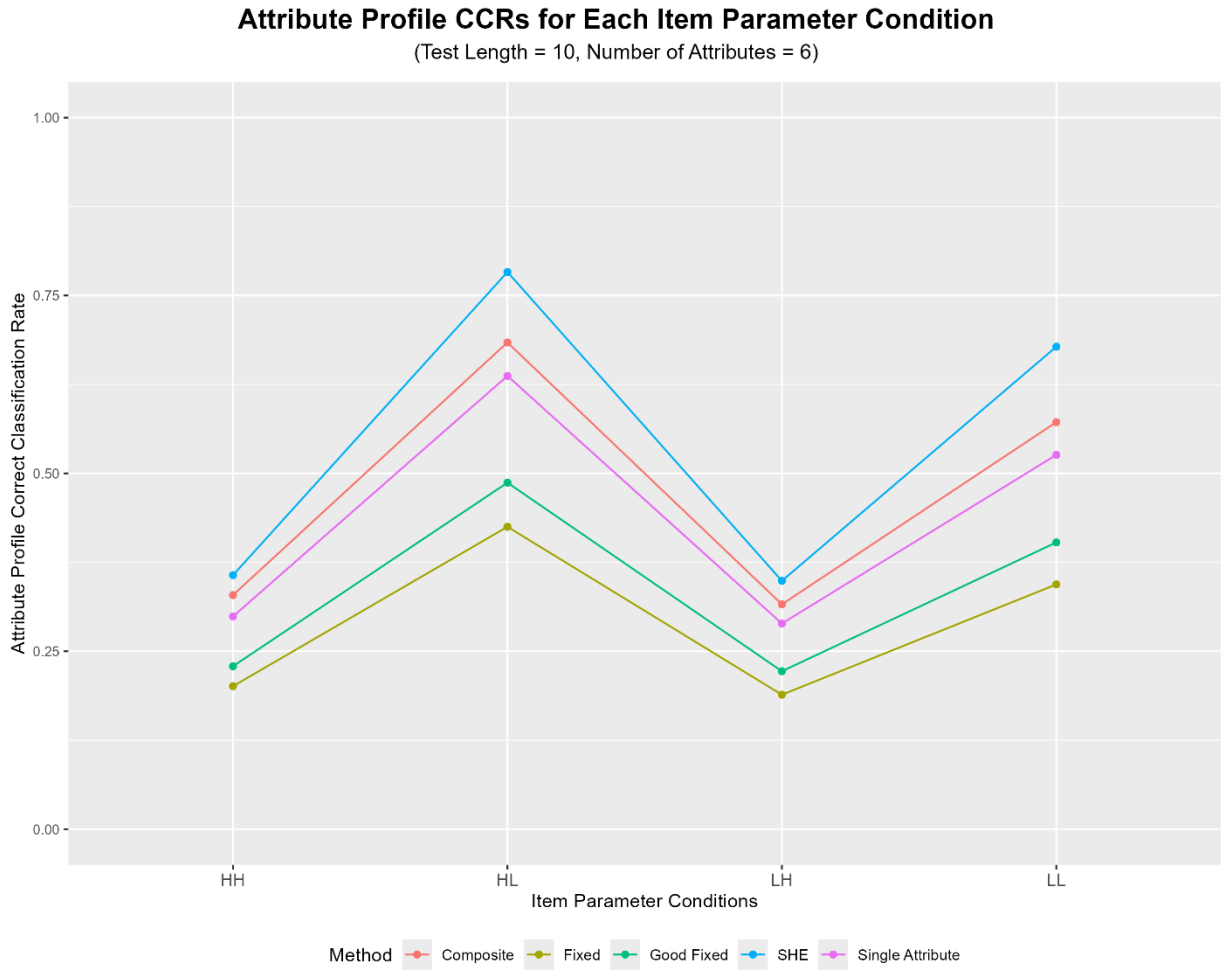
Table 8. Attribute-Level Correct Classification Rates for 6 Attributes with a Test Length of 10

Method	**Attribute-Level CCRs Across All Attributes**			
	HH	**HL**	**LH**	**LL**
Single Attribute	0.815	0.924	0.811	0.893
Composite	0.828	0.934	0.822	0.905
SHE	0.838	0.956	0.833	0.930
Fixed	0.768	0.863	0.759	0.832
Good Fixed	0.784	0.885	0.778	0.856

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods were run for 50 replications, except for Shannon Entropy, which was run for only 10 replications.

Figure 8 displays the attribute profile CCRs for all item parameter conditions for the condition with six attributes and a test length of 10 items. The multinomial extension of SHE achieved the highest attribute profile CCRs across all item parameter conditions, with the largest difference between the multinomial extension of SHE and the other methods being observed for the HL and LL item parameter conditions. The HL and LL item parameter conditions also have the biggest performance gap for attribute profile CCRs between the CD-CAT and fixed form methods.

Figure 8. Attribute Profile Correct Classification Rates for 6 Attributes with a Test Length of 10



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods were run for 50 replications, except for Shannon Entropy, which was run for only 10 replications.

The attribute profile CCRs for SHE ranged from 0.349 to 0.783 for the LH and HL item parameter conditions, respectively. The single attribute and composite theoretical CCR methods had similar attribute profile CCRs , with the composite method performing slightly better, by 0.038 on average, than the single attribute method. Attribute profile CCRs for the composite

method ranged from 0.316 to 0.684 for the LH and HL item parameter conditions, respectively. For the single attribute theoretical CCR method, attribute profile CCRs ranged from 0.289 to 0.637 for the LH and HL item parameter conditions, respectively.

Well-designed fixed forms had attribute profile CCRs that were, on average, 0.046 higher than those of random fixed forms. Attribute profile CCRs for the well-designed fixed form method ranged from 0.222 to 0.487 for the LH and HL item parameter conditions, respectively. For the random fixed form method, attribute profile CCRs ranged from 0.189 to 0.425 for the LH and HL item parameter conditions, respectively. For Table 9, which compares fixed forms to the CD-CAT methods, the multinomial extension of Shannon Entropy outperforms random fixed forms by 0.252 and a well-designed fixed form by 0.206, with the largest gap in performance observed for the HL item parameter condition.

Table 9. Attribute Profile Correct Classification Rates for 6 Attributes with a Test Length of 10

Method	**Attribute Profile CCRs**			
	HH	**HL**	**LH**	**LL**
Single Attribute	0.299	0.637	0.289	0.526
Composite	0.329	0.684	0.316	0.572
SHE	0.357	0.783	0.349	0.678
Fixed	0.201	0.425	0.189	0.344
Good Fixed	0.229	0.487	0.222	0.403

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods were run for 50 replications, except for Shannon Entropy, which was run for only 10 replications.

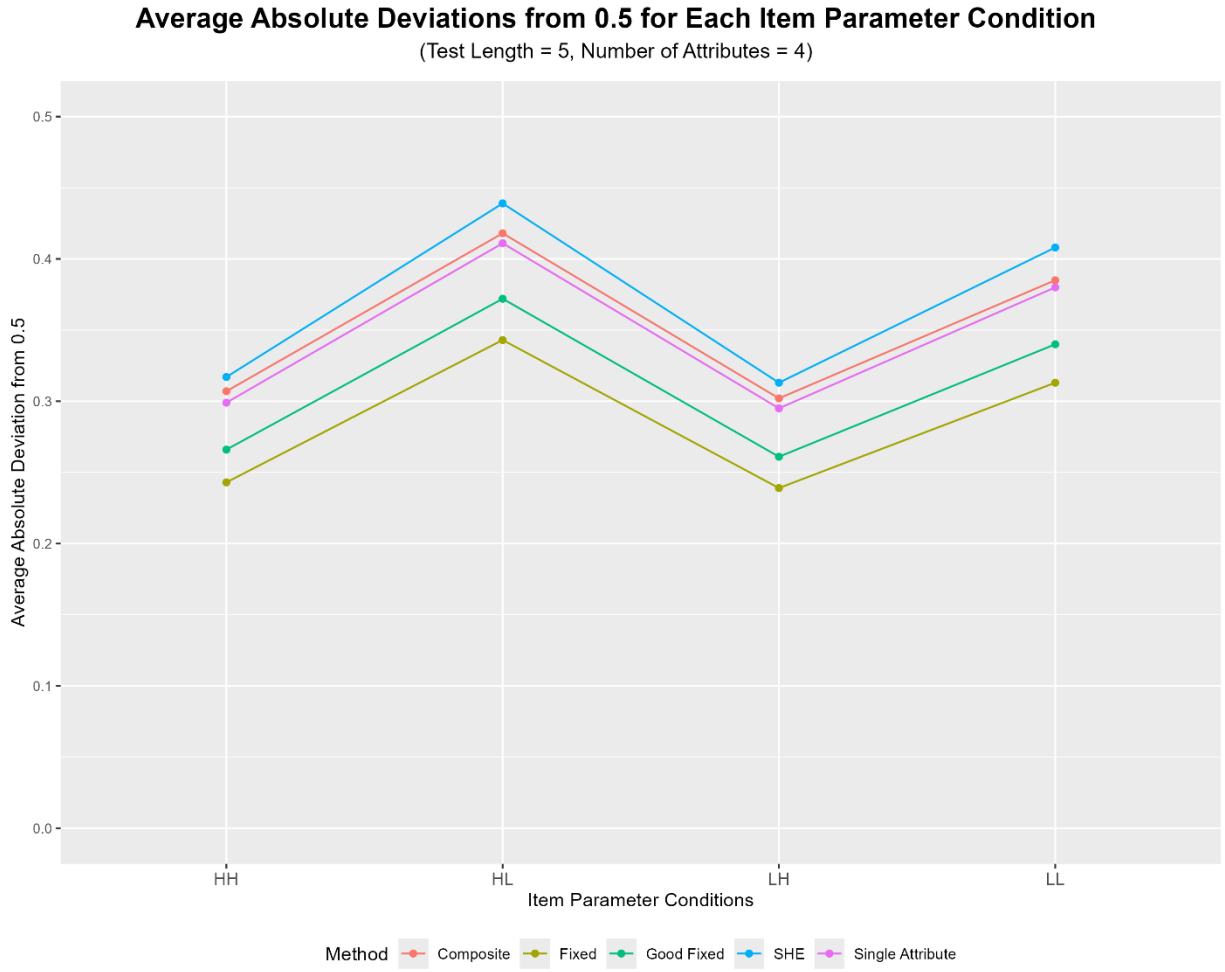
Average Absolute Deviations from 0.5

The average absolute deviations from 0.5 for each of the simulation conditions are presented in Figures 9 to 12. Using the average absolute deviation from 0.5 as an evaluation metric provided an additional way to assess how effectively each CD-CAT item selection method achieved mastery. A posterior probability of mastery near 0.5 conveys little or no information about whether an examinee has mastered the attribute, whereas probabilities farther from 0.5 indicate greater certainty about mastery status.

