FU, YANYAN, Ph.D. Comparison of General Diagnostic Classification Model for Multiple-Choice and Dichotomous Diagnostic Classification Model (2018).
Directed by Dr. Robert A. Henson. 134pp.

A submodel of the general diagnostic classification models for multiple choice (GDCM-MC), the excluding guessing from the correct answer (EGCA) model, was first introduced because the submodel with kernel extended reparameterized unified model (ERUM) can be compared directly to the dichotomous reduced reparameterized unified model (RRUM) without model induced bias.

A simulation study was used to demonstrate this equivalence of the EGCA parameters of the correct options and the RRUM item parameters. At the same time, the simulation study was also used to demonstrate the equivalence of the two models when there were no skills or misconceptions measured by the incorrect options, and show the improvement of the EGCA estimation when distractors are created to provide additional information. The results confirmed the equivalence of the EGCA parameters of the correct options and the RRUM item parameters. The results also show that the correct classification rates (CCRs) and test-level cognitive diagnostic index ($CDI_{\bullet}$) were the same for the two models when there was no informative distractor. Additionally, by including weakly informative distractors, the EGCA showed higher CCRs and $CDI_{\bullet}$ than the RRUM. When the distractors were strongly informative, the EGCA had much higher CCRs and $CDI_{\bullet}$. The studies also showed that CCRs and $CDI_{\bullet}$ increased when the sample size, test length, and item quality increased, as well as when the number of measured test skills and misconceptions decreased.

A real-world example was used to compare the classification differences and predictability of the classification on the selection of the options between the two models in a distractor-driven assessment. The results show that the profile classification agreement was 48%, and the classification based on the EGCA was more correlated with the students' selection of the correct or the misconception-embedded options than the classification based on the RRUM. The results indicate that the EGCA provides more realistic classification than the RRUM. The results of both simulation and the real data studies suggest that the polytomous diagnostic classification models (DCMs), rather than the dichotomous DCMs, should be used when the multiple-choice items have informative distractors.

COMPARISON OF GENERAL DIAGNOSTIC CLASSIFICATION MODEL FOR MULTIPLE-CHOICE AND DICHOTOMOUS DIAGNOSTIC CLASSIFICATION MODEL

by

Yanyan Fu

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
2018

Approved by

Robert Henson_____
Committee Chair

APPROVAL PAGE

This dissertation written by Yanyan Fu 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 Member Dr. John Willse

Committee Member Dr. Randall Penfield

Committee Member Dr. Devdass Sunnassee

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

Firstly, I like to thank my advisor Dr. Bob Henson for providing guidance, inspiration, and assistance that help me to start and finish this dissertation. You are an excellent advisor! Thank you for being patient with me for the past five years. It was a great pleasure to work with you. I also like to thank all my dissertation committee members, Dr. John Willse, Dr. Randy Penfield, and Dr. Devdass Sunnassee for being supportive. Your guidance and suggestions on the dissertation were very beneficial to me!

Secondly, I like to thank my father Shunde Fu and my mother Jinhuan Zhi for motivating me to pursue the Ph.D. degree, and all the effort and sacrifice you have made to support me financially and emotionally all these years. Without your help, I would not be able to make to the finish line. I also like to thank my fiancée Dingling Zhong. You are the sweetest person in my life. Thank you for loving and believing in me.

Lastly, I would like to thank all the "GERMs", especially, Jonathan Rollins, Emma Sunnassee, Juanita Hicks, and Tyler Strachan. Thank you all for accompanying me on this journey. You made my five-years' experience exciting and unforgettable.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

The purpose of an educational assessment is to make inferences about teaching, learning and students' specific area of knowledge (Standard, 2014). Educational assessment scores can indicate if teaching goals have been achieved and students have successfully mastered the knowledge and skills in a specific domain. Scores are often used to inform certification decisions. Educational testing can involve low (e.g., classroom assessment). or high stakes (e.g., licensure exams). Formative assessment, an assessment used to provide learning and teaching feedback, is typically a low-stakes test. The students can learn about strengths and weaknesses in skills, knowledge, and abilities, and teachers use the information to improve their instruction by addressing the areas in which the students need more assistance.

In contrast to low-stakes testing, when a result is used to provide an overall evaluation of students and teachers over a learning period, the assessment is referred to as a summative assessment. Summative assessments are often high stakes. For example, the summative scores of a course can be used to indicate whether students can continue on to higher-level courses. Such high-stakes assessments are often standardized and administered at a large-scale level. Cizek (2001) suggested that high-stakes tests are usually reliable, free of bias, and related to public goals because greater effort has been spent on development and calibration to ensure the quality of the test. Cizek (2001)

also argued that high-stakes testing can be used to accommodate minority or disability groups, help the public to learn about students and school performance, and serve as an accurate piece of information for students to learn their achievement levels in different subjects. Moreover, using high-stakes testing to evaluate the performance of teachers can stimulate educators to improve their instruction and enable more critical classroom assessments (Cizek, 2001).

Standardized summative testing can be productive for students, educators, and other stakeholders at global levels. However, formative assessments do have some advantages over summative assessments. Leighton et al. (2010) researched teachers' beliefs about classroom assessment. They found that educators believed that classroom assessments could provide a better understanding of students' learning process and that such assessments were more likely to trigger students' learning than standardized summative assessments. Klute, Apthorp, Harlacher, and Reale (2017) performed a meta-analysis and found that students given formative assessments have better outcomes than students who have not; moreover, that students have higher math scores when formative assessments are used along with lectures.

Although formative assessments can provide detailed diagnostic feedback, they are not as accurate as high-stakes summative tests, because most educators are not experts in test development. As such, relying on educators to develop formative assessments can be challenging and problematic. Therefore, a need exists to create accurate, formative tests to supply better diagnostic information about student skills, knowledge, and abilities. For example, test developers can use an evidence-centered

design (Mislevy, Almond, & Lukas, 2003) to create a test that targets various skills (Rupp, Templin & Henson, 2010).

In addition to the development of an instrument or test, appropriate models should be used that best extrapolate information from the examinees' responses. For educational assessment modeling, the differences between the goals of summative and formative assessments are reflected by selecting a unidimensional or multidimensional latent-attributes model. For example, item response theory (IRT) models are commonly used to examine unidimensional continuous latent attributes. IRT models are widely used by psychometricians because they can help practitioners examine the item qualities (e.g., the difficulty level of an item) and examinees' continuous attributes (e.g., math ability) separately. However, if it necessary to attain more diagnostic information regarding an examinee's set of abilities (i.e., formative assessment), an extension of unidimensional IRT models—that is, multidimensional IRT (MIRT) models—can be used.

Unfortunately, MIRT models have certain application limitations. For example, the computational burden for MIRT models can be very high when using the expectation-maximization algorithm (Cai, 2010; Han & Paek, 2014). The computation time increases exponentially as the number of dimensions or latent attributes increase, and it can be difficult to estimate the multiple latent abilities (Cai, 2010). Additionally, a large number of examinees and longer tests have been needed to estimate the multiple abilities accurately. As a result, other models have been proposed to address the shortcomings of MIRT models.

One method is to treat the different ability scales as dichotomous instead of continuous. Diagnostic Classification Models (DCMs), which model discrete latent attributes, can be used to provide diagnostic information about examinees' attributes. Similar to IRT models, DCMs also separate item-level properties and latent attributes, which can assist practitioners in examining both item quality and diagnostic information of the examinees. For the past two decades, many general DCMs have been developed (de la Torre, 2010; Henson, Templin, Willse, 2009; von Davier, 2005). Similar to the development of IRT models, DCMs originally focused on modeling dichotomous responses and, more recently, such models have focused on polytomous responses. For example, ordinal scale DCMs (e.g., Likert type, partial credit) or nominal scale /multiple-choice DCMs (de la Torre, 2009a; DiBello, Henson, Stout, 2015; Ozaki, 2015; von Davier, 2005) have been developed to capture more useful information in the selection of the options/ ratings by examinees under the DCM framework. More research has applied DCMs to assessing students' latent attributes and better diagnostic information regarding examinee attributes (Kim, 2011; Shear, 2016; von Davier, 2005). The use of DCMs in future formative assessment is promising.

Despite the development of polytomous models in the IRT and DCM framework, people continue to use dichotomous models even when polytomous responses models have been provided in practice (de la Torre, 2009a; Jiao, Liu, Haynie, Woo, & Gorham, 2012). In particular, ordinal scale and nominal/multiple-choice are often dichotomized and modeled using dichotomous DCMs. If the scale is nominal/multiple-choice items, the correct answer is naturally coded as 1, and the other answers are coded as 0. This method

ignores more specific information (such as partial skill or misconceptions) measured by the incorrect options. That is, *why* an item was answered incorrectly is not directly modeled—only *that* the item was answered incorrectly. If the scale is ordinal, one method used to dichotomize the item is by coding the highest score (or the scores that are above the item mean) as 1, and the score lower than the highest score (or the scores that are at or below mean) as 0. By collapsing the multiple categories into two, specific information as to how an item is missed is ignored.

Focusing only on the correct response and ignoring levels of information provided by different categories can reduce the amount of information from any given item. That said, the actual impact of this loss of information on estimating an examinee's ability has not been well studied. Jiao et al. (2012) compared polytomous IRT models to dichotomous IRT models using real data and a simulation study. The results suggest that, although continuous latent ability is highly correlated between the two models using real data, polytomous models provide smaller standard latent-ability errors. Additionally, (de la Torre, 2009a) compared a polytomous DCM— the MC-DINA—to the corresponding dichotomous DCM—the DINA model—in a simulation study and found better recovery of the latent attributes when using the MC-DINA. The results of these two studies indicate the usefulness of applying polytomous IRT models rather than dichotomous models when polytomous responses have been collected.

Most studies comparing polytomous models to dichotomous models should be interpreted with caution. For example, in the previous two cases, data were specifically simulated using the polytomous model, and the polytomous responses were then

dichotomized. However, in these cases, the dichotomized data could not be assumed to exactly fit the analogous dichotomized model. For example, when data are simulated using the MC-DINA and then dichotomized, it cannot be expected that the corresponding DINA model will simultaneously fit these data. Thus, many studies using dichotomous models for the data that was originally simulated using the polytomous model have at least some bias caused by model misfit in the dichotomous model estimation. Most IRT models for polytomous data and their analogous dichotomous models do not simultaneously fit polytomous models, and the dichotomized response data without a model induced bias. The current study first develops a DCM in which polytomous data can be simulated such that a traditional DCM also fits the data when responses are dichotomized. Thus, the effect of modeling the polytomous responses can be estimated without introducing bias caused by model misfit. In addition, as the real data study was only conducted using IRT models (Jiao et al., 2012), it is necessary to compare the polytomous DCMs to the dichotomous DCMs when no model-induced bias exists and when the real data is used.

The following chapter will review common IRT models and DCMs that can be used for dichotomous data and discuss the polytomous versions of the models. In introducing the polytomous models, similarities and differences between the polytomous and the analogous dichotomous approaches will be explored. Next, the factors that impact the distractors of the multiple-choice items will be examined. Previous literature has shown that informative distractors can provide additional information about examinee skills and misconceptions. It is hypothesized that polytomous DCMs will provide a better

estimation of latent attributes when compared to the corresponding dichotomous DCMs if the distractors are informative.

As previously discussed, there is no polytomous DCM that fits polytomous data while simultaneously the analogous dichotomous DCM also fits the same data after being dichotomized. For example, data simulated using the MC-DINA and then dichotomized will not perfectly fit the DINA. Thus, a submodel of the general diagnostic classification models for multiple-choice options (GDCM-MC; DiBello et al., 2015) is defined such that when data is simulated using this model (fitting the polytomous DCM), the dichotomized data also fit the reduced reparameterized unified model (RRUM; Hartz, Roussos, & Stout, 2002). It is hypothesized that this GDCM-MC submodel and the corresponding dichotomous DCM will perform similarly when no informative distractors exist in the items. The advantage of using the GDCM-MC—or more specifically, the submodel defined in the research—over the dichotomous DCM with respect to item discrimination and attributes recovery is larger when more informative distractors, rather than less informative distractors, are used.

CHAPTER II

LITERATURE REVIEW

In educational or psychological assessment several factors are involved when selecting an appropriate model such as the scale of the item response, dimensionality of the construct measured, and the scale of the abilities/attributes. Of particular importance to this research is the scale used for item responses. For example, the scale (or type of item response) can be binary (dichotomous) in which the correct item answer is typically scored as 1, and the incorrect answer is scored as a 0. However, an item response may have three or more levels of response, and these levels could be nominal or ordinal. For instance, students can obtain partial or semi-partial credit for answering a two-point question partially correct. As a result, the range of the score might be 0, .5, 1, 1.5, and 2 with 0 and 2 being the smallest and largest possible score the student can earn, respectively. The responses in between zero and two are ordered from small to large. If there is no ordering of an item's options, the scale is nominal. In other cases, the item responses may be treated as continuous.