4 Attributes and 5 Items

Figure 9 displays the average absolute deviation from 0.5 for all item parameter conditions, specifically for the condition with four attributes and a test length of five items. Across all item parameter conditions, the multinomial extension of SHE provided the most information about mastery for all attributes, having the highest absolute average deviation from 0.5. The single attribute and composite theoretical CCR methods exhibited similar average absolute deviations from 0.5, with the composite method performing slightly better, by 0.007 on average, than the single attribute method across all attributes. Fixed form performance for the condition with four attributes and a test length of five items differed slightly between random fixed forms and well-designed fixed forms.

Figure 9. Average Absolute Deviations from 0.5 for 4 Attributes with a Test Length of 5



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

The average absolute deviations from 0.5, displayed in Table 10, across all attributes for the multinomial extension of SHE ranged from 0.313 to 0.439 for the LH and HL item parameter conditions, respectively. The average absolute deviations from 0.5 for the composite attribute theoretical CCR methods ranged from 0.302 to 0.418 for the LH and HL item parameter conditions, respectively. For the single attribute theoretical CCR method, average absolute

deviations from 0.5 ranged from 0.295 to 0.411 for the LH and HL item parameter conditions, respectively. Well-designed fixed forms had average absolute deviations from 0.5 that were, on average, 0.026 higher than those of random fixed forms. The average absolute deviations from 0.5 for the well-designed fixed form method ranged from 0.261 to 0.372 for the LH and HL item parameter conditions, respectively. For the random fixed form method, the average absolute deviations from 0.5 ranged from 0.239 to 0.372 for the LH and HL item parameter conditions, respectively. In Table 9, comparing fixed forms to the CD-CAT methods, it can be seen that the multinomial extension of Shannon Entropy performs 0.085 better than random fixed forms and 0.060, better than well-designed fixed forms, on average, with the largest gap in performance observed for the HL and LL item parameter conditions.

Table 10. Average Absolute Deviations from 0.5 for 4 Attributes with a Test Length of 5

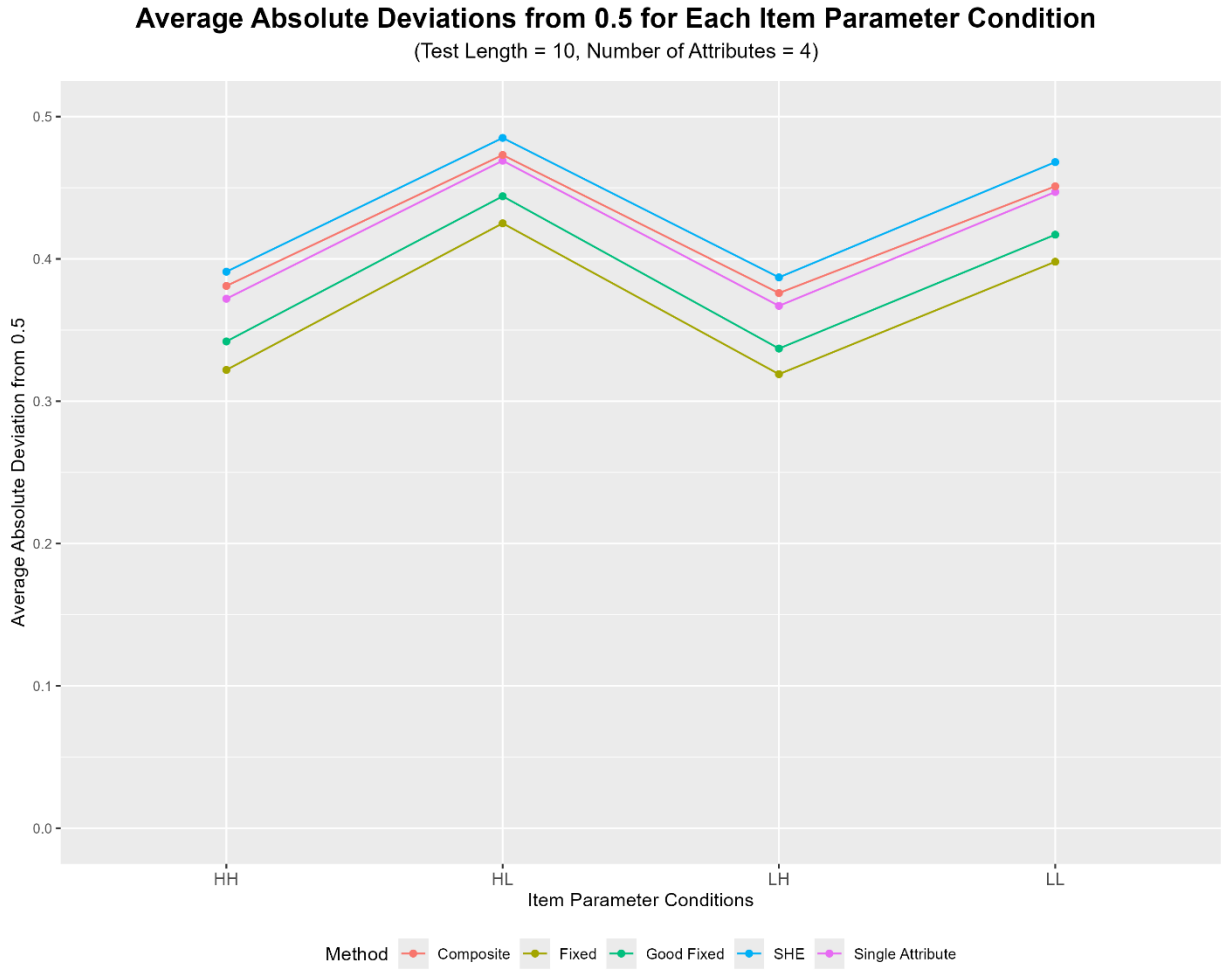
Method	**Average Absolute Deviations from 0.5 for all Attributes**			
	HH	**HL**	**LH**	**LL**
Single Attribute	0.299	0.411	0.295	0.380
Composite	0.307	0.418	0.302	0.385
SHE	0.317	0.439	0.313	0.408
Fixed	0.243	0.343	0.239	0.313
Good Fixed	0.266	0.372	0.261	0.340

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

4 Attributes and 10 Items

Figure 10 displays the average absolute deviations from 0.5 for all item parameter conditions, with a test length of 10 items and four attributes. Across all item parameter conditions, the multinomial extension of SHE provided the most information about mastery for all attributes, having the highest absolute average deviation from 0.5. The single attribute and composite theoretical CCR methods had similar average absolute deviations from 0.5, with the composite method performing slightly better, by 0.007 on average, than the single attribute method for attributes. Fixed form performance for the condition with four attributes and a test length of 10 items differed slightly between random fixed forms and well-designed fixed forms.

Figure 10. Average Absolute Deviations from 0.5 for 4 Attributes with a Test Length of 10



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

Average absolute deviations from 0.5, displayed in Table 11, for the multinomial extension of SHE ranged from 0.387, for LH, to 0.485, for the HL item parameter condition. The average absolute deviations from 0.5 for the composite attribute theoretical CCR methods ranged from 0.376 to 0.473 for the LH and HL item parameter conditions, respectively. For the single attribute theoretical CCR method, the average absolute deviations from 0.5 ranged from 0.367 to

0.469. Well-designed fixed forms had average absolute deviations from 0.5,which were 0.019 higher than those of random fixed forms. For the well-designed fixed-form method, the average absolute deviations from 0.5 ranged from 0.337 to 0.444. For the random fixed form method, individual average absolute deviations from 0.5 ranged from 0.319 to 0.425 for the LH and HL item parameter conditions, respectively. For Table 11, comparing fixed forms to the CD-CAT methods, the multinomial extension of Shannon Entropy performs 0.067 better than random fixed forms and 0.048 better than a well-designed fixed form, on average, with the biggest gap in performance for the LL item parameter condition.

Table 11. Average Absolute Deviations from 0.5 for 4 Attributes with a Test Length of 10

Method	Average Absolute Deviations from 0.5 for all Attributes			
	HH	HL	LH	LL
Single Attribute	0.372	0.469	0.367	0.447
Composite	0.381	0.473	0.376	0.451
SHE	0.391	0.485	0.387	0.468
Fixed	0.322	0.425	0.319	0.398
Good Fixed	0.342	0.444	0.337	0.417

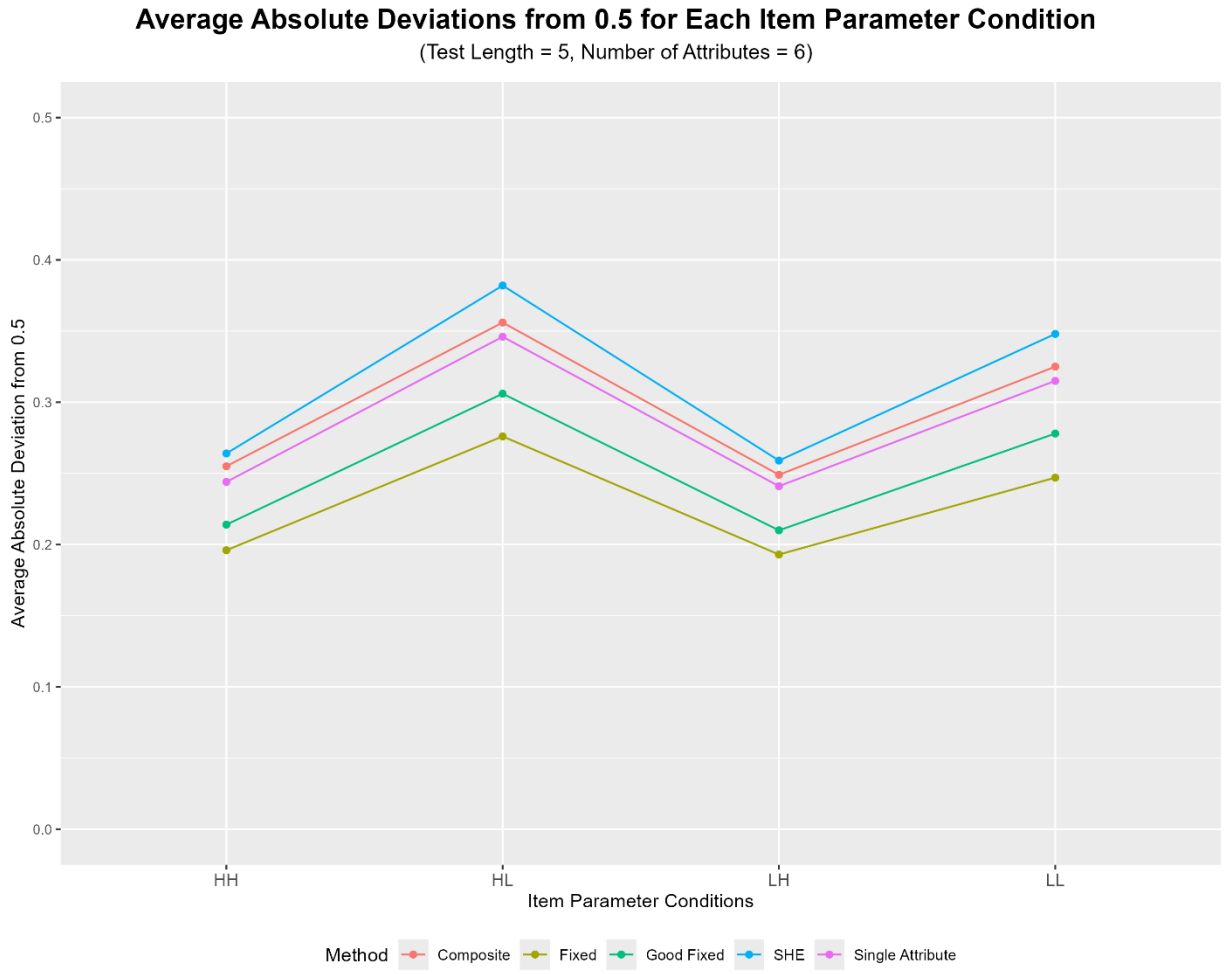
Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

6 Attributes and 5 Items

Figure 11 displays the average absolute deviation from 0.5 for all item parameter conditions, with a test length of five items and six attributes. Across all item parameter conditions, the multinomial extension of SHE provided the most information about mastery for

all attributes by having the highest absolute average deviation from 0.5. The single attribute and composite theoretical CCR methods had similar average absolute deviations from 0.5, with the composite method performing slightly better, 0.010 on average, than the single attribute method for individual attributes. Fixed form performance for the condition with six attributes and five items differed slightly between random fixed forms and well-designed fixed forms.

Figure 11. Average Absolute Deviations from 0.5 for 6 Attributes with a Test Length of 5



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

In Table 12, the average absolute deviations from 0.5 for the multinomial extension of SHE ranged from 0.259 to 0.382 for the LH and HL item parameter conditions, respectively. The average absolute deviations from 0.5 for the composite attribute theoretical CCR method ranged from 0.249 to 0.356 for the LH and HL item parameter conditions, respectively. For the single attribute theoretical CCR method, individual average absolute deviations from 0.5 ranged from 0.241 to 0.346 for the LH and HL item parameter conditions, respectively. Well-designed fixed forms had average absolute deviations from 0.5, 0.024 higher than random fixed forms. The average absolute deviations from 0.5 for the well-designed fixed form method ranged from 0.210 to 0.306. For the random fixed form method, individual average absolute deviations from 0.5 ranged from 0.193 to 0.276. For Table 12, comparing fixed forms to the CD-CAT methods, the multinomial extension of Shannon Entropy performs 0.085 better than random fixed forms and 0.061 better than a well-designed fixed form, on average, with the biggest gap in performance for the HL item parameter condition.

Table 12. Average Absolute Deviations from 0.5 for 6 Attributes with a Test Length of 5

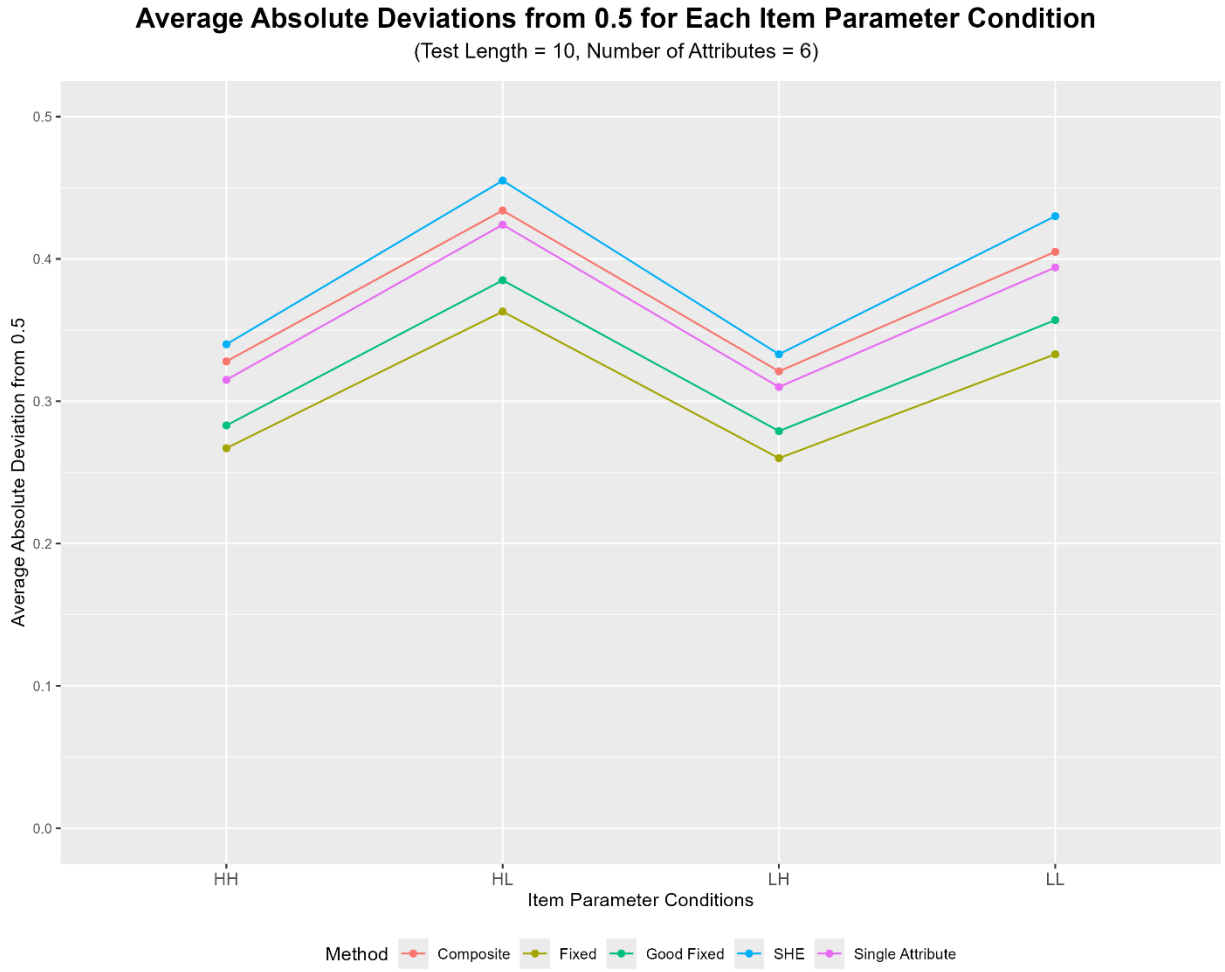
Method	**Average Absolute Deviations from 0.5 for all Attributes**			
	HH	**HL**	**LH**	**LL**
Single Attribute	0.244	0.346	0.241	0.315
Composite	0.255	0.356	0.249	0.325
SHE	0.264	0.382	0.259	0.348
Fixed	0.196	0.276	0.193	0.247
Good Fixed	0.214	0.306	0.210	0.278

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

6 Attributes and 10 Items

Figure 12 displays the average absolute deviation from 0.5 for all item parameter conditions for the condition with six attributes and a test length of 10 items. Across all item parameter conditions, the multinomial extension of SHE provided the most information about mastery for all attributes by having the highest absolute average deviation from 0.5. The single attribute and composite theoretical CCR methods had similar average absolute deviations from 0.5, with the composite method performing slightly better, by 0.011 on average, than the single attribute method for attributes. Fixed form performance for the condition with six attributes and a test length of 10 items differed slightly between random fixed forms and well-designed fixed forms.

Figure 12. Average Absolute Deviations from 0.5 for 6 Attributes with a Test Length of 10



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods were run for 50 replications, except for Shannon Entropy, which was run for only 10 replications.

The average absolute deviations from 0.5, as displayed in Table 13, for the multinomial extension of SHE ranged from 0.333 to 0.455 for the LH and HL item parameter conditions, respectively. The average absolute deviations from 0.5 for the composite attribute theoretical CCR methods ranged from 0.321 to 0.434. For the single attribute theoretical CCR method, average absolute deviations from 0.5 ranged from 0.310 to 0.424 for the LH and HL item

parameter conditions, respectively. Well-designed fixed forms had average absolute deviations from 0.5 that were 0.020 higher than those of random fixed forms. The average absolute deviations from 0.5 for the well-designed fixed form method ranged from 0.279 to 0.385 for the LH and HL item parameter conditions, respectively. For the random fixed form method, average absolute deviations from 0.5 ranged from 0.260 to 0.363 for the LH and HL item parameter conditions, respectively. For Table 13, comparing fixed forms to the CD-CAT methods, the multinomial extension of Shannon Entropy performs 0.084 better than random fixed forms and 0.064 better than well-designed fixed forms, on average, with the largest performance gap observed in the LL item parameter condition.