Psychometric models, in the most general sense, can be first selected based the number of latent traits the instrument has been designed to measure (the dimensionality) in addition to the general characteristics of such traits. Furthermore, the type of item response can influence a researcher's choice of model. In a case in which only one continuous latent trait is measured by the dichotomous/polytomous-scale test,

dichotomous/polytomous IRT models can be used, such as the Rasch model (Rasch, 1960), two-parameters IRT, a partial credit model (Masters, 1982), and a graded response model (Samejma, 1969). In a context in which more than one continuous ability is measured, multidimensional versions of IRT can be used such as compensatory multidimensional IRT models (CMIRT; Reckase, 1985), as well as non-compensatory MIRT models (NCMIRT; Sympson, 1978), and multidimensional partial-credit models (Reckase, 2009). In contrast to IRT models, DCM models are typically thought to be multidimensional models that measure discrete latent traits (commonly called attributes). Furthermore, specific DCMs have been defined to model polytomous and dichotomous data (de la Torre, 2009a; Dibello, Henson & Stout, 2015; Rupp, Templin & Henson, 2009; Ozaki, 2015; von Davier, 2005). A more specific description of models used to score dichotomous examinee responses is next discussed, followed by a discussion of models for polytomously score items.

2.1 Dichotomous Scale

Although both IRT models and DCMs have different assumptions about latent traits (i.e., dimensionality and scale), both models can be applied to a dichotomous-scale item (e.g., right/wrong) in an assessment. Furthermore, as previously noted, an IRT framework can be used to model unidimensional (UIRT) latent space or multidimensional (MIRT) latent space of continuous abilities. However, most often, the UIRT models have been used in educational assessments. In contrast, DCMs are used in settings where more than one dimension of discrete latent attributes is measured by an assessment. First, the traditional unidimensional IRT models will be discussed, followed

by an examination of multidimensional models for dichotomous response data (i.e., multidimensional IRT models and DCMs).

2.1.1 Unidimensional IRT Models

The Rasch model (Rasch, 1960) defines the probability of a correct response through a linear model such that the log odds are predicted as a function of an examinee's unidimensional ability, θ_j. In this model, the logit link is shown as follows:

$$P(x_{ij} = 1|\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1+\exp(\theta_j - b_i)}, (1)$$

where θ_j is the j^{th} examinee's latent ability and b_i is the difficulty parameter of the i^{th} item. The latent ability θ_j is defined to be continuous and normally distributed, and b_i is usually centered at 0 for identifiability purposes. Note that the larger the value of b_i, the harder the item. The one-parameter (1PL; Birnbaum, 1968) model is essentially the same as the Rasch model, because only item difficulty is included as an item parameter. The difficulty parameters b_i are usually centered at 0 for the Rasch model, whereas the latent ability θ_j are usually centered at 0 for identifiability purposes for the 1PL model.

As psychometric research developed further, more complex IRT models were developed. The general form of a three-parameter (3PL; Birnbaum, 1968) model is shown as

$$P(x_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i) * \frac{\exp a_i(\theta_j - b_i)}{1+\exp a_i(\theta_j - b_i)}, (2)$$

where a_i is the discrimination parameter, which is related to how well an item discriminates "high" ability examinees from "low" ability examinees, and the guessing parameter c_i used to measure the probability that low-ability examinees can guess the item's correct response. If the parameter c_i is set to 0 (i.e., no guessing), then the model is equal to the two-parameter (2PL; Choppin, 1983) model.

2.1.2 Multidimensional IRT Models

Multidimensional IRT (MIRT) models are used when it is assumed that more than one continuous latent ability is being measured. MIRT models are usually thought to be confirmatory (although exploratory methods exist), because researchers must specify the latent abilities measured by each item and the type of interaction among the abilities (e.g., compensatory or non-compensatory). The most common MIRT model is the compensatory MIRT (CMIRT; Reckase, 1985). Compensatory models are typically defined such that lacking or being low on one latent trait can be compensated for by having or being higher on another latent ability. The item response function for CMIRT is represented as:

$$P\left(x_{ij}=1|\theta_{j1},\theta_{j2},\ldots,\theta_{jk},a_{j1},a_{j2},\ldots,a_{jk},b_i\right)=c_i+(1-c_i)*\frac{\exp\left(\sum_{k=1}^{K}a_{ik}\theta_{jk}+b_i\right)}{1+\exp\left(\sum_{k=1}^{K}a_{ik}\theta_{jk}+b_i\right)}, \quad (3)$$

where a_{ik} represents the discrimination for the k^{th} latent trait, b_i is difficulty of the i^{th} item, and θ_{jk} represents the k^{th} of the total of K latent abilities of the j^{th} examinee. An item-level guessing parameter c_i can be added to the model.

The non-compensatory multidimensional IRT model (NCMIRT; Sympson, 1978) defines the probability of a correct response to item *i* for examinee *j* in such a way that lacking or being low on one latent ability cannot be compensated by mastering or having higher levels of other latent abilities measured by that item. The item response is defined as:

$$P(x_{ij} = 1|\theta_{j1}, \theta_{j2}, \dots) = c_i + (1 - c_i) \prod_{k=1}^{K} \frac{1}{1+e^{-(a_{ik}\theta_{jk}+b_{ik})}}, \quad (4)$$

where b_{ik} is the difficulty parameters for the k^{th} latent ability.

Recently, DeMars (2016) developed the partially compensatory MIRT that was inspired by the NCMIRT (1978) and Embretson(1984)'s product model. The model includes the item difficulty and discrimination, as well as the interaction effects, of each latent ability. For example, a two-dimensional partially compensatory model can be expressed as:

$$P(x_{ij} = 1|\theta_{j1}, \theta_{j2}, a_{i1}, a_{i2}, b_i) = c_i + (1 - c_i) * \frac{\exp(a_{i1}\theta_{j1}+a_{i2}\theta_{j2}+a_{i3}\theta_{j1}\theta_{j2}+b_i)}{1+\exp(a_{i1}\theta_{j1}+a_{i2}\theta_{j2}+a_{i3}\theta_{j1}\theta_{j2}+b_i)}, \quad (5)$$

where the a_{i3} is the coefficient of the interaction effect between θ_{j1} and θ_{j2}. The model allows for the possibility that any two pairs of latent abilities interact to influence the probability of a correct response, which is similar to the log-linear model diagnostic classification model (LCDM; Henson, Templin & Willse, 2009).

2.1.3 Dichotomous DCMs

DCMs are psychometric models developed to model examinee responses to an assessment created to measure multiple dichotomous latent attributes. As a result, DCMs can also be expressed as constrained latent class models. Specifically, DCMs "classify" the examinees' abilities into groups based on mastery or non-mastery of a set of latent attributes (discrete latent traits). Thus, any two examinees with the same profile of mastery/non-mastery are thought to belong to the same class. Typically, because of their link to latent class models, estimates of the class membership probability are obtained and usually summarized using the probability of mastery for each latent attribute. Note that DCMs are typically thought to be multidimensional in nature, because a DCM with only one discrete attribute is equivalent to a latent class analysis with only two classes. In settings where the stakes are low and diagnostic information on an examinee's set of latent attributes is needed, DCMs can be used to score a formative assessment that can provide feedback to students' attributes. As a result, teachers can tailor the instruction to student weaknesses.

Like most MIRT models, DCMs are typically thought to be confirmatory, which means that the attributes measured by each item should be known (or defined by the researcher). Once the attributes measured by each test item have been determined, an items-by-attributes indicator matrix, called a Q-matrix is defined. The Q-matrix specifically defines whether the k^{th} attribute is measured by the i^{th} item, $q_{ij} = 1$, or not measured, $q_{ij} = 0$. An example of a three-item, three-attribute Q-matrix is shown in Table 1. A value of "1" means that the item measures the attribute and a value of "0"

means that the item does not measure the attribute represented by that column. In this example, the test measures three attributes: item 1 only measures attribute 2, item 2 only measures attribute 1, and item 3 measures attributes 1, 2, and 3.

Table 1. A Dichotomous Q-Matrix

	A1	**A2**	**A3**
Item 1	0	1	0
Item 2	1	0	0
Item 3	1	1	1

Like the discussed MIRT models, DCMs can also be identified as either compensatory and non-compensatory. Each compensatory model typically has a counterpart that is a non-compensatory model. Compensatory DCMs define the probability of a correct response in such a way that an examinee can compensate for a lack of mastery on some attributes measured by the item by having mastered other attributes. In contrast, for non-compensatory DCMs, having some attributes does increase the chance of getting an item correct when examinees are absent of the other attributes required by the item. A researcher must determine which model to use based on prior information regarding the attributes, as well as a prior theory on how an examinee answers each item.

2.1.3.1 DINA

The non-compensatory "deterministic input noisy and" gate (DINA) model is one of the most widely researched DCMs (DeCarlo, 2011; de la Torre, 2009b; Junker & Sijtsma, 2001), and is known for its parsimonious nature and ease of interpretation of the

model parameters (de la Torre, 2011). Under this model, each item has a "slipping" (s_i) and a "guessing" (g_i) parameter. The "slipping" parameter defines the probability of a person incorrectly responding to an item of which he/she has mastered all of the measured attributes. The "guessing" parameter indicates the probability of a person, who has not obtained mastery for that item, correctly responding to that item (i.e., has not mastered at least of one of the measured attributes). The probability of correctly responding to the i^{th} item by the j^{th} examinee is shown as follows:

$$P(X_{ij} = 1) = (1 - \mathrm{s_i})^{\eta_{ij}} g_i^{1-\eta_{ij}} \text{ , (6)}$$

$$\eta_{ij} = \prod_k^K \alpha_{jk}^{q_{ik}} \text{ ,}$$

where α_{jk} is binary variable indicating if the j^{th} examinee has mastered (α_{jk} =1) the k^{th} attribute or not mastered (α_{jk}= 0). The q_{ik} indicates if the k^{th} attribute is measured (q_{ik} =1) by the i^{th} item or not (q_{ik}= 0). Please note that $\alpha_{jk}^{q_{ik}}$ indicates whether the mastery of k^{th} attribute of the j^{th} examinee influences the probability of a correct response for the i^{th} item. Specifically, if an attribute is measured by the item, and the examinee has mastered the k^{th} attribute, then $\alpha_{jk}^{q_{ik}}$ = 1, otherwise, if the j^{th}examinee has not mastered the k^{th} attribute measured by the item, $\alpha_{jk}^{q_{ik}}$ = 0. Note that in equation 6, it is assumed that $0^0 = 1$. The value *K* represents the total number of attributes, and the product of all $\alpha_{jk}^{q_{ik}}$ suggests that each attribute measured by the item is required for answering the item correctly.

If the value of s_i is high, there is a high chance for examinees that have mastered the measured attributes of the item to miss it. Similarly, if the value of g_i is high, there is a high chance for examinees who do not master the required attributes by the item to answer the item correctly.

2.1.3.2 DINO

The compensatory DINO model (Templin & Henson, 2006) is the counterpart of the non-compensatory DINA. Like the DINA model, the DINO defines the probability of correctly responding to the item as a function of a slipping parameter s_i and a guessing parameter g_i. Thus, the DINO defines the probability of the correct response as:

$$P(X_{ij} = 1) = (1 - \mathrm{s_i})^{\zeta_{ij}} g_i^{1-\zeta_{ij}}, \quad (7)$$

$$\zeta_{ij} = 1 - \prod_k^K (1 - \alpha_{jk})_{jk}^{q_{ik}} \ ,$$

where ζ_{ij} represents whether any attribute required by the i^{th} item is mastered by the j^{th} examinee. If *any* attribute is mastered by the examinee, ζ_{ij} equals 1, otherwise ζ_{ij} equals 0. This model is compensatory in nature because mastering any measured attribute will result in a high chance of a correct response (regardless of which attribute is mastered), even if all other attributes have not been mastered. In addition to ζ_{ij}, the interpretations of the item parameters of the DINO are similar to those of the DINA, with the exception of who is identified as being in the "item master" group.

Although the DINA or DINO models are relatively simple models to use because of ease of interpretation and relatively parsimonious parametrization, they can also be

relatively restrictive. Specifically, both the DINA and the DINO models do not allow for differentiable contribution of attributes to the probability of a correct response. Thus, the two models assume that all attributes measured by an item contribute the same probability of a correct response within an item, which is rarely the case. Other models have been developed to overcome this shortcoming.

2.1.3.3 RRUM

The reduced reparametrized unified model (RRUM; Hartz, Roussos, & Stout, 2002) is a DCM that describes a more complex interaction between the attributes and probability of a correct response. The RRUM estimates how each attribute impacts the response probability. The added complexity has been found to be useful in real-world settings, such as in applications to assess students' language attributes (Kim, 2011). For the RRUM, each item has a baseline probability of being answered correctly, π_i^*, assuming that all measured attributes for that item have been mastered. People who lack any of the measured attributes are then "penalized" by a factor of r_{ik}^*. Because the penalty is defined specifically for that item and attribute, each attributes' contribution is not assumed to be the same for all attributes. The item-response probability function of the RRUM is defined below:

$$P(X_{ij} = 1) = \pi_i^* \prod_{k=1}^{K} r_{ik}^{*\,(1-\alpha_{ik})q_{ik}} , (8)$$

where q_{ik} is the k^{th} attribute measured by the i^{th} item, *K* is the total number of the attributes, π_i^* can be interpreted as the base portion of the i^{th} item assuming mastery of

all measured attributes, and r_{ik}^{*} is the "penalty" for non-mastery of the k^{th} attribute. Lacking the k^{th} attribute required by the i^{th} item will result in the probability of a correct response π_i^{*} being adjusted by a factor of r_{ik}^{*}. The RRUM model provides information on how mastery/non-mastery of a single attribute can impact an item-response probability.