Table 13. Average Absolute Deviations from 0.5 for 6 Attributes with a Test Length of 10

Method	Average Absolute Deviations from 0.5 for all Attributes			
	HH	HL	LH	LL
Single Attribute	0.315	0.424	0.310	0.394
Composite	0.328	0.434	0.321	0.405
SHE	0.340	0.455	0.333	0.430
Fixed	0.267	0.363	0.260	0.333
Good Fixed	0.283	0.385	0.279	0.357

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods were run for 50 replications, except for Shannon Entropy, which was run for only 10 replications.

Average Item Exposure Rate

Figures 13 to 16 display the average item exposure rates for each CD-CAT method. Item exposure was calculated at five intervals: 0%, between 0% and 25%, between 25% and 50%,

between 50% and 90%, and between 90% and 100%. The first items for both theoretical CCR methods were selected by identifying the item with the largest theoretical CCR for one of the measured attributes. For the multinomial extension of SHE, the first item chosen was the item that minimizes the expected Shannon Entropy when the posterior probability of each attribute pattern is equally likely. As every examinee received the same first item for each selection method, there is at least one item that was seen by 100% of examinees for each study condition.

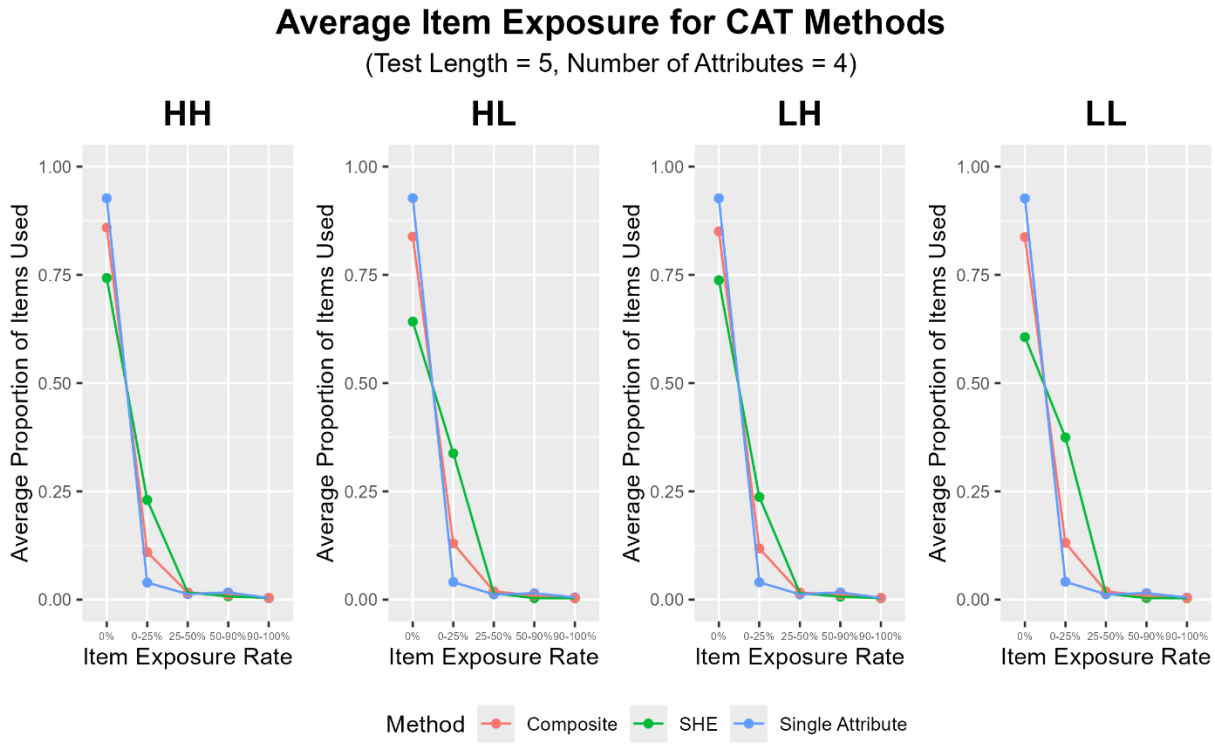
To find the average proportion used for each of the intervals, first, the counts of how many examinees received each unique item were calculated for each replication. Next, within each replication, the proportion of items used within each interval was counted. For 0%, the number of items used for each replication was subtracted from 300, the total number of items in the bank. For the 0% to 25% range, the proportion of items administered to between 1 and 249 examinees was calculated. To calculate between 25% and 50%, the proportion of items administered to 250 to 499 examinees was used. For 50% to 90%, the proportion of items administered to between 500 and 899 examinees was calculated. Lastly, to calculate between 90% and 100%, the proportion of items administered to 900 to 1000 examinees was used. After these calculations were completed at the replication level, the results were averaged across replications for each condition.

4 Attributes and 5 Items

Figure 13 displays the average item exposure rates for the conditions of four attributes and five items. Across all item parameter conditions, the single attribute theoretical CCR method used the least number of items from the item bank, with an average proportion of 0.927 of items being used 0% of the time (i.e., about 92.7% of the items were never used). The composite theoretical CCR method used more items from the bank on average but still had between 0.837

and 0.859 of items used 0% of the time across the four item parameter conditions, with the LL condition using the highest proportion of items and the HH condition using the least. Using the largest number of items was the multinomial extension of SHE, but it still had between 0.606 and 0.743 of items that were never used across conditions.

Figure 13. Average Item Exposure Rate for 4 Attributes with a Test Length of 5



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

The majority of items from the item banks used for each of the CD-CAT methods were used by at most 25% of the examinees for the conditions with four attributes and five items, as shown in Table 14. Among the different item parameter conditions, the largest proportion of items were used under the LL item parameter condition. The multinomial extension of SHE

method used the most items from the bank, averaging from 0.230, for the HH condition, to 0.375 for the LL condition. In the middle, the composite theoretical CCR method used between 0.110 and 0.131 of items in the bank. With the fewest number of items used, the single attribute theoretical CCR method utilized only between 0.039 and 0.041 items from the bank.

Very few items were used by more than 25% of the examinees, as displayed in Table 14. For an item exposure rate between 25% and 50%, all three CD-CAT methods had similar percentages of examinees who saw items. The single attribute theoretical CCR method averaged between 0.012 for the LL, LH, and HL conditions, to 0.013 of items used for the HH condition. For the composite theoretical CCR method, the proportion of items used averaged between 0.016 for the HH and LH conditions to 0.019 for the HL and LL item parameter conditions. Lastly, the multinomial extension of SHE had an average proportion of items used between 0.014 for the LL condition and 0.017 for the HH item parameter condition. When items were exposed to 50% to 90% of examinees, the single attribute theoretical CCR method has the highest proportion of items used, with between 0.015 for the HL and LL conditions and 0.017 for the HH and LL item parameter conditions. For the composite theoretical CCR method, the proportion of items used was between 0.009 for the HL and LL conditions and 0.011 for the LH item parameter condition. The multinomial extension of SHE has a proportion of items used ranging from 0.003 for the HL condition to 0.008 for the HH item parameter condition. Less than 0.010 of items were exposed to between 90% and 100% of examinees in each of the item parameter conditions and each CD-CAT method.

Table 14. Average Item Exposure Rate for 4 Attributes with a Test Length of 5

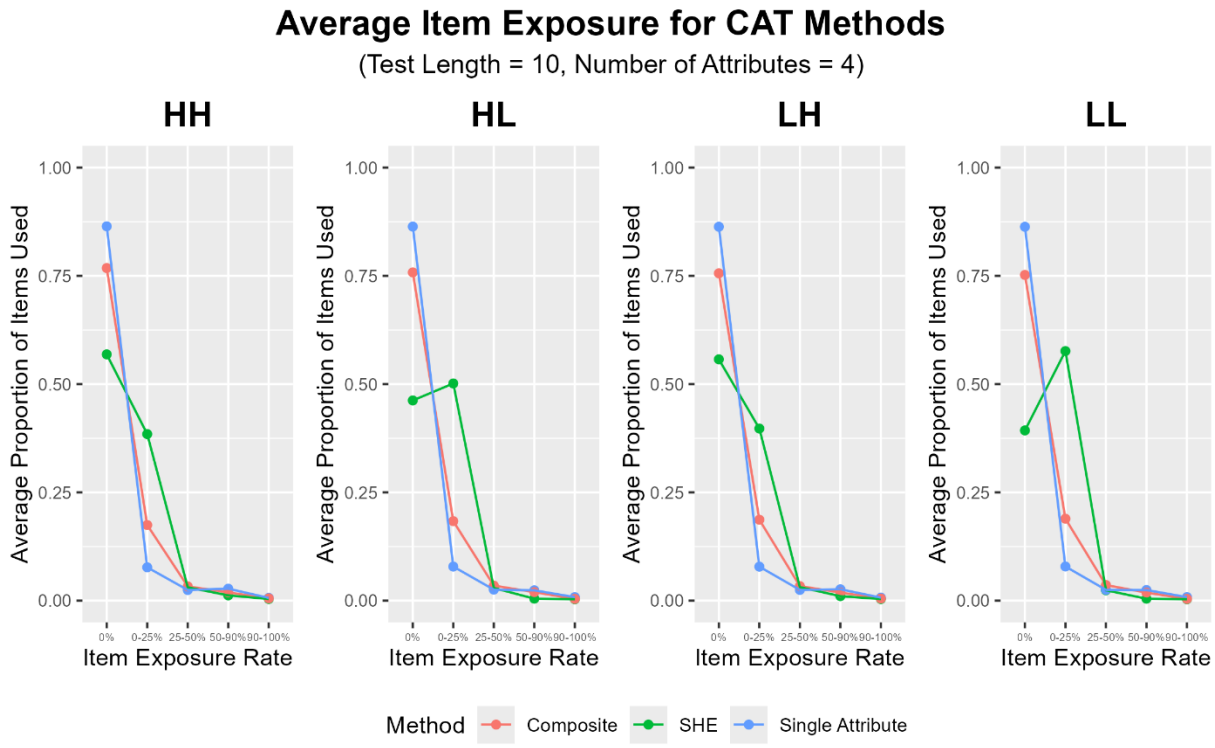
Method	Average Proportions of Items Used				
	0%	**0-25%**	**25-50%**	**50-90%**	**90-100%**
HH					
Single Attribute	0.927	0.039	0.013	0.017	0.004
Composite	0.859	0.110	0.016	0.010	0.004
SHE	0.743	0.230	0.017	0.008	0.004
HL					
Single Attribute	0.927	0.041	0.012	0.015	0.005
Composite	0.838	0.130	0.019	0.009	0.004
SHE	0.642	0.338	0.015	0.003	0.003
LH					
Single Attribute	0.927	0.040	0.012	0.017	0.004
Composite	0.851	0.118	0.016	0.011	0.004
SHE	0.738	0.237	0.015	0.007	0.004
LL					
Single Attribute	0.926	0.041	0.012	0.015	0.005
Composite	0.837	0.131	0.019	0.009	0.004
SHE	0.606	0.375	0.014	0.004	0.003

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

4 Attributes and 10 Items

Figure 14 displays the average item exposure rates for the condition with four attributes and 10 items. Given that the examinees received more items, fewer items were unused from the item bank. Across all item parameter conditions, the single attribute theoretical CCR method used the fewest number of items from the item bank, with an average proportion of 0.860 of items being used 0% of the time (i.e., about 86.0% of the items were never used), across all item parameter conditions. The composite theoretical CCR method used more items from the bank on average, but still had a proportion between 0.752 and 0.768 of items used 0% of the time across the four item parameter conditions. The LL condition used the most items, and the HH condition used the least. Using the largest number of items was the multinomial extension of SHE, with only a proportion of between 0.393 and 0.569 of items not used for the LL and HH conditions, respectively.