For the RRUM, a combination of the value of parameters π^{*} and r^{*} can be used to describe how well an item can discriminate between examinees that have mastered the attributes and those who have not. A high value of π^{*} (the base probability) indicates a higher level of base probability between examinees who have the attributes and those that do not. If the value of π^{*} is low, both students who master and do not master the attributes measured by the item will have approximately the same probability of answering the item correctly. The low value of r_k^{*} (the penalty probability) indicates a higher discrimination. The probability of answering an item correctly for examinees who master the k^{th} attribute is π^{*}, and is $\pi^{*}r_k^{*}$ for examinees who do not master. The higher the value of r_k^{*} is, the larger reduction in the probability of examinees who do not master, which means the larger difference between the examinees who master the k^{th} attribute and who do not. Thus, it is easier to separate the two groups with a low value of r_k^{*} than with a high value of r_k^{*} . Therefore, for a good-quality item, the π^{*} is high, and r_k^{*} is low, but for a poor-quality item, the π^{*} is low, and r_k^{*} is high.

2.1.3.4 CRUM

The compensatory RUM (CRUM, Hartz, 2002) is the RRUM counterpart, and it is similar to the compensatory MIRT model. Thus, the model indicates that mastering

other measured attributes compensates for an item lacking one measured attribute. The probability of correctly answering an item with a logit link is shown as:

$$P(X_{ij} = 1) = \frac{\exp(\lambda_{i0} + \sum_k^K \lambda_{ik} \alpha_{jk} q_{ik})}{1 + \exp(\lambda_{i0} + \sum_k^K \lambda_{ik} \alpha_{jk} q_{ik})}, \quad (9)$$

where λ_{i0} is the intercept of the i^{th} item, and λ_{ik} is the coefficients of the product of $a_{jk} q_{ik}$. If the k^{th} attribute is measured by the i^{th} item, the q_{ik} is 1 otherwise is 0.

The aforementioned specific models have simple parameterization that can assist researchers to in making judgments concerning item quality based on the value of the parameters. However, when prior knowledge of the attributes and how these attributes are used to answer an item are unknown to the researchers, selection of compensatory and non-compensatory is not easy to determine for some attributes or specific items. Also, if, a combination of both compensatory and non-compensatory attributes are measured by a test, choosing a specific model would be a difficult task to do and possibly unreasonable.

Within the last two decades, a series of general DCMs have been developed: the general diagnostic model (GDM; von Davier, 2005), the log-linear cognitive diagnosis classification model (LCDM; Henson, Templin, & Willse, 2009), and the G-DINA model (de la Torre, 2011). As a result, general DCMs can be estimated, and the estimated parameters can serve as a potential guide for a specific model selection for a set of latent attributes. If possible, the reduction of the general model to a simple model may improve estimation because of the reduction of the number of parameters estimated in the model.

2.1.3.5 GDM

The GDM (von Davier, 2005), which is the similar to the compensatory RUM, also models an examinee's response using an intercept and latent attributes' main effects. However, the GDM can be extended to polytomous attributes (a latent attribute having two or more levels) as opposed to only two levels (mastery and non-mastery). The GDM and the CRUM both focus on the additive attributes that can impact the probability of answering an item correctly. Similar to the IRT model discussed, the GDM uses a logit link function to build the connection between a dichotomous outcome and latent dichotomous attributes. The function probability of a correct response using the GDM is:

$$P(X_i = 1) = \frac{\exp(\beta_i + \gamma_i^T h(q_i, a))}{1 + \exp(\beta_i + \gamma_i^T h(q_i, a))}, \quad (10)$$

$$h(q_i, a) = (q_{i1}a_1, \ldots, q_{iK}a_K),$$

$$\gamma_i^T = \begin{matrix} r_{i1} \\ \vdots \\ r_{iK} \end{matrix}.$$

One limitation of the discussed applications of the GDM is that it is typically presented as an additive model without interactions. As a result, the GDM may not capture the interaction of effects among the latent attributes measured by the items. If these interactions are large, the omission model misfit and lead to model misfit and misclassification of latent attributes if such interaction effects are large. The LCDM addresses this limitation.

2.1.3.6 LCDM

The LCDM is a general model that can be used to fit all of the specific aforementioned DCMs. The LCDM defines the probability of a correct response as a function of all main attributes effects plus all two- or more-way attributes interaction (Henson, Templin, &Willse, 2009). It also uses the logit link function for modeling the dichotomous response pattern. In fact, the LCDM can be defined in a similar way as the GDM in equation 11, however instead of defining $\gamma_i^T h(q_i, a)$ as a sum of only main effects, the LCDM also considers all possible attribute interactions. When $\gamma_i^T h(q_i, a)$ is extended the, the LCDM defines the probability of a correct response as:

$$P(X_{ij} = 1) = \frac{\exp\left(\lambda_{i0} + \sum_k^K \lambda_{ik}\alpha_{jk}q_{ik} + \sum_{v=1}^K \sum_{u>v} \lambda_{ivu}\alpha_{jv}q_{iv}\alpha_{ju}q_{iu} + \cdots\right)}{1 + \exp\left(\lambda_{i0} + \sum_k^K \lambda_{ik}\alpha_{jk}q_{ik} + \sum_{v=1}^K \sum_{u>v} \lambda_{ivu}\alpha_{jv}q_{iv}\alpha_{ju}q_{iu} + \cdots\right)}, \quad (11)$$

Thus, in addition to the intercept and the main effects, where λ_{ivu} is the coefficient of the two-way interaction effect between the two attributes measured by the i^{th} item. If more than 2 attributes are measured by the item, then all possible two-way effects could be included in addition to all three-way effect and so on Because each parameter is at the item level, the number of parameters increases exponentially as the number of the attributes increase.

The model contains an item specific intercept, the main effects of each attribute measured by the i^{th} item, and the two-ways or more-ways interaction among the measured attributes. By constraining the parameters in the LCDM, the model can be transformed to a DINA, DINO, RRUM, and specific cases of a GDM (Henson et al.,

2009). For example, if all but the highest order interaction effects are constrained to 0, the LCDM will be mathematically equivalent to the DINA model, while the LCDM will be equivalent to the CRUM model or the typical presentation of the GDM for modeling dichotomous attributes and dichotomous responses by omitting the interaction effects.

The models such as the UIRT, MIRT, GDM, and LCDM use a traditional link, the logit link, to model dichotomous. However, there are times when alternative links may provide some benefit. For example, if modeling the log-probability, a product may be expressed a linear sum of main effects as opposed to requiring an interaction term. Because several link functions could be used for categorical data analysis (Agresti, 2007), one other model has been proposed by de la Tore (2011). The generalized DINA model (G-DINA; de la Torre, 2011) with a log link function has been proposed. de la Torre (2011) allows for a number of links, when this link is equal to logic link then the parametrization is identical to the LCDM. However, he discusses additional links such as the log link and the identity link. Despite the fact that these different links may lead to different estimates, the G-DINA is a saturated model just as the LCDM and thus, the results of the two models should be equivalent (de la Torre, 2011).

The IRT and DCM models previously discussed are typically applied to item response data that are dichotomous. However, when the scored responses are polytomous (i.e., ordinal or multinomial) and thus have more than two levels, varying levels of diagnostic information can be obtained. Ignoring the levels of information could potentially lead to less accuracy in estimating the examinee attributes (Jiao et al., 2012).

There are IRT models, and DCMs that have been developed for polytomous response data.

2.2 Polytomous Scale

Although the literature typically uses the term polytomous to refer to models for an ordinal scale, the term technically refers to an item response with more than two categories. Thus, this document uses the phrase“polytomous item” to indicate an item that is either ordinal or nominal (commonly referred to as a multinomial model). Many polytomous IRT models have been developed to measure an item with multiple scoring categories such as nominal, interval and ordinal scales (Andrich, 1978; Masters, 1982; Muraki, 1992; Samejima, 1969; Thissen &Steinberg, 1984). On the other hand, in recent years, researchers have started to develop polytomous models in the DCM framework. Several polytomous DCMs can now be used for the analysis of this type of data can now be used for the analysis of this type of data (de la Torre, 2009; DiBello, Henson, & Stout, 2015; Ozaki, 2015).

2.2.1 Ordinal Scale IRT Models

There are two different types of commonly-used polytomous IRT models for items that are scored using an ordinal scale: the adjacent category model and the cumulative model (Penfield, 2014). The adjacent category model includes the partial credit model (PCM; Masters, 1982), the generalized partial credit model (GPCM; Muraki, 1992), and the rating scale model (RSM; Andrich, 1978). The cumulative model is referred to as the graded response model (GRM; Samejima, 1969).

2.2.1.1 PCM

The PCM consists of step-wise item response functions. Each step function models the probability of choosing an item's adjacent categories, and the number of steps is equal to the total number of score categories minus one. A step function can be defined using the 1PL IRT model, and can be expressed as follows:

$$logit(\Psi_{it}) = a_i\theta_j + b_{it}, \quad (12)$$

where *t* represents the t^{th} step of the i^{th} $item$, a_i represents the discrimination of θ_j of the t^{th} step, and b_{it} is the difficulty of the t^{th} step of the i^{th} item. The Ψ_{it} step represents the probability of choosing *t* and *t*-1 for the i^{th} item.

For instance, an item that contains four categories (0,1, 2, and 3) has a total of three steps. The first step, step 1, models the probability of choosing 0 versus 1, the second step compares 1 versus 2 assuming that 0 has not been selected. The final step in this example (step 3) models the probability of choosing 2 versus 3 assuming that neither 0 nor 1 have been selected. The step function of the PCM assumes that the discrimination parameters of all steps are equal to 1, which is the feature of the Rasch model. The probability of choosing each category can be solved using the step function. The difficulty of each step b_{it} is different for each item, and the higher value of the b_{it} is, the less difficult the step. It is assumed that as the latent abilities' level increase, the probability of being in higher "steps" increases.

Using algebraic manipulation, the probability of choosing an item response using the PCM can be defined as follows

$$P(X_{ij} = x) = \frac{\exp(\sum_{c=0}^{x}(\theta_j + b_{ic}))}{\sum_{k=0}^{m_i}(\exp(\sum_{c=0}^{k}(\theta_j + b_{ic}))}, x = 0, 1, 2, \ldots, m_i. \quad (13)$$

$$c = 0, \sum_{c=0}^{0}(\theta_j + b_{i0}) = 0$$

where x is the value (number of categories) of an item that the j^{th} examinee responds, θ_j the latent attribute of the examinee j, m_i is the number of steps of the i^{th} item, and b_{ic} is the category difficulty of the c^{th} category of i^{th} item.

2.2.1.2 RSM

The RSM (Andrich, 1978) is a constrained version of the PCM. As in the previously described models, it uses adjacent categories steps to model the probability of choosing the adjacent categories given an examinee's latent ability. The steps of each item have the same discrimination. However, the distances between each step are fixed for each item, and the difficulty varies across items. This constraint allows different items to have steps that are separated by the same distances but vary with respect to difficulty across items. The response function is shown as follows:

$$P(X_{ij} = x) = \frac{\exp(\sum_{c=0}^{x}(\theta_j + d_i + t_c))}{\sum_{k=0}^{m_i}(\exp(\sum_{c=0}^{k}(\theta_j + d_i + t_c))}, x = 0, 1, 2, \ldots, m_i \quad (14)$$

$$c = 0, \sum_{c=0}^{0}(\theta_j + d_i + t_0) = 0 \ ,$$

where t_c is the distance between each step of an item and the i^{th} item difficulty. Note that the t_c is only different between steps and does not change across items. The RSM can be useful for the Likert-type scales when all the items use the same Likert-type scale and the distance between each step and the item difficulty can be assumed to be equal across all items.

2.2.1.3 GPCM

Compared with the PCM, the GPCM (Muraki, 1992) relaxes the assumption of constraining all the discrimination parameters across items on a test to be 1. The discrimination parameters are freely estimated across all items under this model. Both the PCM and the GPCM allows for different categories or steps to have different levels of difficulty. The response function of the GPCM can be shown as follows:

$$P(X_{ij} = x) = \frac{\exp(\sum_{c=0}^{x}(a_i\theta_j + b_{ic}))}{\sum_{k=0}^{m_i}(\exp(\sum_{c=0}^{k}(a_i\theta_j + b_{ic}))} , x = 0, 1, 2, \dots, m_i \text{ ,(15)}$$

$$c = 0, \sum_{c=0}^{0}(a_i\theta_j + b_{i0}) = 0 .$$

2.2.1.4 GRM

The GRM (Samejima, 1969) uses cumulative steps to model the responses rather than adjacent categories of each item. It uses 2PL as the step function, but each step function represents the probability of choosing that category above that value versus the probability of selecting a category less. The step function can be shown as follows:

$$logit(\Psi_{it}) = a_i\theta_j + b_{it}, \text{ (16)}$$

where Ψ_{it} is the $P(X_{ij} \geq t)$. For example, when there are four categories (i.e., 0,1,2, and 3) in an item, Step 1 means that the probability of choosing categories 1, 2, and 3, versus the probability of selecting 0. Step 2 models the probability of choosing category 2 or 3 versus the probability of selecting 0 or 1, and so forth. Similar to the GPCM, the GRM assumes that each step of an item has the same discrimination parameter. However, each step of the GRM has a unique level of difficulty, and from lower steps to higher steps, the level of difficulty increases. As the level of the latent attribute increases, the probability of moving from lower steps to higher steps also increases. The response function is defined as follows:

$$P_i(X_{ij} = j) = \Psi_{ij}(\theta_j) - \Psi_{i(j+1)}(\theta_j), \quad (17)$$

$$\text{If j =0, } \Psi_{ij}(\theta_j)=1.$$

2.2.2.5 Multidimensional GPCM

A unidimensional polytomous IRT models can be easily extended to multidimensional cases (Chen, 2017; de la Torre, 2009b). In the framework of a CMIRT, each item can measure more than one dimension. However, polytomous MIRT models often assume that test items are simple-structure where only one latent ability is measured by an item (Chen, 2017). For example, a multidimensional GPCM can be shown as follows:

$$P(X_{ij} = x) = \frac{\exp(\sum_{c=0}^{x}(a_{i(d)}\theta_{(d)j} + b_{i(d),c}))}{\sum_{k=0}^{m_{i(d)}}(\exp(\sum_{c=0}^{k}(a_{i(d)}\theta_{(d)j} + b_{i(d),c}))}, x = 0, 1, 2, \ldots, m_{i(d)}., \quad (18)$$

$$\text{when } c = 0, \sum_{c=0}^{0}\left(a_{i(d)}\theta_{(d)j} + b_{i(d)0}\right) = 0$$

where d represents the dimension measured by the test. Note that the $a_{i(d)}$, $b_{i(d)c}$, and $\theta_{j(d)}$ are unique to the i^{th}item of the d^{th} dimension.