Figure 14. Average Item Exposure Rate for 4 Attributes with a Test Length of 10



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

For the condition with four attributes and 10 items, the majority of items from the item banks used for each of the CD-CAT methods were used by at most 25% of the examinees, as shown in Table 15. Amongst the different item parameter conditions, most items were used under the LL item parameter condition, while the HH item parameter condition was used the least. The multinomial extension of the SHE method used the most items from the bank, averaging a proportion between 0.385 for the HH condition and 0.576 for the LL condition. Having item usage in the middle for the three CD-CAT methods, the composite theoretical CCR method used between 0.175 and 0.189 of items in the bank, for the HH and LL item parameter

conditions, respectively. With the lowest number of items used, the single attribute theoretical CCR method averaged a proportion of only 0.078 of the items used from the bank under 25% of the time.

Fewer than 0.060 of the items were used by more than 25% of the examinees, as displayed in Table 15 for each of the CD-CAT methods. For an item exposure rate between 25% and 50%, all three CD-CAT methods had similar percentages of examinees who saw items. The single attribute theoretical CCR method had an average proportion of 0.078 across all item parameter conditions. For the composite theoretical CCR method, average proportions of items used ranged between 0.033 for the HH condition and 0.036 for the LL item parameter condition. Lastly, the multinomial extension of SHE had an average proportion of items used between 0.024 for the LL condition and 0.031 for the HH and LH item parameter conditions. When items were exposed to 50% to 90% of examinees, the single attribute theoretical CCR method had the highest item usage rate, averaging about 0.026 across all item parameter conditions. For the composite theoretical CCR method, the proportion of items used was about 0.019 across all item parameter conditions. The multinomial extension of SHE had the proportion of items used between 0.004 for the LL and HL conditions and 0.010 for the HH and LH item parameter conditions. Less than 0.010 of items were exposed to between 90% and 100% of examinees in each of the item parameter conditions, as shown in Table 15.

Table 15. Average Item Exposure Rate for 4 Attributes with a Test Length of 10

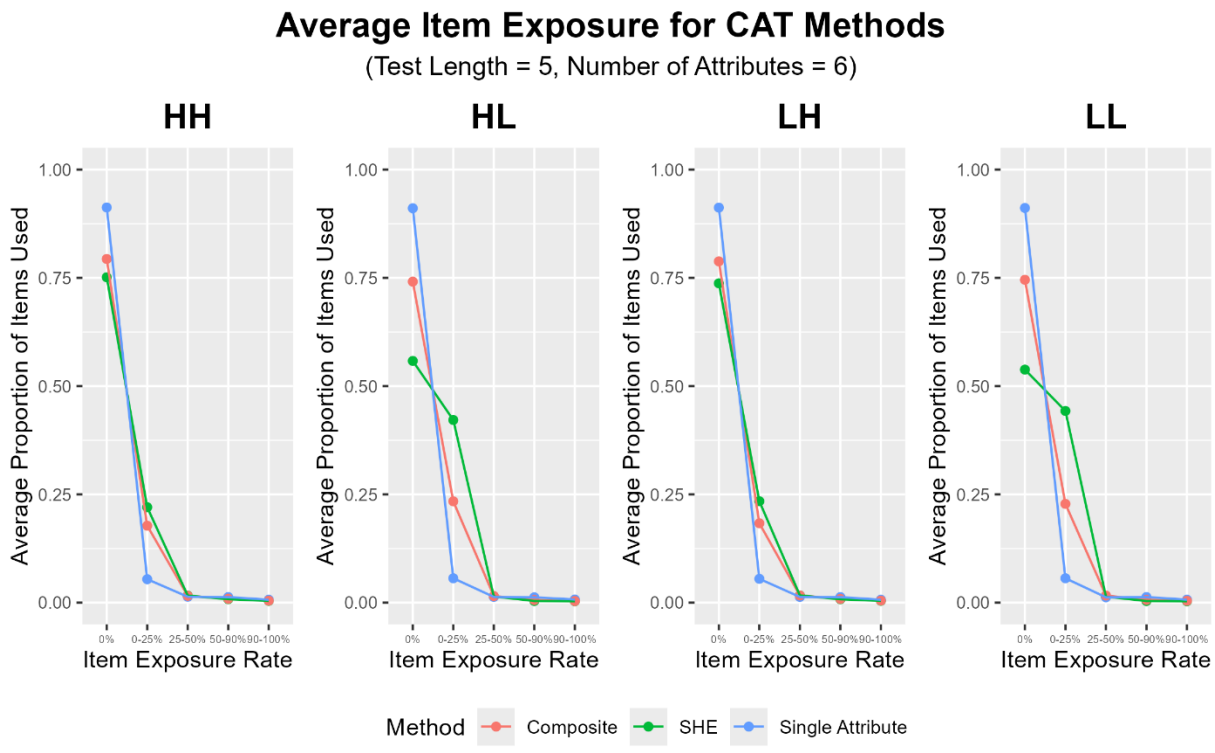
Method	Average Proportions of Items Used				
	0%	**0-25%**	**25-50%**	**50-90%**	**90-100%**
HH					
Single Attribute	0.864	0.077	0.025	0.028	0.006
Composite	0.768	0.175	0.033	0.019	0.005
SHE	0.569	0.385	0.031	0.012	0.004
HL					
Single Attribute	0.864	0.079	0.025	0.024	0.008
Composite	0.758	0.184	0.035	0.020	0.004
SHE	0.462	0.502	0.029	0.005	0.003
LH					
Single Attribute	0.863	0.078	0.025	0.026	0.007
Composite	0.756	0.187	0.034	0.019	0.005
SHE	0.557	0.397	0.031	0.010	0.004
LL					
Single Attribute	0.863	0.079	0.025	0.025	0.008
Composite	0.752	0.189	0.036	0.019	0.004
SHE	0.393	0.576	0.024	0.004	0.003

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

6 Attributes and 5 Items

Figure 15 displays the average item exposure rates for the condition with six attributes and five items. Across all item parameter conditions, the single attribute theoretical CCR method used the least number of items from the item bank, averaging a proportion of 0.912 of items never being used. For the HH and LH item parameter conditions, the composite theoretical CCR method and the multinomial extension of SHE yielded similar proportions of items not used from the item bank, at 0.794 and 0.751, respectively, for the HH item parameter condition, and 0.788 and 0.737, respectively, for the LH item parameter condition. The multinomial extension of SHE utilized more items from the bank for the HL and LL item parameter conditions, with only 0.558 and 0.538 of the items used from the bank, respectively.

Figure 15. Average Item Exposure Rate for 6 Attributes with a Test Length of 5



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

The majority of items from the item banks used for each of the CD-CAT methods were used by at most 25% of the examinees for the conditions with six attributes and five items, as shown in Table 16. Amongst the different item parameter conditions, most items were used under the LL and HL item parameter conditions, while the HH item parameter condition was used the least. The multinomial extension of the SHE method used the most items from the bank, averaging a proportion between 0.220, for the HH condition, and 0.443 for the LL condition. The next method with the largest item usage was the composite theoretical CCR method, which averaged a proportion between 0.178 and 0.234 of items used from the bank, for the HH and HL item parameters conditions, respectively. With the lowest number of items used, the single attribute theoretical CCR method utilized an average proportion of 0.056 of the items from the bank, under 25% of the time across all item parameter conditions.

Fewer than 0.040 of the items were used by more than 25% of the examinees, as displayed in Table 16 for each of the CD-CAT methods. For an item exposure rate between 25% and 50%, all three CD-CAT methods yielded similar proportions, with less than 0.020 of examinees seeing items. The single attribute theoretical CCR method had an average proportion of about 0.013 of items used across all item parameter conditions. For the composite theoretical CCR method, items usage averaged 0.016 for all item parameter conditions. Lastly, the multinomial extension of SHE had an average proportion of items used of 0.015 across the item parameter conditions. When items were exposed to 50% to 90% of examinees, the single attribute theoretical CCR method had the highest item usage rate, averaging a proportion of

0.013 across all item parameter conditions. For the composite theoretical CCR method, the item usage averaged a proportion of 0.007 for all item parameter conditions. The multinomial extension of SHE had a proportion of items used of 0.004 for the HL and LL conditions, and 0.008 for the HH and LH item parameter conditions. Less than 0.010 of items were exposed to between 90% and 100% of examinees in each of the item parameter conditions, as shown in Table 16.