2.2.2 Nominal /Multiple-Choice IRT Models

2.2.2.1 NRM

Multiple-choice items usually involve non-ordering nominal responses; other than the correct options, the incorrect options are not particularly ordered. The nominal response model (NRM; Bock, 1972) is proposed to model the correct and incorrect options of a multiple-choice item. Similar to the ordinal category model, a step function is used to model the probability of the correct option against the probability of an incorrect option,

$$logit(\Psi_{it}) = -a_{it}\theta_j - b_{it}, t = 1, 2, \dots, m_i, \text{ (19)}$$

where Ψ_{it} is the probability of $X_{ij} = 0$ against the probability of $X_{ij} = t$. Zero is the correct option.

Each step has a discrimination parameter and a difficulty parameter. The discrimination parameter a_{it} of the t^{th} step of the i^{th} item indicates the discrimination level of the incorrect option against the correct option, and difficulty parameter b_{it} estimates the log odds ratio of choosing the correct option against the incorrect option (Bock, 1972; Penfield, 2014). The probability of selecting an option is defined in the equations 20 and 21:

For correct options:

$$P(X_{ij} = 0) = \frac{1}{1+\sum_{c=1}^{m_i} \exp(a_{ic}\theta_j + b_{ic})}; (20)$$

and for incorrect options:

$$P(X_{ij} = h) = \frac{\exp(a_{ih}\theta_j + b_{ih})}{1+\sum_{c=1}^{m_i} \exp(a_{ic}\theta_j + b_{ic})}, h = 1,2, \dots, m_i. (21)$$

where m_i is the total number of the h^{th} option of the i^{th} item minus one.

2.2.2.2 Modified NRM

However, the Bock's NRM (1972) does not consider the possibility of randomly guessing, even if that person's ability is to be considered relatively low (Penfield, 2014). A modified NRM model was proposed to model instances when an examinee "does not know" the answer of a multiple-choice item and engages in random guessing (Samejima, 1979). Based on this model, selecting an option is determined by a mixture of the influence of the latent ability and random guessing.

An imaginary "does not know the correct answer" option is created and denoted as 0. The probability of selecting an option is given in equation 22.

$$P'(X_{ij} = h) = P(X_{ij} = h) + \frac{1}{m_i} * P(X_{ij} = 0)$$

$$= \frac{\exp(a_{ih}\theta_j + b_{ih})}{\sum_{c=0}^{m_i} \exp(a_{ic}\theta_j + b_{ic})} + \frac{1}{m_i} * \frac{\exp(a_{i0}\theta_j + b_{i0})}{\sum_{c=0}^{m_i} \exp(a_{ic}\theta_j + b_{ic})}, h = 1,2, \dots, m_i, (22)$$

where m_i is the number of options. The first part is the probability of selecting any option except for the imaginary "does not know correct answer" option, and second part is the guessing portion that is the product of weight $1/m_i$ and the probability of imaginary "does not know correct answer" option. The model suggests that the probability of choosing an option is influenced by the probability of selecting an option based on one's ability and random guessing. The model also requires that the probability of the imaginary "does not know correct answer" option decreases as the ability increases.

Samejima's NRM (1979) assumes the weights of the guessing portion are the same across all options, whereas Thissen and Steinberg (1984) introduced a different NRM that estimates different weights that depends on the option characteristics. The newly modified NRMs can model random guessing from the multiple-choice response data. As a result, these models can be potentially useful in modeling item response behavior in a multiple-choice setting when examinees may also be guessing.

2.2.2.3 Multidimensional NRM

Similarly, the multidimensional NRM model can be written as follows (assuming the correct answer is represented by category 0):

For correct answer

$$P(X_{ij} = 0) = \frac{1}{1+\sum_{c=1}^{m_{i(d)}} \exp(a_{i(d),c}\theta_{(d)j}+b_{i(d),c})}; (23)$$

and for incorrect options:

$$P(X_{ij} = h) = \frac{\exp(a_{i(d),h}\theta_{(d)j} + b_{i(d),h})}{1 + \sum_{c=1}^{m_{i(d)}} \exp(a_{i(d),c}\theta_{(d)j} + b_{i(d),c})}, h = 1,2, \ldots, m_{i(d)} \text{ . (24)}$$

The models can recover the correlation between the measured latent attributes (Chen, 2017) because the model assumes each item only measures one dimension. However, similar to the CMIRT, when many items measure more than one latent ability, the polytomous MIRT may have difficulty to recover the latent correlation (Han & Paek, 2014). Also, because the polytomous MIRT models have more parameters, and the latent abilities are on the continuous scale, a large sample size may be required to recover the parameters and latent abilities well. In contrast, the requirement of sample size is lower in the polytomous DCM framework because the latent attributes are not on a continuous scale (Templin & Bradshaw, 2013).

2.2.3 Polytomous Scale DCMs

For the past decade, many DCM models have been developed to model polytomous response data. For example, the polytomous general diagnostic model (pGDM; von Davier, 2005) can be used for an ordered scale (e.g., Likert-Scale). Most of the other developed polytomous DCMs are used to model multiple-choice data such as the MC-DINA(de la Torre, 2008), the SICM model (Bradshaw and Templin, 2014), GDCM-MC(DiBello, Henson, Stout, 2015), and structured MC-DINA models(MC-S-DINA; Ozaki, 2015) use a multinomial model approach.

2.2.3.1 Polytomous GDM

The von Davier (2005)'s pGDM adopted the partial credit IRT model. The formula is shown in equation 25. Note that pGDM subsumes the aforementioned dichotomous GDM in which x is only 0 or 1.

$$P(X_{ij} = x) = \frac{\exp(\beta_{ix} + \sum_{k=1}^{K} x\gamma_{ik} q_{ik} a_{kj})}{1 + \sum_{y=1}^{m_i} \exp(\beta_{iy} + \sum_{k=1}^{K} y\gamma_{ik} q_{ik} a_{kj})}, \quad (25)$$

$$x = 0, 1, \ldots, m_i \ ,$$

where a_{kj} is the k^{th} attribute of the j^{th} examinee, m_i is the the number of options of the i^{th} item minus 1, and q_{ik} is a Q-matrix entry at the item-level that indicates whether the k^{th} attribute is measured by the i^{th} item. The parameters γ_{ik} is the coefficient of the latent attribute, and β_{ix} is the intercept for specific category x. Each $x\gamma_{ik}$ is weight of the main effect of the k^{th} attribute of the i^{th} item, meaning that the attributes measured by higher score have larger weight. Such weighting method can be limited because the attributes' main effects may not be a function of score categories. In addition, the pGDM only assumes main effects of the attributes, which are measured at the item-level, and can be limited to cases where a large interaction of the attributes can influence the probaiblity of selecting an option.

Multiple-choice items are frequently used in an educational assessment. Multiple-choice items are often scored as correct or incorrect and, as such, treated as binary. However, the incorrect answers (i.e., distracters) of an item can be embedded with additional information about an examinee's ability. Specifically, if the created distracters

are based on a potential set of reasons why a student may not know how to respond to an item, then the way in which the student incorrectly responds can also give information about certain skills or misconceptions. If this additional information can be modeled or extracted, the diagnosis of examinee attributes may be improved so that educators can have a better understanding of examinee attributes without adding additional items.

2.2.3.2 SICM

The scaling individuals and classifying misconceptions (SICM) model (Bradshaw & Templin, 2014) was developed to model multiple-choice items. The SCIM model combined the IRT and DCM in that a continuous latent ability serves as the general unidimensional ability, and the multiple dichotomous misconceptions serve as the nuisance dimensions that leads to local dependency among the items (Bradshaw & Templin, 2014). It uses the log ratio between the probability of answering incorrect options and the probability of answering the correct option as the dependent variable, and the latent covariates (i.e., the latent continuous trait and dichotomous misconceptions) as predictors. The probability of selecting the correct option defined as a function of only an examinee's continuous ability. This probability of a correct response provides a baseline probability, whereas the incorrect options only measure misconceptions but not the continuous latent attribute. The response function of selecting an incorrect option is defined as:

$$P\left(X_{ij} = x_{ic}\right) = \frac{\exp\left(\lambda_{x_{ic},0} - \lambda_{x_{iC},\theta}(\theta_e) + \boldsymbol{\lambda}_{\boldsymbol{x_{ic}}}^{\boldsymbol{T}}\boldsymbol{h}\left(\boldsymbol{\alpha_j}, \boldsymbol{q}_{\boldsymbol{x_{ic}}}\right)\right)}{\sum_{h=1, h\neq H}^{m_i} \exp\left(\lambda_{x_{ih},0} - \lambda_{x_{iH},\theta}(\theta_e) + \boldsymbol{\lambda}_{\boldsymbol{x_{ih}}}^{\boldsymbol{T}}\boldsymbol{h}\left(\boldsymbol{\alpha_j}, \boldsymbol{q}_{\boldsymbol{x_{ih}}}\right)\right)}, \quad (26)$$

where *c*(numerator) and *h* (denominator) are the notation for the same incorrect option, *C* (numerator) and *H* (denominator) are the same notation for the correct option. x_{ic} represents the c^{th} incorrect option of i^{th} item, m_i is the total number of incorrect options of the i^{th} item, $\boldsymbol{q}_{\boldsymbol{x_{ic}}} = q_{x_{i1}} \quad q_{x_{i2}} \quad \cdots \quad q_{x_{ic}}$, is the Q vector that indicates whether the misconception is measured by the c^{th} option across all misconceptions of a test, $\lambda_{x_{ic},0}$ is intercept that represents the logit of the incorrect option over correct option, and $\lambda_{x_{iC},\theta}$ is the coefficent of continous scale θ represented by the correct option *C(or H)*, and $\boldsymbol{\lambda}_{\boldsymbol{x_{ic}}}^{\boldsymbol{T}}\boldsymbol{h}(\boldsymbol{\alpha_j}, \boldsymbol{q}_{\boldsymbol{x_{ic}}})$ consists of all the main effects and interaction effects of the discrete misconceptions measured by the c^{th} incorrect option.

The Q-matrix of the SICM only includes misconceptions measured by each incorrect option because the correct option only measures the continuous ability and not any misconceptions (Bradshaw & Templin, 2014). For example, a Q-matrix of three incorrect options that measure four attributes can be shown in Table 2, in which each distractor measures at least one misconception. Although the SCIM can be a useful tool for measuring multiple dichotomous misconceptions, it does not allow for the possibility of items to directly measure dichotomous skills of interest.

Table 2. The Q-Matrix of an Item.

Option No.	**A1**	**A2**	**A3**	**A4**
0	-	-	-	-
1	1	0	0	0
2	0	1	0	0
3	1	0	1	0

2.2.3.3 MC-DINA

Other multiple-choice DCMs have been developed to allow for the possibility of modeling multiple skills. For example, the multiple-choice DINA (MC-DINA) model (de la Torre, 2009a) was proposed to model the attributes measured by each multiple-choice item. Like the SICM, the Q-matrix of the MC-DINA model is option-based (i.e., a Q-matrix vector is defined for each option). However, unlike the SICM, the MC-DINA considers the attributes measured by each option *including* the correct choice. Because it is a DINA-based model, each item-option coded Q-vector identifies a latent class "group", which may contain more than just one profile, that should be attracted to that option. The number of latent class "groups" of an item is equal to the number of uniquely coded Q-vectors plus one because a reference group, group 0, is defined that has not mastered any set of measured attributes for any of the options. Group 0 will have an equal probability of choosing each option of the item (i.e., random guessing), and other latent class groups will have a modeled parameter describing the probability of choosing the "attractive" option (i.e., in which the item-option Q-matrix matches the examinees' attribute profile).

The MC-DINA estimates the probability of choosing an option given a latent group. Simulation results (de la Torre, 2009a) show that the MC-DINA has better attribute profile classification rates than the dichotomous DINA model. However, there are no model parameters that specifically define the property of each option, which makes it difficult for practitioners to evaluate the quality of each distracter.