Table 16. Average Item Exposure Rate for 6 Attributes with a Test Length of 5

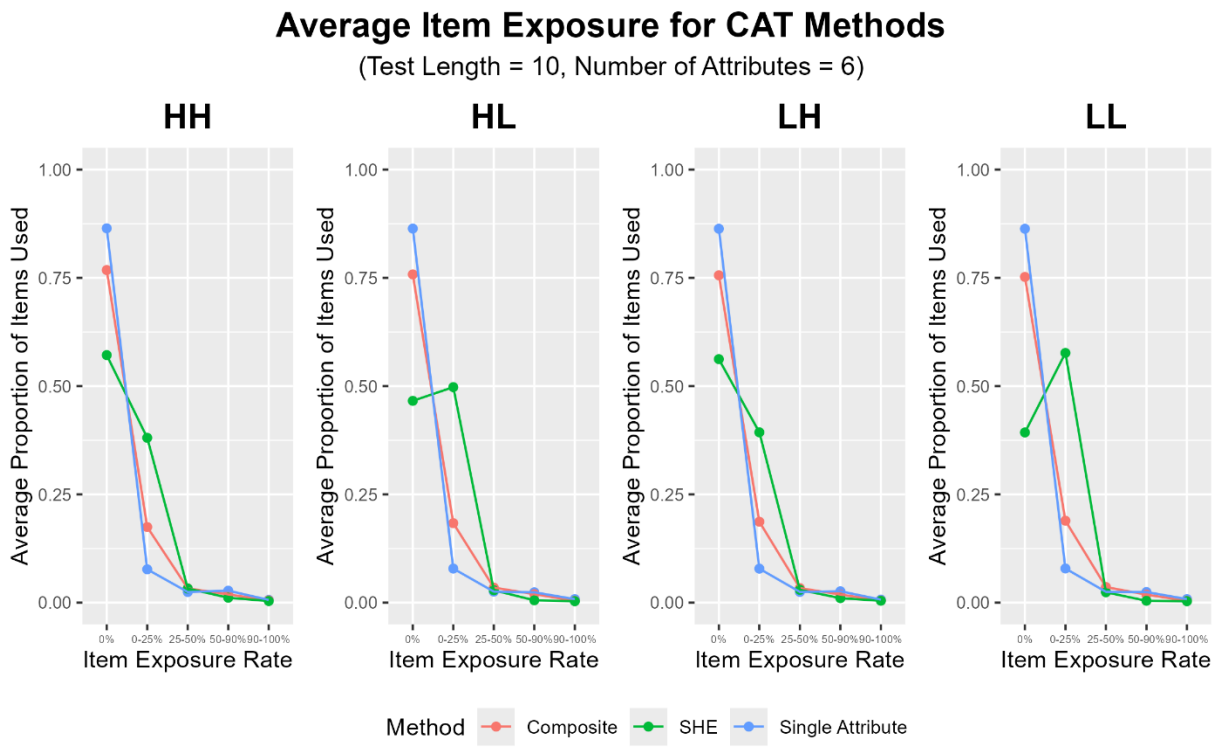
Method	Average Proportions of Items Used				
	0%	0-25%	25-50%	50-90%	90-100%
HH					
Single Attribute	0.913	0.054	0.014	0.013	0.007
Composite	0.794	0.178	0.016	0.009	0.005
SHE	0.751	0.220	0.016	0.008	0.004
HL					
Single Attribute	0.911	0.056	0.013	0.012	0.007
Composite	0.741	0.234	0.015	0.006	0.004
SHE	0.558	0.422	0.014	0.004	0.003
LH					
Single Attribute	0.912	0.055	0.013	0.013	0.007
Composite	0.788	0.183	0.017	0.009	0.005
SHE	0.737	0.234	0.017	0.008	0.004
LL					
Single Attribute	0.911	0.056	0.012	0.013	0.007
Composite	0.746	0.228	0.016	0.006	0.004
SHE	0.538	0.443	0.014	0.004	0.003

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods ran for 50 replications.

6 Attributes and 10 Items

Figure 16 displays the average item exposure rates for the condition of six attributes and ten items. Across all item parameter conditions, the single attribute theoretical CCR method used the least number of items from the item bank, averaging a proportion of 0.0835 of items never being used. The composite theoretical CCR method used more items from the bank on average; however, the proportion of items never used ranged between 0.620 and 0.665 across the four item parameter conditions, with the LL condition using the most items and the HH condition using the fewest. Using the largest number of items was the multinomial extension of SHE, with a proportion greater than 0.700 for the LL condition and approximately 0.460 for the HH item parameter condition.

Figure 16. Average Item Exposure Rate for 6 Attributes with a Test Length of 10



Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods were run for 50 replications, except for Shannon Entropy, which was run for only 10 replications.

The majority of items from the item banks used for each of the CD-CAT methods were used by at most 25% of the examinees, as shown in Table 17, for the condition with six attributes and 10 items. Among the different item parameter conditions, most items were used under the LL item parameter conditions, while the HH and LH item parameter conditions were used the least. The multinomial extension of the SHE method used the most items from the bank, averaging a proportion between 0.409 for the HH condition and 0.648 for the LL condition. The next method with the largest item usage is the composite theoretical CCR method, which had a proportion of items used between 0.284 and 0.329 in the bank, for the HH and LL item parameter conditions, respectively. With the lowest number of items used under 25% of the time across all item parameter conditions, the single attribute theoretical CCR method averaged a proportion of 0.110.

Fewer than 0.060 items were used by more than 25% of the examinees, as displayed in Table 17 for each of the CD-CAT methods. For an item exposure rate between 25% and 50%, all three CD-CAT methods yielded similar percentages, with fewer than 0.040 of examinees seeing items. The single attribute theoretical CCR method averaged a proportion of 0.024 of items used across all item parameter conditions. For the composite theoretical CCR method, item usage averaged a proportion between 0.031 for the LH and HH conditions and 0.036 for HL and LL item parameter conditions. Lastly, the multinomial extension of SHE had an average proportion of item used between 0.023 and 0.030 across the item parameter conditions. When items were exposed to 50% to 90% of examinees, the single attribute theoretical CCR method had the

highest item usage rate, averaging between 0.023 and 0.027 across all item parameter conditions. For the composite theoretical CCR method, the item usage averaged between 0.011 and 0.015 for all item parameter conditions. The multinomial extension of SHE used between 0.004 and 0.013 for the HL condition and the HH item parameter condition, respectively. Fewer than 0.010 of items were exposed to between 90% and 100% of examinees in each of the item parameter conditions, and each CD-CAT method, as displayed in Table 17.

Table 17. Average Item Exposure Rate for 6 Attributes with a Test Length of 10

Method	Average Proportions of Items Used				
	0%	0-25%	25-50%	50-90%	90-100%
HH					
Single Attribute	0.836	0.106	0.024	0.027	0.008
Composite	0.665	0.284	0.031	0.015	0.005
SHE	0.544	0.409	0.030	0.013	0.005
HL					
Single Attribute	0.831	0.112	0.025	0.023	0.008
Composite	0.622	0.327	0.036	0.011	0.004
SHE	0.329	0.640	0.024	0.004	0.003
LH					
Single Attribute	0.837	0.105	0.024	0.026	0.007
Composite	0.654	0.296	0.031	0.015	0.005
SHE	0.528	0.428	0.029	0.010	0.005
LL					
Single Attribute	0.832	0.111	0.025	0.024	0.008
Composite	0.620	0.329	0.036	0.011	0.004
SHE	0.286	0.684	0.023	0.005	0.003

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods were run for 50 replications, except for Shannon Entropy, which was run for only 10 replications.

Average Time

When conducting a simulation study to compare the efficiency of each method, the time it takes to run each method is crucial in determining the computational load. For each CD-CAT method, the average number of days per replication was calculated. For the single attribute and composite attribute theoretical CCR methods, the time was reported in minutes but was converted to days in Table 18 for comparison with the multinomial extension on SHE. Across all conditions, the multinomial extension of SHE was the slowest and most computationally intense method, averaging 1.613 days for the condition with four attributes and a test length of five items to 22.862 days for the condition with six attributes and 10 items. The single attribute and composite attribute theoretical CCR methods had similar average times across all conditions. For the single attribute theoretical CCR method, the average times ranged from 0.002 days for the condition with attributes and five items to 0.024 days for the conditions with six attributes and 10 items. For the composite attribute theoretical CCR method, the average times ranged from 0.002 days for the condition with 4 attributes and five items to 0.025 days for the condition with six attributes and 10 items.

Table 18. Average Time Per Replication

Method	Average Time Per Replication (In Days)			
	HH	**HL**	**LH**	**LL**
4 Attributes, 5 Items				
Single Attribute	0.002	0.002	0.002	0.002
Composite	0.002	0.002	0.002	0.002
SHE	1.615	1.611	1.612	1.615
4 Attributes, 10 Items				
Single Attribute	0.006	0.006	0.006	0.005
Composite	0.006	0.006	0.006	0.005
SHE	4.032	4.055	4.069	4.071
6 Attributes, 5 Items				
Single Attribute	0.009	0.009	0.009	0.009
Composite	0.011	0.011	0.010	0.010
SHE	6.262	6.197	5.943	5.732
6 Attributes, 10 Items				
Single Attribute	0.025	0.025	0.021	0.023
Composite	0.025	0.025	0.025	0.024
SHE	22.632	23.403	25.131	20.282

Note. HH: $\pi \sim U(0.75,0.95), r \sim (0.3,0.5)$; HL: $\pi \sim U(0.75,0.95), r \sim (0.1,0.3)$; LH: $\pi \sim U(0.5,0.75), r \sim (0.3,0.5)$; LL: $\pi \sim U(0.5,0.75), r \sim (0.1,0.3)$. All methods were run for 50 replications, except for Shannon Entropy, which was run for only 10 replications for the condition with 6 attributes and 10 items.

CHAPTER V: DISCUSSION

This dissertation study examined the performance of various CD-CAT methods for item selection under the GDCM ERUM-MC model, a multinomial diagnostic classification model. Specifically, this study looked to answer three questions. First, the performance of the multinomial extension of the expected Shannon Entropy procedure was evaluated. Next, the performance of two newly proposed CD-CAT methods, based on theoretical correct classification rates, was compared to the multinomial extension of expected Shannon Entropy. Lastly, the performance of the three CD-CAT methods was compared to a randomly designed test and, most importantly, to a well-designed fixed form test, to introduce realistic conditions that have not been previously studied, under the GDCM ERUM-MC model.