2.2.3.4 MC-S-DINA

To make the model easier to use, Ozaki (2015) proposed several MC-DINA model alternatives, namely, multiple-choice structural DINA (MC-S-DINA) models. In these models, a random guessing portion was added to the model that is based on the probability of missing the option in which the measured attributes are matched with an examinee's mastered attributes. The models also include parameters that indicate the probability of "slipping" and missing the option for examinees who have mastered all the required attributes, which describes the item difficulty. The MC-S-DINA II, one of the three proposed models, is discussed here. For the MC-S-DINA II, each option has a "slipping" parameter. The probability of choosing the c^{th} option, given an attribute profile of the j^{th} examinee, is shown as:

$$P(X_{ij} = c|\alpha_j) = \gamma_{ij}(1-\delta_{ic})^{\eta_{ijc}}\left(\frac{\beta_{ij}}{C-1}\right)^{(1-\eta_{ijc})} + \frac{(1-\gamma_{ij})}{C}, \quad (27)$$

$$\eta_{ijc} = \prod_{k=1}^{K}(2 - 2^{(\alpha_{jk}-q_{ikc})^2}),$$

$$\gamma_{ij} = \sum_{c=1}^{C}\eta_{ijc}\left[1 - \prod_{k=1}^{K}(1-\alpha_{jk})\right],$$

$$\beta_{ij} = \sum_{c=1}^{C}\delta_{ic}\eta_{ijc},$$

where *C* is the total number of options, and *K* is the total number of skills measured by the item. η_{ijc} is equal to 0 if the required attributes of an option are not fully mastered by the examinee, otherwise it is equal to 1. γ_{ij} is 0 if the examinee does not have any attributes measured by any item options, and the examinee has equal probability of choosing any option. β_{ij} is equal to the probability of missing the option in which all the

measured attributes are mastered by the j^{th} examinee, and it is weighted by the number of options minus 1 (C-1).

The other two MC-S-DINA models are similar to the MC-S-DINA II. One model has only one "slip" parameter at the item level, while the other model has more parameters that model even more attribute interactions and options than the MC-S-DINA II. Simulation results show that the proposed models have better attribute profile recovery than the MC-DINA model (Ozaki, 2015), suggesting that the models can be potentially useful in diagnosing students' strengths and weaknesses using the multiple-choice items.

2.2.3.5 GDCM-MC

Concurrently with Ozaki's (2015) work, a family of general diagnostic classification models for multiple-choice options (GDCM-MC; DiBello et al., 2015) was also proposed, which are central to this research. The GDCM-MC defines a general framework where it is assumed that how an examinee responds to a multiple-choice item depends on the mastery of skills and misconceptions. Note that because the model specifically focused on both skills and misconceptions DiBello et al. (2015) referred them generically as attributes that were either possessed or not. For consistency, this terminology will not be adopted when discussing this model and the submodel defined for this research. Instead, the terms attribute and mastery/non-mastery will continue to be used. Under this framework, a function is used to define the attractiveness of each option. Using the GDCM-MC, the function used to identify the attractiveness of each option can be related to any of the abovementioned dichotomous DCM models (although slightly modified). Note that, although this is not entirely the case, the idea of modeling each

option is “as though” they are pseudo -items. For example, the GDCM-MC could use a version of the RRUM, DINA, or even the LCDM. This function was referred to as the “Kernal Function” by DiBello et al. (2015) and was notated as, $F_{ih}(\alpha)$. Furthermore, the GDCM-MC uses an option-based Q-matrix, but made a modification to allow for a specific focus on both skills and miscoceptions. Finally, similar to Ozaki (2015) MC-S-DINA models, the GDCM-MC allows for the potential of random guessing, which is weighted by the number of options.

Although the choice of “Kernel Functions” in GDCM-MC will not be discussed further, a dichotomous DCM previously discussed in this document could be used. However, a more thorough discussion of the Q-matrix for the GDCM must be provided. An additional entry was necessary for the GDCM-MC because the GDCM-MC measures both skills, and misconceptions and mastery or non-mastery of either may simultaneously make one option attractive and another unattractive. Specifically, entries specify the pattern of skills and misconceptions that make that option most attractive. Thus, the third entry of an “N” was added to suggest that a particular attribute did not directly impact the attractiveness of a given option. An example of the GDCM-MC Q-matrix for a single four-option item that measures three attributes (of which some could be skills and other misconceptions) is shown below.

Table 3. A GDCM-MC Q-Matrix of an Item

Option No.	**A1(Skill 1)**	**A2(Skill 2)**	**A3(Misconception)**
1	1	1	0
2	N	1	N
3	1	N	N
4	N	N	1

Each row of the Q-matrix represents an option, and each column of the Q-matrix represents an attribute (skill or misconception). In total, there are four options and three attributes measured by this item. The entry of a value 1 means that mastery of the attribute would result in higher attraction to that option, whereas an entry of 0 means that lacking mastery of that attribute is will result in a higher attraction to it. Generally, matching the 1's and 0's in a Q-matrix option lead to a higher attraction of that option. An entry of "N" means that the attribute is irrelevant for that option, and as such, does not directly influence the attractiveness of that option.

Given the Q-matrix, the GDCM-MC consists of a weighted combination of a cognitive portion and a guessing portion. The cognitive portion is very similar to a typical multinomial model that is a part "divided by the total". Whereas the guessing portion assumes that, when an examinee does guess, all options are equally attractive, and thus, the probability of selecting any option is the inverse of the number of options for that item. Given these two parts, the GDCM-MC defines the probability of selecting the h^{th} option of the i^{th} item as follows:

$$P_i(h|\boldsymbol{\alpha}) = \frac{F_{ih}(\boldsymbol{\alpha})}{S_{\alpha}}\omega_{i\boldsymbol{\alpha}} + \frac{1}{H_i}(1 - \omega_{i\boldsymbol{\alpha}}) \text{ , (28)}$$

$$\omega_{i\boldsymbol{\alpha}} = \min\{1, \sum_{h'=1}^{Hi} F_{ih}(\boldsymbol{\alpha})\},$$

$$S_\alpha = \sum_{h=1}^{Hi} F_{ih}(\boldsymbol{\alpha}).$$

In equation 28, the probability of selecting a given option is a weighted sum between the probability of cognitively selecting that option, $\frac{F_{ih}(\alpha)}{S_\alpha}$, and the probability of selecting that option when guessing, $\frac{1}{H_i}$. Specifically, $F_{ih}(\alpha)$ is the kernel function that represents the attractiveness of the h^{th} option of the i^{th} item, and can be any aforementioned dichotomous DCM (this particular research will use the RRUM). Note that kernel $F_{ih}(\alpha)$ is related to the probability of cognitively choosing that option, thus, this portion is modeled in the same way as a typical dichotomous DCM, and S_α is the sum of kernel $F_{ih}(\alpha)$ across all options. H_i indicates the number of options for the i^{th} item. As was mentioned, the value $\omega_{i\boldsymbol{\alpha}}$ defines a weight that is placed on the cognitive portion of the model relative to the guessing portion. Notice that this weight depends on S_α which is the sum of the kernel function across all options. Because the kernels represent the attractiveness of each option, when the "attraction" of the options, in general, is relatively high, $S_\alpha > 1$. As a result, $\omega_{i\alpha} = 1$ and, thus, all weight is placed on the cognitive ability (i.e., the response probability is $\frac{F_{ih}(\boldsymbol{\alpha})}{S_\alpha}$). If, however, all options are not that attractive, $S_\alpha < 1$ and as a result $\omega_{i\alpha} < 1$. When $\omega_{i\alpha} < 1$, at least some portion of the probability of selecting an option will be due to random guessing. Most importantly to this research, in general, the probability of selecting any option is not equal to any dichotomous function alone because of the guessing part for an option or the division of S_α. Thus, even when a

specific kernel is selected, the corresponding dichotomous model will not perfectly fit the polytomously scored data after recoding it as a dichotomous (right/wrong) response.

Although the GDCM-MC provides a useful framework for a number of possible models, one must choose a kernel, usually based on the particular items. DiBello et al. (2015) specifically discuss the estimation of the GDCM-MC when using the RRUM as the kernel, $F_{ih}(\alpha)$. When using the RRUM as a kernel function, the model is named the extended reparameterized unified model (ERUM). Using the RRUM as a base (RRUM; Hartz, Roussos, & Stout, 2002), the kernel is defined as:

$$F_{ih}(\alpha) = \pi_{ih} \prod_{k=1|q_{ihk} \neq N} r_{ihk}^{|q_{ihk} - \alpha_k|} . \quad (29)$$

Both the ERUM and RRUM use π and r to model the respondents' item response behavior given the respondents' latent attribute profile $\boldsymbol{\alpha}$. π_{ih} can be viewed as the attractiveness of the h^{th} option of i^{th} item for an examinee with a mastery profile that matches that option Q-matrix vector. Notice that the ERUM extends the function of r_{ihk} to penalize not only the lack of mastery for the measured k^{th} attribute, but also for possibly mastering the k^{th} attribute (e.g., specific misconception). In other words, the r_{ihk} of the ERUM indicates the reduction of probability for choosing the h^{th} option of the i^{th} item if an examinee's k^{th} attribute does not match the specified attribute measured by the option. This matching is calculated as the $|q_{ihk} - \alpha_k|$, which only equals 0 when the two are identical and 1 otherwise. If the two values match, $r^0 = 1$ and thus no penalty is applied. Otherwise, the attractiveness is reduced by a factor of r, which

$0 < r < 1$. The probability of choosing an option is related to the value of π and r. A high π value indicates that an examinee whose attribute profile matched the measured Q-vector of an option will increase the attractiveness of that option, and low r value means that any mismatched attribute can dramatically reduce the attractiveness of that option.

2.3 The Impact of Distractors

Distractors can have a great impact on item functioning. Many past studies have focused on how to appropriately model distractors, examining the quality of distractors, types of distractors, and the quantity of distractors (Ali, Carr, & Ruit, 2016; Cizek & O'Day, 1994; Kubinger, Holocher-Ertl, Reif, Hohensinn, & Frebort, 2010; Pachai, DiBattista, & Kim, 2010; Sideridis, Tsaousis, & Al Harbi, 2017). Informative distractors were found to be not ignorable, and models that explicitly account for distractors would provide a better fit (Sideridis et al., 2017). Sideridis and colleagues (2017) examined several ways of modeling informative distractors. The methods used in the study included (1) adapting both a Rasch and partial credit models to model the informative distractors within items; (2) modeling the informative distractors as separate items; and (3) modeling the items that were combined with a testlet model. The study also examined the items using a Rasch model without including the low-ability group, which could have a higher chance guessing than the high-ability group. The results show any method other than treating the informative distractors as separate items can provide a better fit when compared the use of standard Rasch model. However, given the different data used in the methods compared, model fit cannot be fairly judged.

Studies have found that adding non-informative distractors does not affect the multiple-choice responses. However, using informative distractors can increase the reliability of a test and the difficulty of an item to the level of a free-response format. Cizek and O'Day (1994) studied non-functioning item options, which are options rarely chosen by examinees. They compared a test that contained five-option items, each of which had a non-functioning option, to a test that contained four-option items that did not include the non-functioning option. They found that the item's discrimination and difficulty parameters, when calibrated using a 2PL model, were not significantly different. Additionally, the reliability of the two tests was not significantly different. Ali et al. (2016) replaced non-functioning distractors of multiple-choice items with examinees' partial answers from a free-response format of the same questions, which is a way of increasing the information of distractor. This new form was compared to the multiple-choice items with non-functioning distractors. The difficulty of the items increased, becoming closer to the difficulty observed in the free-response format. In this case, the internal consistency of the test also improved after the replacement.

Other studies have examined the impact of a "none of the above" (NOTA) option as a distractor or even a correct option. Pachai et al. (2010) found that using a NOTA option as a distractor can decrease the discrimination of an item because the NOTA option can attract both the higher-ability group and the lower-ability group. In addition, using NOTA as a correct option can increase the item difficulty, which does not help improve the quality of items. Similarly, Kubinger et al. (2010) found that a five-option multiple-choice item that requires an examinee to select two options to obtain credit tends

to be more difficult than a six-option multiple-choice item that requires only one correct answer. Additionally, there was no difference in item difficulty between a free-response format item and the five-option item. The results either make distractors more informative or increase the number of options, which can reduce the chance of guessing and thus lead to an increase in item difficulty.

In general, it does appear that adding informative distractors can improve reliability and even limit the chances of guessing. Increases in test reliability will lead to decreases in measurement error, which is an indication of measured-ability improvement. Studies have shown that distractors can provide useful information for more accurately estimating an examinee's ability level. The next section addresses research related to how distractors impact ability estimation.

2.4 Abilities Estimation Distractors

Given an assessment with multiple-choice items, there are at least two approaches for scoring: binary scoring and polytomous scoring (i.e., using an ordinal or multinomial model). Binary scoring approaches only consider whether the examinee correctly answers the item. As a result, how the examinee misses the item (which distractor is chosen) is ignored. Modeling approaches that use polytomous scoring treat each incorrect option as a piece of information. The incorrect options can be treated as ordered (Jiao et al., 2012) or nominal (Bradshaw & Templin, 2014) data. Although the suitable model will be selected based on the scoring approach characteristics, polytomous models tend to be more complex than dichotomous-response models. As a result, the question of whether information obtained from a polytomous scoring approach can be justified by the

increased complexity of a polytomous model should be explored. That is, does scoring examinees using polytomously scored multiple-choice items meaningfully improve estimates of an examinee's ability?