Key Findings and Conclusion

Consistent with previous studies (Cheng, 2009; Wang, 2013; Xu et al., 2003; Xu et al., 2016; Yigit et al., 2019), a key finding was that each of the CD-CAT methods performed similarly, with no more than a 10% improvement when compared to a randomly constructed fixed form test. Because each of those studies employed a fixed-form design based on random selection, which is not typically used to build tests in practice, this study aimed to investigate a more realistic approach using a well-designed fixed form. The goal of this study was to introduce three new CD-CAT item selection methods and determine whether there is a difference in performance between them and a well-designed fixed form test constructed from attributes with the highest theoretical CCRs. The multinomial extension of expected Shannon Entropy had higher CCRs than the single attribute and composite theoretical CCR methods, as well as both fixed form methods. In addition, in this simulation study, it was shown that a well-designed fixed form had CCRs that were only 0.058 less than the multinomial extension of SHE and even a

smaller difference for the single attribute and composite attribute theoretical CCR methods, This results demonstrates that computer adaptive testing might not always be the most effective way to administer a test for diagnostic classification modeling. If the items are strategically picked, the fixed form performs almost as well as a CAT test.

The longer the test, the higher the CCRs became. CCRs are expected to increase because a longer test has more items to provide information about whether an examinee is a master or not of a particular attribute. Trying to correctly classify six attributes with only five items, where each item only measures two skills and one misconception, can be difficult, so the more attributes on a test, the higher the test length should be to increase the rate of correctly classifying examinees into masters and nonmasters.

It was also shown that the CCRs and average absolute deviations from 0.5 for the four attribute conditions were greater than those for the six attribute conditions. Additionally, CCRs and average absolute deviations from 0.5 were greater for conditions with a test length of 10 compared to those with a test length of 5. Across all attribute and test length conditions, the HL item parameter conditions correctly classified examinees most often. Because the HL item parameter condition is when items are most informative and the options have a high discrimination, it is to be expected that it was also the condition with the highest rate of correctly classifying examinees into masters and nonmasters. The condition with the lowest CCRs was the LH condition, where the items are the least informative and the options have a weak discrimination. Additionally, the HH condition, where items are most informative but options have weak discrimination, had CCRs and average absolute deviations from 0.5 that were close to those of the LH condition, but slightly higher.

When test governing bodies consider administering a computer-adaptive test as an alternative to a fixed form test, the primary concern, or reason for a possible switch, is item exposure and ensuring that examinees do not receive the same items. The goal is to ensure that the full item bank is utilized and that examinees receive a personalized test based on what they know and do not know about a topic. For this study, the single attribute and composite theoretical CCR methods did not adequately control for item exposure, as fewer than a quarter of the 300 items were utilized. In contrast, the multinomial extension of the Shannon Entropy procedure performed better with respect to item exposure, using an average of half of the item bank. This led to examinees seeing fewer of the same items for SHE compared to the two theoretical CCR procedures. When few items are used from the bank, examinees see very similar tests, resulting in less personalization of the tests. For both test length conditions with four attributes, the LL item parameter condition, where the items are least informative and options have high discrimination, used the fewest items from the item bank. In contrast, in the HH condition, where items are the most informative, there is an easier time of detecting mastery, and the options have weak discrimination, the most items were used in the bank for both test length conditions, measuring four attributes. For the six-attribute condition with a test length of five items, the HH item parameter condition, used the least number of items from the bank. In contrast, the condition that used the most items from the bank differed between LL, LH, and HL for each of the three CD-CAT methods, SHE, composite attribute, and single attribute, respectively. For the condition with six attributes and a test length of 10, the HH item parameter condition used the least number of items for SHE and the composite attribute methods, while the LH condition used the least number of items for the single attribute method. Lastly, the greatest number of items from the bank were used under the LL condition for the composite attribute and

SHE methods. In contrast, the single attribute theoretical CCR methods used the most items under the HL item parameter conditions when there were six attributes and a test length of 10.

Research Question 1: Multinomial extension of expected SHE

One of the new methods introduced in this study was the multinomial extension of the expected Shannon Entropy procedure. Given that the expected Shannon Entropy procedure has been one of the primary methods used and compared against in the literature for dichotomous DCMs, it was necessary to extend this method to a multinomial model to explore its performance. Overall, the multinomial extension of the Shannon Entropy procedure's performance on multiple criteria being evaluated exceeded both the single and composite attribute theoretical CCR methods and the random and well-designed fixed form methods.

To summarize the performance of the multinomial extension of SHE, the average for each evaluation criterion across all replications within each condition was calculated, followed by the calculation of the minimum, maximum, and average across the averages for the various conditions. First, the multinomial extension of SHE had attribute-level CCRs above 0.757 across all simulation conditions, with the largest attribute-level CCRs reaching 0.985 and an average proportion of 0.876 of examinees being correctly classified as masters or nonmasters across all attributes. Second, the average absolute deviations from 0.5 were above 0.259, with the largest absolute deviation at 0.485, which is nearly the maximum value of 0.5, and with an overall average of about 0.376. Lastly, the multinomial extension of the expected Shannon Entropy procedure had the best item exposure, selecting an average of 0.439 of the item bank for examinees, with proportions ranging from 0.249 to 0.714.

Research Question 2: CD-CAT method comparison

Comparing the multinomial extension of the expected Shannon Entropy procedure to the newly proposed single and composite attribute theoretical CCR methods, the multinomial extension of the SHE outperforms the two theoretical CCR methods, with the biggest gap in performance between the single attribute theoretical CCR method and the multinomial extension of the expected Shannon Entropy procedure.

To summarize the performance of the three CD-CAT methods, the average for each evaluation criterion across all replications within each condition was calculated. Then the minimum, maximum, and average values were determined across the averages for the various conditions within each CD-CAT method. First, the multinomial extension of SHE correctly classified examinees, on average, 0.024 better than the single attribute theoretical CCR method, with the largest difference being 0.036 and the smallest difference being 0.015. The multinomial extension of SHE correctly classified examinees only about 0.015 better than the composite attribute theoretical CCR method, with the largest difference being 0.027 and the smallest being 0.008. Although the improvement from the multinomial extension of SHE to the two theoretical CCR methods appears small, the improvement is more significant in context. With the multinomial extension of SHE's average CCRs at 0.876, values are approaching the maximum possible value of 1, leaving little room for improvement. Thus, even a 0.10 or 0.05 increase reflects a notable improvement in classification accuracy.

Second, in terms of average absolute deviation from 0.5, the multinomial extension of SHE, on average, performed 0.024 better than the single attribute theoretical CCR method, with the smallest difference being 0.016 and the largest being 0.036. Compared to the composite attribute theoretical CCR method, the multinomial extension of SHE had higher average absolute

deviations, on average, by 0.016, with the largest difference being 0.026 and the smallest being 0.009. Lastly, with respect to item exposure, the multinomial extension of SHE used, on average, 0.323 more of the bank than the single attribute method, with the difference ranging from 0.162 to 0.546. In comparison to the composite attribute theoretical CCR method, the multinomial extension of SHE used 0.192 more of the item bank on average, with a range of 0.043 to 0.359.

Despite the multinomial extension of SHE having higher attribute-level CCRs and average absolute deviations from 0.5, it may not be the best choice, as it takes a substantially longer time to run one replication, a difference of days versus minutes. The difference between the multinomial extension of SHE and the two theoretical CCR methods was marginal, so it may not be worth the difference in performance due to the computational demand. When working with real data, organizations would want a method that is both reliable for correctly classifying examinees into masters and nonmasters and time-sensitive, so one of the two theoretical CCR methods may be a better choice.

Research Question 3: Well-designed fixed form vs. CD-CAT

One of the major interests of this study was to determine whether a few CD-CAT methods performed better than a well-designed fixed form text built based on large theoretical CCRs for individual attributes. Prior research (Cheng, 2009; Wang, 2013; Xu et al., 2003, 2016; Yigit et al., 2019) has only used a fixed form based on random item selection to compare with each of the CD-CAT methods, which may not be a realistic comparison. When teachers or K-12 organizations design tests, they may need to consider certain parameters for the form, and items would not typically be chosen at random to measure content. In this sense, a well-designed fixed form could be seen as a more realistic testing condition. The largest difference in performance was between the multinomial extension of SHE and the well-designed fixed form. On average,

the multinomial extension of SHE correctly classified examinees 0.058 better than the well-designed fixed form, while the difference in the average absolute deviation was about 0.058. Second, the composite attribute theoretical CCR method outperformed the well-designed fixed form by 0.043 in correctly classifying examinees into masters and nonmasters, and by approximately 0.042 in terms of the average absolute deviation from 0.5. Lastly, when comparing the single attribute theoretical CCR and well-designed fixed form methods, the single attribute outperformed the well-designed fixed form by about 0.034, on average, in correctly classifying examinees, and for the average absolute deviation from 0.5. Prior research (Xu et al., 2003; Xu et al., 2016; Yigit et al., 2019) focused on interpreting results in terms of attribute profile CCRs and the degree of difference between CAT and random fixed forms, with results aligning with this study. By focusing on the attribute profile CCRs rather than individual attributes, the studies concluded that there were better gains with CAT, and it may be the overall choice of method to use for building a test. Examining the attribute level CCRs across all attributes in this study reveals a marginal difference between CAT and fixed forms, with the difference decreasing when compared to a well-designed fixed form. This suggests that CAT and a well-designed fixed form can yield outcomes that are similar to each other.

Limitations and Future Research

The simulation conditions were selected in a way to best represent conditions that are believed to be realistic. However, there are still potential limitations that should be addressed in future studies. First, the main limitation of this study was not being able to run the full 50 replications for the multinomial extension of the Shannon entropy procedure for the condition with six attributes and 10 items due to the time required to run one replication of the CD-CAT procedure. For example, the condition with a test length of 10 items, one replication took on

average 22.862 days. Due to the nature of the expected SHE procedure, after each examinee receives an item, the expected Shannon Entropy is calculated for every item still remaining in the item bank. These calculations are very computationally intense. In comparison, a typical condition for the single attribute or composite attribute took 0.025 days, or approximately 36 minutes, for the condition with six attributes and 10 items. Also, as the number of attributes being measured increased, the time needed to estimate attribute mastery also increased. If all 50 replications of the multinomial extension of Shannon Entropy were run across all simulation conditions, a more well-informed comparison could be made with other CD-CAT and fixed forms methods, as the evaluation criteria would be averaged across all the same replications. Even though all 50 replications were not run for the condition with six attributes and 10 items, the 10 replications are still a good estimate of performance for the multinomial extension of SHE in comparison to the other methods.