Several studies have focused on whether to score items polytomously when partial credit can be given to examinees with responses that are not entirely incorrect. Grunert and colleagues (2013) examined how the distribution of total scores on a chemistry test was affected when changing from dichotomous scoring to polytomous scoring and found that different scoring methods did result in changes in students' rank order. Although the relative order of students was impacted, the normality of the total score distribution was unchanged. The results of this study suggest that polytomous scoring could change a student's estimated ability level.

Jiao and colleagues (2012) examined an assessment with only dichotomously scored items and found that polytomous scoring of some items did not change the estimation of examinees' latent abilities. The authors scored some of the items polytomously and compared the results to estimates obtained using only dichotomous-scored items. The results showed that the ability scores from the two methods were highly correlated (r = .94). Additionally, simulation studies were run to explore the difference between scoring examinees using polytomously scored items or dichotomously scored items. Data were simulated from both the PCM and Rasch models using parameters from a real data analysis. The approaches were then used to conduct the anlaysis in the following way. The two types of originally simulated data were analyzed using the PCM and Rasch, respectively. The simulated polytomous responses were then

dichotomized, and the dichotomous Rasch model was used to analyze the data simulated using Rasch and dichotomized data. Simulation results showed examinee abilities estimated from the first approach had a lower standard error when compared to the abilities estimated using the second approach. However, because some of the data were generated from the PCM, the higher accuracy of latent-ability estimates using the PCM may have been due to a model misfit.

A comparison of polytomous versus dichotomous models has also been conducted using the DCM framework. For example, de la Torre (2009a) compared examinee estimates obtained using polytomously scored items that were analyzed using the MC-DINA to examinee estimates obtained using the same items when dichotomously scored and analyzed using the DINA. In this study, the data were simulated using the MC-DINA. A 30-item, 4-option, 5-attribute test was used in this scenario. The first set of 10 items only had one attribute measured by the correct options, and no attributes were measured by the incorrect options. The second set of 10 items measured two attributes, and the third set of 10 items measured three attributes. The last two sets had attributes measured by some or all of the incorrect options. de la Torre (2009a) suggested that the first ten items should produce the same results if estimated by either the MC-DINA or the DINA model because the distractors provided no information about the examinee's latent attributes. The data were analyzed first using the MC-DINA and compared to data analysis using the DINA with dichotomized data. The correct option Q-matrix was used in the DINA model. The results showed that the MC-DINA provides better classification accuracy of latent attributes than the DINA model. This study had a similar problem to

the previous study(Jiao et al., 2012); that is, the advantage of the MC-DINA in classification accuracy could be due to a misfit of the DINA model.

The MC-DINA is a multiple-choice extension of the DINA model and can be used to diagnose examinees' attribute profiles that match with attribute profiles measured by item options. The model can be limited in application because it only considers the attribute profiles matched with profiles measured by options, and measured attributes must be non-compensatory. On the other hand, the GDCM-MC is a more flexible model, as it is not limited to profiles only measured by the options of items (ERUM) and non-compensatory attributes.

Although polytomous DCMs have been developed, the same items could be scored dichotomously and fit with a much simpler model in most cases. Thus, there is a need to specifically address the question of whether the added complexity of a polytomous model is justified. Further research should examine if polytomous modeling approaches provide better estimates of examinees' attributes.

In the current research, the focus is specifically on model recovery and, more importantly, on examinees' classifications (i.e., estimation of examinees' abilities). In contrast, comparing model fit between polytomous and dichotomous models is difficult because typical measures of relative fits, such as AIC or BIC, are based on different data. Other indices that show polytomous and dichotomous DCM performance may also serve as an alternative comparison. For instance, the cognitive diagnostic index (*CDI*; Henson &Douglas, 2005), based on Kullback-Leibler information, can be used to understand how well an item discriminates between high- ability and low-ability groups. Because the *CDI*

is based on the Kullbach-Leibler distance between attribute patterns, which is defined for both dichotomously and polytomously scored items, the index can be computed for both polytomous and dichotomous data. Also, as both the polytomous and dichotomous forms of *CDI* rely on model parameters, it may be an indirect tool with which to compare the performance of polytomous and dichotomous scoring models.

2.5 *CDI*

Henson and Douglas (2005) introduced the *CDI* for dichotomous DCMs. The *CDI* measures an item's overall discrimination power between attribute-mastery profiles. It can indicate an item's usefulness in examinee-profile estimation (Henson, Rousso, Douglas & He, 2008). Henson et al. (2008) have shown that the attribute-level *CDI* is positively associated with CCRs, which means that the higher the *CDI*, the better the item performs. One advantage of the *CDI* is that it provides a unified approach to measuring the value of an item that can be computed regardless of the model used. Specifically, the *CDI* is expressed as a function of the Kullbach-Leibler information (KLI), which can be expressed as a function of the conditional-probability distribution of the item given the attribute profile, as opposed to only relying on differences of specific item parameters. In addition, the *CDI* naturally extends to polytomous models. (Henson, DiBello & Stout, 2015). For these reasons, this study will use the index as a measure of item quality and as one method to quantify the improvement of an item when considering polytomous responses as opposed to dichotomously scored responses. Because the *CDI* depends on the KLI, a brief discussion is given, and the *CDI* is then defined.

The KL information (KLI) formula of the j^{th} item is defined as:

$$KLI_i(u,v) = \sum_{h=1}^{H_i}\left[P(X_i = h|\boldsymbol{\alpha}_u) * \ln(\frac{P(X_i = h|\boldsymbol{\alpha}_u)}{P(X_i = h|\boldsymbol{\alpha}_v)})\right] , \quad (30)$$

where H_i is the number of options for the i^{th} item (for dichotomous responses there are only two values of *h*), and *h* is the specific option of the item.

$KLI_i(u,v)$ is the weighted sum of the logarithm difference between the option-response probability conditional on the facet pattern u and v across H_i option(s). The weight is the probability of choosing the h^{th} option of the i^{th} item conditional on the attribute pattern u. Note that there should be only two conditional option response probabilities, $P(X_i = 1)$ and $P(X_i = 0)$, for the KLI used in the dichotomous. Because $KLI_i(u,v)$ is defined for all possible pairs of facet patterns, the CDI_i of the i^{th} item is weighted by all possible facet patterns,

$$CDI_i = \frac{1}{\sum_{u\neq v} h(\alpha_u,\alpha_v)^{-1}} \sum_{u\neq v}[h(\alpha_u,\alpha_v)^{-1} KLI_i(u,v)] , \qquad (31)$$

where $h(\alpha_u,\alpha_v)$ is the Hamming distance between two facet patterns, u and v.

CDI_i measures the discrimination of the i^{th} item. Test-level *CDI* ($CDI_\bullet$), which is the sum of CDI_i of all the items of a test, can also be used to indicate the quality of the entire test (Henson & Douglas, 2005; Henson et al., 2008). In this study, $CDI_\bullet$, which is the sum of the CDI_i of all items on a test as well as on CDI_i, will be used to examine the test-level discrimination,

$$CDI_\bullet = \sum_{i=1}^{I} CDI_i . \quad (32)$$

The previous research also showed that $CDI_{\bullet}$ could be positively non-linearly associated with classification accuracy (Henson et al., 2008). As a result, this research further extends this research to examine the relationship between the $CDI_{\bullet}$ and CCRs when comparing dichotomous and polytomously scored assessments.

2.6 Research Aims

Research Aim 1. The first aim is to develop a polytomous model such that if the polytomous data fits this model, it must also be true that the dichotomized data (e.g., scored right/wrong) will simultaneously fit a dichotomous DCM.

Research Aim 2: The second aim is to demonstrate the advantage of modeling the polytomously scored items as opposed to only using the dichotomously scored information. Furthermore, the aim is to explore the effect of varying levels of information provided by the polytomous data by varying the information in the distractors (options that are incorrect). Thus, three conditions will be considered: (a) when there are no informative distractors measured in the incorrect options, the polytomous model should have the same correct classification rates (CCRs) as when using the dichotomously scored items; (b) when the distractors are somewhat informative as the correct option, the polytomous scored items should have higher CCRs than their dichotomous version; and (c) when the distractors are as informative as the correct option, the polytomously scored items should result in higher CCRs than the somewhat-informative distractors and their dichotomous version. The results should be generalizable across different conditions such as sample size, test length, the number of facets, and item quality.

Research Aim 3: To show that $CDI_{\bullet}$ will behave similarly as CCRs across different sample sizes, test lengths, attribute numbers, and quality of distractors and items. The log of $CDI_{\bullet}$ will be positively associated with CCRs. Because the true value of CCRs is unknown in educational assessments, the value of $CDI_{\bullet}$ in this study can, therefore, be used to indicate the quality of an assessment.

CHAPTER III

METHOD

In this chapter, the EGCA model is first introduced as a submodel of the GDCM-MC. Thus, the EGCA is a polytomous DCM. In addition, this model is developed in such a way that if the polytomous responses are modeled and fit the EGCA, then it must also be true that, after being dichotomized, the same responses will fit the RRUM. Thus, the model will allow for study of the informative distractor benefit without being confounded by model fit. A simulation study and a real data study plan are then described in addition to defining the indices used for analyzing the results.

3.1 Excluding Guessing from Correct Answer Model

As defined in the literature review, the GDCM-MC defines the probability of selecting an option as a weighted combination of a cognitive portion and a random guessing portion. Although conceptually, such a model can be useful, when studying the potential benefit of using polytomously scored items, the GDCM-MC has a limitation. Specifically, if data are simulated from the GDCM-MC, then the rescored dichotomous items will not follow any known DCM. As a result, any comparison with respect to CCRs between the two models used to score items will be confounded by model misfit in addition to any information that may be obtained by the polytomous data.

The EGCA GDCM-MC is a submodel of the GDCM-MC and defined by excluding the guessing portion of the GDCM from the correct option. Note that guessing

is still functionally the same as the GDCM-MC for all other incorrect options. Thus, the probability of an examinee choosing any incorrect option is modeled as a function of both the cognitive portion of the GDCM-MC and random guessing for the option (see equation 33). If there is no information (attributes) in the incorrect option, the probability of choosing the option will be reduced to the probability of selecting that option by random guessing. The option with no diagnostic information is referred to as "unlinked." In contrast, an option with diagnostic information is "linked." That is, in a linked option, mastery or non-mastery of a set of attributes do directly influence the attractiveness of that option as opposed to only random guessing.

The EGCA uses the same Q-matrix as the GDCM-MC and has three constraints. The first constraint requires that the correct option be entirely modeled using only the cognitive portion of the GDCM-MC. That is, guessing is excluded from the calculation of the correct option (i.e., the probability of selecting the correct option is completely modeled by the kernel function). The second constraint requires the kernel function, $F_{ih}(\boldsymbol{\alpha})$, of an unlinked incorrect option(s) to be equal to 0. The third constraint is accomplished by requiring that $S\alpha < 1$ in the original GDCM-MC. Recall that when $S\alpha < 1$ it must also be true that $\omega i\alpha < 1$, which results in at least some guessing being modeling for the probability of selecting an incorrect response. Given these three constraints of the GDCM-MC, the EGCA defines the probability as:

$$P_i(h|\boldsymbol{\alpha}) = \begin{cases} \frac{F_{ih}(\boldsymbol{\alpha})}{S_\alpha}\omega_{i\boldsymbol{\alpha}}, \text{when } h = \text{correct option} \\ \frac{F_{ih}(\boldsymbol{\alpha})}{S_\alpha}\omega_{i\boldsymbol{\alpha}} + \frac{1}{H_i'}(1-\omega_{i\boldsymbol{\alpha}}), \text{when } h = \text{incorrect option linked} \\ \frac{1}{H_i'}(1-\omega_{i\boldsymbol{\alpha}}), \text{when } h = \text{incorrect option unlinked} \end{cases}, \quad (33)$$

$$\omega_{i\boldsymbol{\alpha}} = \min\{1, S_\alpha\},$$

$$S_\alpha = \sum_{h=1}^{Hi} F_{ih}(\boldsymbol{\alpha}),$$

where $H_i' = H_i - 1$ because the correct option is excluded from guessing. Additionally, the guessing portions are required in the incorrect options. Thus, $\omega_{i\boldsymbol{\alpha}}$ has to be 1, which means that S_α has to be smaller than 1 in order to make $\omega_{i\boldsymbol{\alpha}}$ is 1. Equation 33 can therefore be simplified as:

$$P_i(h|\boldsymbol{\alpha}) = \begin{cases} F_{ih}(\boldsymbol{\alpha}), \text{when h} = \text{correct option} \\ F_{ih}(\boldsymbol{\alpha}) + \frac{1}{H_i'}(1-S_\alpha), \text{when h} = \text{incorrect option linked} \\ \frac{1}{H_i'}(1-S_\alpha), \text{when h} = \text{incorrect option unlinked} \end{cases}. \quad (34)$$

Note that by defining the EGCA this way, the correct option-response function is always the same as if the data were recoded as correct and incorrect (i.e., dichotomously scored) and then parameterized using the dichotomous model, kernel, $F_{ih}(\boldsymbol{\alpha})$. If there is no additional information in the incorrect options, $F_{ih}(\boldsymbol{\alpha})$ will become 0 for those incorrect options, which means that the EGCA is the same as $F_{ih}(\boldsymbol{\alpha})$ of the correct option. If there is additional information in the distractors (i.e., $F_{ih}(\boldsymbol{\alpha}) \neq 0$), then the EGCA directly models the diagnostic information above and beyond the correct response option modeled by $F_{ih}(\boldsymbol{\alpha})$. Notably, if none of the incorrect options provide information,

the EGCA will be equivalent to the dichotomous $F_{ih}(\boldsymbol{\alpha})$ because: (a) the correct option of the EGCA with $F_{ih}(\boldsymbol{\alpha})$ and the RRUM has the same response function; and (b) there are no additional parameters.

Finally, the EGCA, just as was the case in the GDCM-MC, must have the kernel function defined for calibration or application. In this research, the kernel, $F_{ih}(\alpha)$, used is the RRUM. Thus, in each option, there will be a π, and r is as many as number of attributes as measured by the option.

3.2 Simulation Study

Given the EGCA with the ERUM kernel, the current study aims to compare the EGCA with the dichotomous RRUM through both a simulation study and a real-world data example. Recall that the primary focus of this research is to study the effect of using dichotomously scored items versus polytomously scored items on correct classifications rates (i.e., the estimation of examinees' attributes). In doing so, factors of the simulation study must be considered such that the results will be useful and generalizable to a fairly broad set of application. Thus, the simulation will consider factors that have been shown to impact the quality of examinee estimates.

3.2.1 Research Factors

Previous research has shown that factors such as sample size, number of items, number of attributes, and the value of item parameters can impact correct classification rates (CCRs) in the dichotomous DCMs (Bradshaw & Templin, 2014; Fu, Rollins, & Henson, 2016; Shu et al., 2013). Specifically, the literature suggests that increasing the sample size or the test length can lead to higher CCRs. Tests that measure a large number

of attributes tend to have lower CCRs compared to tests that measure fewer attributes given the same test conditions (Bradshaw & Templin, 2014; Fu et al. 2016). In addition, item quality can impact CCRs, as such quality is often referred to as item discrimination and can usually be expressed as a function of the item parameters. For example, item quality is associated with the magnitude of the intercept and the slopes of latent attributes in addition to any interaction effect when using the LCDM (Bradshaw & Templin, 2014). Most relevant to this research, it has been shown that item quality is directly related to the combination of π and r parameters when using the RRUM; specifically, a high π and low r condition indicate a high-item quality (Fu, Rollins, & Henson, 2016). The findings show that CCRs are directly impacted by the abovementioned factors researched in the dichotomous DCM framework. The factors that impact can also impact CCRs of the GDCM-MC ERUM (DiBello et al., 2015; Naumenko, Fu, Henson, Stout, & DiBello, 2016). Moreover, previous research has shown that the informativeness of distractors may impact the latent-ability estimation of the latent abilities (Jiao et al., 2012). It is believed that informativeness of distractors can also impact CCRs using EGCA. These factors were manipulated for this study. The following paragraphs will describe and detail the different levels considered for each of these factors. In addition, multiple replications of the same condition were considered for the simulation study.

3.2.1.1 Manipulation of Item Quality

The values of π and r were manipulated for the correct option and linked incorrect options. The π values are set to 0 for unlinked incorrect options and, as a result, the r values do not matter because they would be multiplied by $\pi = 0$ and are

constrained to 1. A good quality item for the RRUM is when π is high and r is low (Fu et al., 2016). In a previous study (DiBello et al., 2015), the π and *r* values were drawn from *Unif* (0.65, 0.95) and *Unif* (0.1, 0.5) respectively. The study further distinguished the values of π and *r* for each linked option. High and low levels of the π values were drawn from *Unif* (0.7, 0.9) and *Unif* (0.5, 0.7), respectively. High and low levels of the r parameters were drawn from *Unif* (0.2, 0.5) and *Unif* (0.05, 0.2), respectively. The two conditions of π were crossed with the two conditions of r to create a total of four conditions of item quality. These four conditions are High/Low (H/L, good quality), High/High (H/H, medium quality), Low/Low (L/L, medium quality) and Low/High (L/H, poor quality).

The proposed π and r can further influence the ERUM kernel $F_{ih}(\boldsymbol{\alpha})$, which can in turn affect the value of S_α. The EGCA requires $S_\alpha < 1$ to make the kernel $F_{ih}(\boldsymbol{\alpha})$ of the correct option exactly the same as its analogous dichotomous DCM. If the proposed π and *r* lead to $S_\alpha > 1$, the parameters estimated EGCA would not be correct. The simulated value of π of all the options were divided from the maximum value of S_α of the entire attribute profiles for that item so that $S_\alpha < 1$.

3.2.1.2 Manipulation of Distractors

The current study also manipulated the informativeness of the distractors simulated using the EGCA. There were three levels of the informativeness of the distractors: strong, weak, and no information (none). The amount of information for any distractors was manipulated by changing the π and r. Specifically, the values of π of the

strong (good quality) distractors were in the same range of the values of π of the correct option. For example, if the values of π of the correct options were drawn from a *Unif* (0.5, 0.7) distribution, then the values of π of the distractors were also drawn from a *Unif* (0.5, 0.7) distractor. To define the weak distractor conditions, the π values were simulated to be lower than the π for the correct option. Specifically, if the values of π for the correct options were drawn from a *Unif* (0.5, 0.7) distribution, the values of π of distractors would first be first drawn from *Unif* (0.5, 0.7), and then divided by 2 to resemble a bad distractor that could not distinguish the attributes.

3.2.1.3 Simulation of Attributes

Using the following steps, the attribute mastery profile was simulated. First, random numbers were simulated from a multivariate normal distribution with dimensions equal to the number of measured attributes in which the first half are skills, and the second half are misconceptions. The means and a variances-covariances matrix of the multivariate distribution are $\mathbf{0}$ and $\mathbf{\Sigma}$ respectively. The diagonal of $\mathbf{\Sigma}$ is all 1s, which is the variance of the multivariate distribution, and the off-diagonal is covariance (or correlation because the variances are 1s.) between skills, misconceptions and skills, and misconceptions. The skills were assumed to be correlated with a magnitude of *Unif* (.2, .4), misconceptions were assumed to be correlated with a magnitude of *Unif* (.3, .5), and the correlation between misconception and skills were assumed to be negative with a value of *Unif* (-.5, -.7). In addition to the association between attributes, it is believed that not all attributes would be equally difficulty. Thus, the proportion of mastery on each

attribute was simulated from a *Unif* (.4, .6) distribution for each condition replication. Using these proportions, the cutoff point for each attribute was computed based on the z-scores corresponding to the generated mastery probability, as each dimension followed a univariate normal distribution. For instance, the z-score that corresponds to the mastery probability .5 is 0. If the generated random number is larger than 0, the attribute is mastered (1); otherwise, the attribute is not mastered (0).

3.2.1.4 Q-Matrix Generation

In this study, four and six attributes Q-matrices were used. The four-attribute condition included two skills and two misconceptions measured by the whole test, and the six-attribute condition included three skills and three misconceptions. After determining the number of attributes and whether they are skills or misconceptions, the Q-matrix was randomly generated. Different Q-matrices were generated for each replication and condition. In order to make the Q-matrices realistic, the correct option required the skills measured by the item to be present and required the misconceptions measured by the item to be lacking or absent, and there were no more than three attributes measured by an item. The incorrect options required that not all skills measured by the item must be mastered or that the misconceptions measured by the item must be mastered. These constraints are similar to the "realistic" constraints used by DiBello et al. (2015). By using these constraints, the Q-matrices resembled the multiple-choice items in educational assessments. An example of a four-attributes Q-matrix of a GDCM-MC is shown in Table 4. In this example, which was also the case in the current simulation study, the correct option is the first option for every item.

3.2.1.5 Recoding Data

Given the Q-matrix and the specified conditions, data were simulated from the EGCA with the ERUM kernel by using the polytomous Q-matrix and then estimated using the EGCA. Next, to calibrate the data set using the RRUM, the data needed to be rescored as right/wrong (0/1 dichotomous scoring). Because the data was intentionally simulated such that the first item was "correct," the data was rescored such that if the first option was scored as a 1 ("correct"), and all other options were scored as a 0 ("incorrect").

3.2.1.6 Recoding Q-Matrix

To calibrate the RRUM, a Q-matrix for the dichotomous model must also be defined. There are two challenges in redefining a Q-matrix for the RRUM when originally simulating data from the EGCA. First, the EGCA defines a Q-matrix entry for every option, whereas the RRUM only defines a Q-matrix for each item. Second, the EGCA allows three entries in the Q-matrix (0, 1, and N), whereas the RRUM only allows for two entries (0 and 1). The following discussion describes the method used to recode the Q-matrix for the RRUM from the EGCA.

The EGCA is defined in such a way that the probability of selecting the correct option is *always* defined as $F_{ih}(\boldsymbol{\alpha})$. Thus, the Q-vectors of correct options of all items for the EGCA are only needed for the RRUM. However, in addition to ignoring the Q-matrix entries for the distractors, one additional change must be made. Specifically, the EGCA allows for three types of entries (i.e., 0, 1, N) whereas the RRUM allows for two types of

entries (i.e., 1, 0) and thus the number of entry types for the EGCA Q-matrix must be changed.

Using the EGCA, the Q-matrix entry of 1 and 0 mean that the attribute directly influences attractiveness and N means it does not. In the RRUM, the Q-matrix entry of 1 means that the attribute influences the attractiveness and 0 means it does not. Therefore, in order to use the correct option Q-vector for the EGCA, all Ns are coded as 0s (because 0 in the RRUM matrix means that it does not influence), and all 0s and 1s are coded as 1 in the RRUM (because they influence the attractiveness). The correct option Q-vectors only have entries 0s and Ns for misconceptions. As such, Q-vector entries 0s are coded as 1s, which means that the idea of misconceptions need to be reexpressed as skills, and those skills would be "Does not possess that misconception". Thus, the RRUM uses the correct option Q-vectors with misconceptions entries 0s recoded to 1s and Ns recoded to 0s. In order to make the classification comparable between the EGCA and the RRUM, the estimated misconceptions are recoded back from 1s to 0s and 0s to 1s after the attributes are estimated by the RRUM.

Tables 4 and 5 provide an example of a possible Q-matrix for the EGCA (Table 4) and how it is recoded for the RRUM (Table 5). Note that the correct option in the Q-matrix for the EGCA is the first option. For example, the entry for item 1 option 1 of A1 (skill) is 1 for the EGCA, which is the same in the RRUM Q-matrix. The entry for item 1 option 1 of A2 (skill) and A3 (misconception) are Ns for the EGCA, and it is coded as 0s in the Q-matrix for the RRUM. Lastly, the entry for item 1 option 1 of A4 (misconception) is 0 for the EGCA, and it is coded as 1 in the Q-matrix for the RRUM.

Table 4. Four-Attribute Q-Matrix with One Unlinked Option for the EGCA

Item No.	Option No.	Skills		Misconceptions	
		A1	A2	A3	A4
1	1	1	N	N	0
1	2	0	N	N	N
1	3	N	N	N	1
1	4	N	N	N	N
2	1	N	1	0	N
2	2	N	0	N	N
2	3	N	N	1	N
2	4	N	N	N	N

Table 5. Recoded Q-Matrix for the RRUM Based on the EGCA Q-Matrix in Table 4

Item No.	Skills		Misconceptions	
	A1	A2	A3	A4
1	1	0	0	1
2	0	1	1	0

3.2.1.7 Estimation Algorithm

The Metropolis-Hastings (MH) within the Gibbs Sampling Markov Chain Monte Carlo (MCMC) estimation algorithm was programmed in FORTRAN to estimate both the RRUM and the EGCA. This program is a modified version of the FORTRAN program originally developed for estimation of the GDCM-MC (DiBello, Stout, & Henson, 2015). The chain length for all MCMC estimations was 5,000 with a 4,000 burn-in, and two chains with random starting values were used.

3.2.1.8 Summary of Simulation Conditions

In the simulation study, two other factors were manipulated. Specifically, the number of simulees (i.e., 1,000 and 2,000) and the number of items (i.e., 20 and 40). The factorial design of this study contained two levels of sample size, two levels of test

lengths, two levels of attribute sizes, four levels of π and r conditions (i.e., H/H, H/L, L/H, L/L), three levels of distractor quality (i.e., strong, weak and none), and two different estimating models (i.e., EGCA and RRUM) resulting in 2 x 2 x 2 x 4 x 3 x 2 = 192 conditions in total (see Table 6). Finally, each condition was replicated 50 times. Given the large number of factors manipulated in this study, it was believed that the variability of each condition would not be large for each condition.