Second, because this was a simulation study, the conditions simulated were previously used in many CD-CAT research studies. These conditions were also employed to attempt to represent real data scenarios. The benefits of using a multinomial model for CD-CAT can be particularly helpful for analyzing short tests, such as formative assessments in the classroom. In the classroom, formative assessments are typically short, consisting of 5 to 10 items, but may only measure a few attributes. These constraints can make it challenging to generalize results for the simulation condition, as it is not ideal for a short assessment that measures six attributes.

For future research, the first goal would be to run the full number of replications of all CD-CAT methods to facilitate a more comprehensive comparison across the results of this study. Along with running all replications, having more test length and attribute conditions will help generalize the results to have broader applications to real data and assessments, such as having a

test length of 15, 20, or even 30 items. Having a higher test length will allow for the opportunity to correctly classify attributes at a higher rate and also allow for another attribute condition, possibly eight attributes.

In this study, it was found that the multinomial extension of SHE was more effective with four than with six attributes. Only two studies (Xue et al., 2003; Xue et al., 2016) compared expected Shannon Entropy to other CD-CAT methods at different attribute levels, and the results were similar for correctly classifying examinees at the varying attribute levels when comparing SHE to the KL-algorithm and other methods. A future study could investigate varying the number of attribute levels, with at least three different conditions, to determine if the multinomial extension of SHE is more effective with a small number of attributes or not.

Second, being able to conduct a research study and compare the GDCM ERUM-MC model to another multiple-choice option-based scoring model, the MC-DINA (de la Torre, 2009) that has been used in a previous study using the Jensen-Shannon Divergence Index (Yigit et al., 2019), would be good to see how it performs amongst the newly introduced methods. Incorporating the Jensen-Shannon Divergence Index into the list of CD-CAT methods would add another data point and tie the study to a method that has already been used in research studies.

Overall, when seeking the method that best classifies examinees into masters and nonmasters, the multinomial extension of SHE should be the first choice. When considering the computational demand, the multinomial extension of SHE may not be the best solution, while the single attribute and composite attribute theoretical CCR methods had significantly less computational demand and still performed well in correctly classifying examinees. When considering all factors, the composite attribute theoretical CCR method is my suggested

approach for a CD-CAT method to use with a multinomial diagnostic classification model, as it outperforms the single attribute theoretical CCR method slightly.

REFERENCES

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. Springer.

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*(3), 403-425. https://doi.org/10.1007/s11336-013-9350-4

Bao, Y., & Bradshaw, L. (2018). Attribute-level item selection method for DCM-CAT. *Measurement: Interdisciplinary Research and Perspectives*, *16*(4), 209–225. https://doi.org/10.1080/15366367.2018.1436824

Burke, M. J., & Henson, R. (2008). LCDM user's manual. *Greensboro: University of North Carolina at Greensboro*.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined Polytomous attributes. *Applied Psychological Measurement*, *37*(6), 419–437. https://doi.org/10.1177/0146621613479818

Chen, J., & Zhou, H. (2017). Test designs and modeling under the general nominal diagnosis model framework. *PLOS ONE*, *12*(6), e0180016. https://doi.org/10.1371/journal.pone.0180016

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*(4), 619-632. https://doi.org/10.1007/s11336-009-9123-2

Cheng, Y. (2010). Improving Cognitive Diagnostic Computerized Adaptive Testing by Balancing Attribute Coverage: The Modified Maximum Global Discrimination Index

Method. *Educational and Psychological Measurement*, *70*(6), 902–913. https://doi.org/10.1177/0013164410366693

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*(3), 163–183. https://doi.org/10.1177/0146621608320523

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130. https://doi.org/10.3102/1076998607309474

de la Torre, J. (2010). "The partial-credit DINA Model," in *Paper Presented at the International Meeting of the Psychometric Society* (Athens).

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333-353. https://doi.org/10.1007/bf02295640

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595-624. https://doi.org/10.1007/s11336-008-9063-2

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179-199. https://doi.org/10.1007/s11336-011-9207-7

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement*, *39*(1), 62–79. https://doi.org/10.1177/0146621614561315

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F.

Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (1st ed., pp. 361–390). Erlbaum.

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. *Cognitive Diagnostic Assessment for Education*, 242–274. https://doi.org/10.1017/cbo9780511611186.009

Haberman, S. J., von Davier, M., & Lee, Y.-H. (2008). Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions (ETS Research Report no. RR-05–16). 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*(4), 301-321. https://doi.org/10.1111/j.1745-3984.1989.tb00336.x

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*.

Henson, R., DiBello, L., & Stout, B. (2018). A generalized approach to defining item discrimination for DCMs. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 18–29. https://doi.org/10.1080/15366367.2018.1436855

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*(4), 262–277. https://doi.org/10.1177/0146621604272623

Henson, R., Roussos, L., Douglas, J., & Xuming He. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement*, *32*(4), 275–288. https://doi.org/10.1177/0146621607302478

Henson, R., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191-210. https://doi.org/10.1007/s11336-008-9089-5

Hsu, C., Wang, W., & Chen, S. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement*, *37*(7), 563–582. https://doi.org/10.1177/0146621613488642

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with Nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272. https://doi.org/10.1177/01466210122032064

Kaplan, M., De La Torre, J., & Barrada, J. R. (2015). New Item Selection Methods for Cognitive Diagnosis Computerized Adaptive Testing. *Applied Psychological Measurement*, *39*(3), 167–188. https://doi.org/10.1177/0146621614554650

Karelitz, T. M. (2004). Ordered category attribute coding framework for cognitive assessments (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign.

Kuo, B., Chen, C., & De la Torre, J. (2017). A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Applied Psychological Measurement*, *42*(3), 179-191. https://doi.org/10.1177/0146621617722791

Kuo, B., Pai, H., & de la Torre, J. (2016). Modified cognitive diagnostic index and modified attribute-level discrimination index for test construction. *Applied Psychological Measurement*, *40*(5), 315–330. https://doi.org/10.1177/0146621616638643

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*(3), 205–237. https://doi.org/10.1111/j.1745-3984.2004.tb01163.x

Li, Y., Huang, C., & Liu, J. (2023). Diagnosing primary students' reading progression: Is cognitive diagnostic computerized adaptive testing the way forward? *Journal of Educational and Behavioral Statistics*, *48*(6), 842-865. https://doi.org/10.3102/10769986231160668

Lin, C., & Chang, H. (2019). Item selection criteria with practical constraints in cognitive diagnostic computerized adaptive testing. *Educational and Psychological Measurement*, *79*(2), 335-357. https://doi.org/10.1177/0013164418790634

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

MacReady, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *2*(2), 99. https://doi.org/10.2307/1164802

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187-212. https://doi.org/10.1007/bf02294535

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Psychology Press.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

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

Sessoms, J., Stout, W., & Henson, R. (2019). *GDCM Simulation Software User Manual*.

Shear, B. R., & Roussos, L. A. (2017). Validating a distractor-driven geometry test using a generalized diagnostic classification model. *Social Indicators Research Series*, 277–304. https://doi.org/10.1007/978-3-319-56129-5_15

Stout, W., Henson, R., DiBello, L., & Shear, B. (2019). The Reparameterized unified model system: A diagnostic assessment modeling approach. *Methodology of Educational Measurement and Assessment*, 47–79. https://doi.org/10.1007/978-3-030-05584-4_3

Stout, W., Henson, R., & DiBello, L. (2023). Three psychometric-model-Based option-scored multiple choice item design principles that enhance instruction by improving quiz diagnostic classification of knowledge attributes. *Psychometrika*, *88*(4), 1299-1333. https://doi.org/10.1007/s11336-022-09885-3

Stout, W., Henson, R., & DiBello, L. (2022). Optimal classification methods for diagnosing latent skills and misconceptions for option-scored multiple-choice item quizzes. *Behaviormetrika*, *50*(1), 177–215. https://doi.org/10.1007/s41237-022-00172-0

Tatsuoka, C. (2002). Data Analytic Methods for Latent Partially Ordered Classification Models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *51*(3), 337–350. https://doi.org/10.1111/1467-9876.00272

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345-354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (1st ed., pp. 327-360). Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. Routledge.

Templin, J. (2004). Generalized linear proficiency models for cognitive diagnosis (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign.

Templin, J. (2006). *CDM user's guide*. Unpublished manuscript.

Templin, J., Henson, R. A., Rupp, A. A., Jang, E., & Ahmed, M. (2008). "Cognitive diagnosis models for nominal response data," in *Paper Presented at the National Council on Measurement in Education* (New York, NY).

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. https://doi.org/10.1037/1082-989x.11.3.287

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

von Davier, M. (2005). A general diagnostic model applied to language testing data (ETS Research Report no. RR-05-16). Princeton, NJ: Educational Testing Service

Von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287–307. https://doi.org/10.1348/000711007x193957

Wang, C. (2013). Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing with Short Test Length. *Educational and Psychological Measurement*, *73*(6), 1017–1035. https://doi.org/10.1177/0013164413498256

Wang, C., Chang, H.-H., & Huebner, A. (2011). Restrictive Stochastic Item Selection Methods in Cognitive Diagnostic Computerized Adaptive Testing: Restrictive Stochastic Methods in CD-CAT. *Journal of Educational Measurement*, *48*(3), 255–273. https://doi.org/10.1111/j.1745-3984.2011.00145.x

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, *52*(4), 457–476. https://doi.org/10.1111/jedm.12096

Xu, G., Wang, C., & Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 291–315. https://doi.org/10.1111/bmsp.12072

Xu, X., Chang, H., & Douglas, J. (2003). A simulation study to compare CAT strategies for cognitive diagnosis. Annual Meeting of the American Educational Research Association, Chicago, IL.

Yigit, H. D., Sorrel, N. A., & De la Torre, J. (2019). Computerized Adaptive Testing for Cognitively Based Multiple-Choice Data. *Applied Psychological Measurement*, *43*(5), 388–401.

Yu, X., Cheng, Y., & Chang, H. (2019). Recent developments in cognitive diagnostic computerized adaptive testing (CD-CAT): A comprehensive review. In M. von Davier & Y. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 307–331). Springer Nature.

Zheng, C., & Chang, H. (2016). High-efficiency response distribution–based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *40*(8), 608–624. https://doi.org/10.1177/0146621616665196

Zheng, C., & Wang, C. (2017). Application of binary searching for item exposure control in cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *41*(7), 561–576. https://doi.org/10.1177/0146621617707509