Table 6. Simulation Conditions

Factors	Conditions
Examinees	1000, 2000
Test Length	20, 40
Attributes	4 (2 skills +2 misconceptions), 6 (3 skills + 3 misconceptions)
$\boldsymbol{\pi}$	High (*Unif* (0.7, 0.9)) Low (*Unif* (0.5, 0.7))
r	High (*Unif* (0.2, 0.5)) Low (*Unif* (0.05, 0.2))
Distractor Information	Strong, Weak, Non-informative(None)
Model	EGCA, RRUM

3.2.2 Indices

3.2.2.1 PC

To ensure cross-chain convergence using the Gelman and Rubin R statistic (GRR; Gelman & Rubin, 1992), when the GRR statistics was between 1 and 1.3, the

corresponding parameters were treated as convergence. For each replication of a condition, the number of parameters that have the GRR greater than 1.3 is summed as x. The total number of parameters for that condition is X. The proportion convergence (PC) was used to examine the percentage of convergence. The formula is shown as follows:

$$PC = 1 - \frac{x}{X}. \quad (35)$$

3.2.2.2 CCRs

The profile- and attribute-level correct classification rates (pCCRs and aCCRs) were used to detect any differences in accuracy between the two models across various conditions. pCCRs are defined as the proportion of examinees that were correctly classified across all *K* attributes by the model. aCCRs reflect the proportion of correct classification for averaged attributes across all attributes and examinees. pCCRs and aCCRs are defined as,

$$pCCRs = \frac{\sum_j^N E(\boldsymbol{a}_j = \widehat{\boldsymbol{a}_j})}{N}, \quad (36)$$

$$aCCRs = \frac{\sum_j^N \sum_k^K E(a_{jk} = \hat{a}_{jk})}{N * K}, \quad (37)$$

where *N* is the total number of simulees, *K* is the number of attributes measured by an assessment, $\boldsymbol{a}_j$ is the attribute profile of the j^{th} simulee, and a_{jk} is the j^{th} simulee's k^{th} attribute. E is the expectation function. When the condition inside is met, E(.) is 1, otherwise, E(.) is 0.

3.2.2.3 $CDI_{\bullet}$

$CDI_{\bullet}$ was used to check how tests discriminate the attributes. The correlation between the log of $CDI_{\bullet}$ and pCCRs and aCCRs for the EGCA and the RRUM, which can indicate the strength of the relationship between the log of $CDI_{\bullet}$ and the two CCRs, was obtained, In addition, $CDI_{\bullet}$ across different conditions was obtained. This can be used as a guideline for determining the values of $CDI_{\bullet}$ that corresponds to the highest or lowest CCRs.

3.2.2.4 MAD and Correlation

The current study also compared the estimated and true parameters of the EGCA using the mean absolute difference (MAD) and correlation. The MAD is shown as:

$$MAD = \frac{\sum_r^R \sum_p^P |\hat{x}_{pr} - x_{pr}|}{R * P}, \quad (38)$$

where p is the p^{th} parameters, P is the total number of parameters, r is the r^{th} replication, R is the total number of replication $\hat{x}_{pr}$ is the estimated parameter, and x_{pr} is the true parameter. When comparing the two models' parameters, the MAD can be expressed as shown:

$$MAD = \frac{\sum_r^R \sum_p^P |\hat{x}_{pr} - \hat{x}'_{pr}|}{R * P}, \quad (39)$$

where $\hat{x}'_{pr}$ is the estimated parameter of another model.

The MAD and correlation between the estimated and true π and r were examined across all conditions for the EGCA. The EGCA is the model used to simulate the data,

and it has been previously defined to be mathematically equivalent to the RRUM when the incorrect options are non-informative. Furthermore, the parameters for the correct option should be equivalent when using both the polytomous data and the dichotomously scored data. The MAD and correlation between the parameters (i.e., π and r) of the EGCA and the RRUM were compared because the two models are expected to have the same parameters.

3.3 Real Data Study

A distractor-driven assessment from a previous study (Shear & Roussos, 2016) was used to compare the EGCA and the RRUM. The assessment contained 12 items: 5 were multiple-choice (MC) items, and the other 7 were selected-response (SR) items. A total of 2,011 examinees completed the assessment. The MC items have four options and require selecting one option, whereas the SR items have five to six binary choices and require the selection of binary choices in a certain way to correctly answer the items.

For the MC items, an examinee may select an incorrect option embedded with a misconception or other incorrect options that do not have a misconception. As for the SR items, answering a certain combination of binary choices may indicate the examinees have a misconception. Shear and Roussos (2016) found that not every option was chosen by the examinees in the MC items. A recoding method was used so that the 12 items only contained three options: one correct option and the two incorrect options (one only measures the misconception option, and the other is just a regular incorrect option). An illustration of the FR item and its coding method is illustrated in Figure 1. The assessment was intended to measure one skill and two misconceptions. Shear and

Roussos (2016) used relative fit indices Akaike information criterion (AIC) and Bayesian information criterion (BIC) to examine different Q-matrices. The Q-matrix that led to the best relative fit in the study by Shear and Roussos was used for the EGCA in this study (See Table 7).

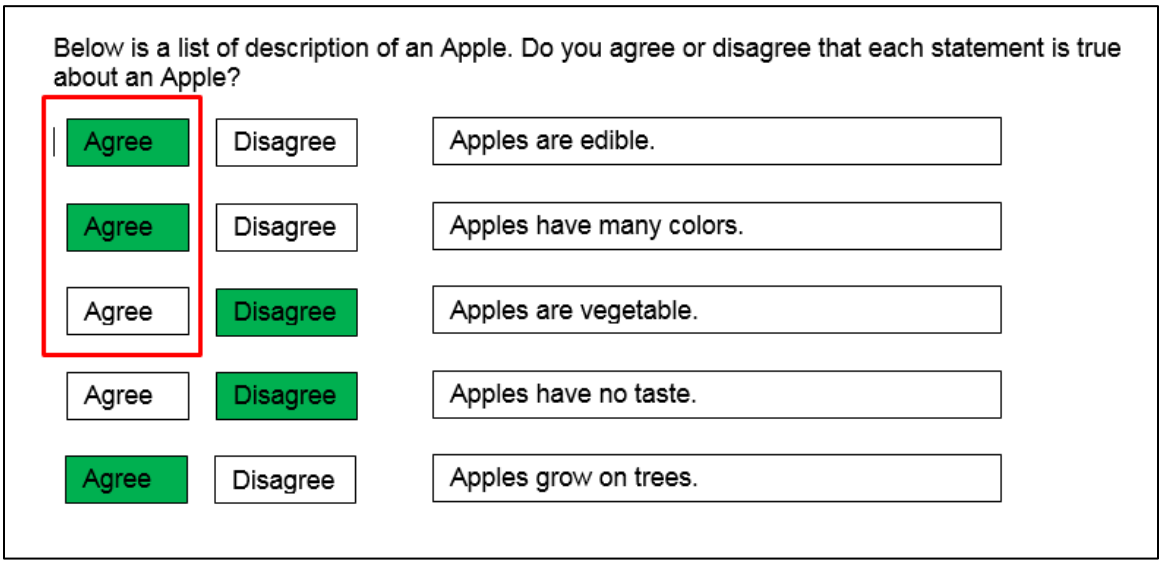

Figure 1. An Illustration of a Selected-Response Item.
Note. If all green options were selected, the answer is correct, meaning that the examinee selected the correct option. If the three options in the red rectangle were selected, it suggests that the examinee chose a misconception option. For other cases, it was considered that the examinees only selected incorrect options.

When using the RRUM model, the correct option Q-vectors of each item were used for the Q-matrix. In addition, when creating the Q-matrix for the RRUM, all misconceptions of the correct option were recoded to 1s, as previously described. Also, because the RRUM is a dichotomous model, the examinees' correct responses were recoded to 1s, and the other responses were recoded to 0. The recoded Q-matrix and responses were used in the RRUM. Recall that when using the RRUM in this way, the

misconceptions are coded as a "skill" that represents not having the misconception. Thus, to ensure that the estimation of the misconceptions is interpreted the same as between the EGCA and the RRUM estimated misconceptions, the RRUM estimated misconceptions were recoded from 1s to 0s and 0s to 1s. Furthermore, because the kernel in the EGCA was the same as in the RRUM, it was possible to use an algorithm that was close to the MCMC algorithm. To ensure convergence, three chains with 15,000 chain lengths and 10,000 burn-in options were used.

Because the true classification accuracy is unknown in this case, a number of indices were used to determine the differences between the EGCA and the RRUM. Specifically, comparisons were made between basic descriptions of the estimation of attribute patterns, statistics of the relationship between the response and attribute classification, and finally discrimination indices CDI_i and $CDI_{\bullet}$. First, the study examined the posterior and classification distributions of each attribute (i.e., the estimated proportion of mastery) between the EGCA and the RRUM. In order to evaluate the posterior distribution of each attribute for the two models, the absolute deviance (AD) between the estimated attribute probability and .5 for an attribute was summed. The sum of AD across all examinees can indicate the amount of discrimination for the attribute. The higher AD means the higher the discrimination of the classification. The sum of AD was compared between the two models across all three attributes for the two models. Similarly, to examine the attribute distribution, the sum of the mastery of each attribute was used and compared.

Secondly, descriptive statistics were used to compare the responses to the estimate attribute profile, which could illustrate whether the estimated attribute profile is associated with actual responses. In this particular example, each correct item option measures examinees who master the skill and lack the corresponding misconception (i.e., the correct option for items 1–5 measures misconception 1 and for items 6-12, misconception 2 is measured). The Q-matrix entry for the correct option is shown in Table 7. As shown, the examinees that match this pattern are expected to select this option much more frequently. To explore this relationship, a binary variable was used to recode the classification first. Specifically, if an examinee was diagnosed to have mastered Skill 1 and lacked Misconception 1, it was counted as 1 for that person and otherwise as 0. The binary variable was correlated with the number of correct options selected for the first five items because these items only measured Skill 1 and Misconception 1. Similarly, another binary variable was used to indicate whether an examinee had Skill 1 and lacked Misconception 2, which was correlated the number correct options selected for the last seven items.

Additionally, each misconception option of the first five items and the last seven items measured two separate misconceptions (see Table 7). The number of misconception options that measured a misconception selected by each examinee was correlated with the corresponding diagnosis of the corresponding misconception. In other words, whether the examinee had Misconception 1 or 2 was correlated with the number of misconception options selected by the examinee for items 1–5 or items 6–12. The abovementioned correlations were compared between the EGCA and the RRUM. A strong correlation

indicates the connection between the diagnosis and the selection of corresponding options. A weak correlation indicates the lack of connection between the diagnosis and the selection of corresponding options. Lastly, $CDI_{\bullet}$ and CDI_i for each item were estimated using the EGCA and the RRUM. The values were compared. The larger value indicate a higher test discrimination or item discrimination.

Table 7. The Q-Matrix for a Distractor-Driven Assessment

Item No.	**Option No.**	**A1(Skill 1)**	**A2(Misc. 1)**	**A3(Misc. 2)**
Item 1-5	Incorrect	0	N	N
	Misconception	N	1	N
	Correct	1	0	N
Item 6-12	Incorrect	0	N	N
	Misconception	N	N	1
	Correct	1	N	0

CHAPTER IV

RESULTS

In the current study, a submodel of the GDCM-MC, the EGCA, was first defined so that the model was equivalent to the analogous DCM when no distractor information exists. Given this model, a simulation study was used to study the added benefit of modeling the information distractors across various conditions (e.g., sample size, quality of item). Finally, a real-world dataset was analyzed using the two models.

In this chapter, the results of the simulation study and real data analysis were presented and discussed. The simulation study was completed specifically to study the effect of modeling on the classification of examinee's attribute profile. First, the convergence of item parameters for both EGCA and the RRUM were examined to ensure that the results can be reliably interpreted. The recovery of item parameters of the EGCA was then examined to ensure that the sub-model function well. Next, because it was assumed that the parameters of RRUM would consist of the same as the parameters of the correct option of the EGCA, the parameters of the correct option of the EGCA and the parameters of the RRUM were compared. Finally, the classification accuracy (i.e., pCCRs and aCCRs) and $CDI_{\bullet}$, were compared between the two models across various conditions. In the real data study, the classification and $CDI_{\bullet}$ between the two models were compared. It was expected that the two models would have different classification,

and the EGCA would provide a more polarized classification than the RRUM. The association between the classification and the selection of the correct options or misconception options were analyzed. It was expected that the correlation for the EGCA would be higher than the RRUM because the classification using the EGCA is more realistic than the RRUM.

4.1 Simulation Study

In this section, the item parameters convergence between the two models is first discussed, then followed by the MAD and correlation between the true and estimated item parameters for the EGCA model. Next, the MAD and correlation between the correct option parameters of the EGCA and the parameters of the RRUM were compared, and they were expected to be equivalent. Finally, pCCRs, aCCRs, and $CDI_{\bullet}$ between the EGCA and the RRUM across different conditions are explored in this section.

4.1.1 Item Parameters Convergence

As discussed in the Methods section, each replication convergence is evaluated using the GRR, which is computed for each item parameter. Thus, the proportion item parameters that would be considered converged (using the GRR) are computed for each replication within a condition. The median proportion of convergence (PC) across the 50 replications of each condition is shown in Figure 2 and Table 8. In general, the median PC was above .9 for both models for all conditions, with most of the conditions close to 1. In cases in which the median PC was low (e.g., .9), the boxplot showed many outliers that did not converge well; thus, longer MCMC chains may be needed for such cases.

Also, when the test length and sample size increased, the median PC increased, and when the number of attributes measured by a test increased the mean PC decreased.

The convergence of non-informative distractors condition is similar to the other distractor conditions for four attribute conditions. However, the non-informative distractor conditions had more outliers that did not converge as well as the strongly and weakly informative distractor conditions when there were six attributes on a test for both the EGCA and the RRUM (see Figure 2). This could be due to the fact that the non-informative distractors did not have the least information to estimate the large attributes, and therefore, was more likely to have a convergence problem. The results suggest, however, that longer chains should be used for such outlier cases. Moreover, although the median PC was not affected by the item quality in general when the test measured six attributes, the number of outlier cases for the non-informative distractor condition was the highest for the high π and low r condition, followed the low π and low r condition, which could affect the interpretation of the results (e.g., CCRs and $CDI_{\bullet}$) for these cases.

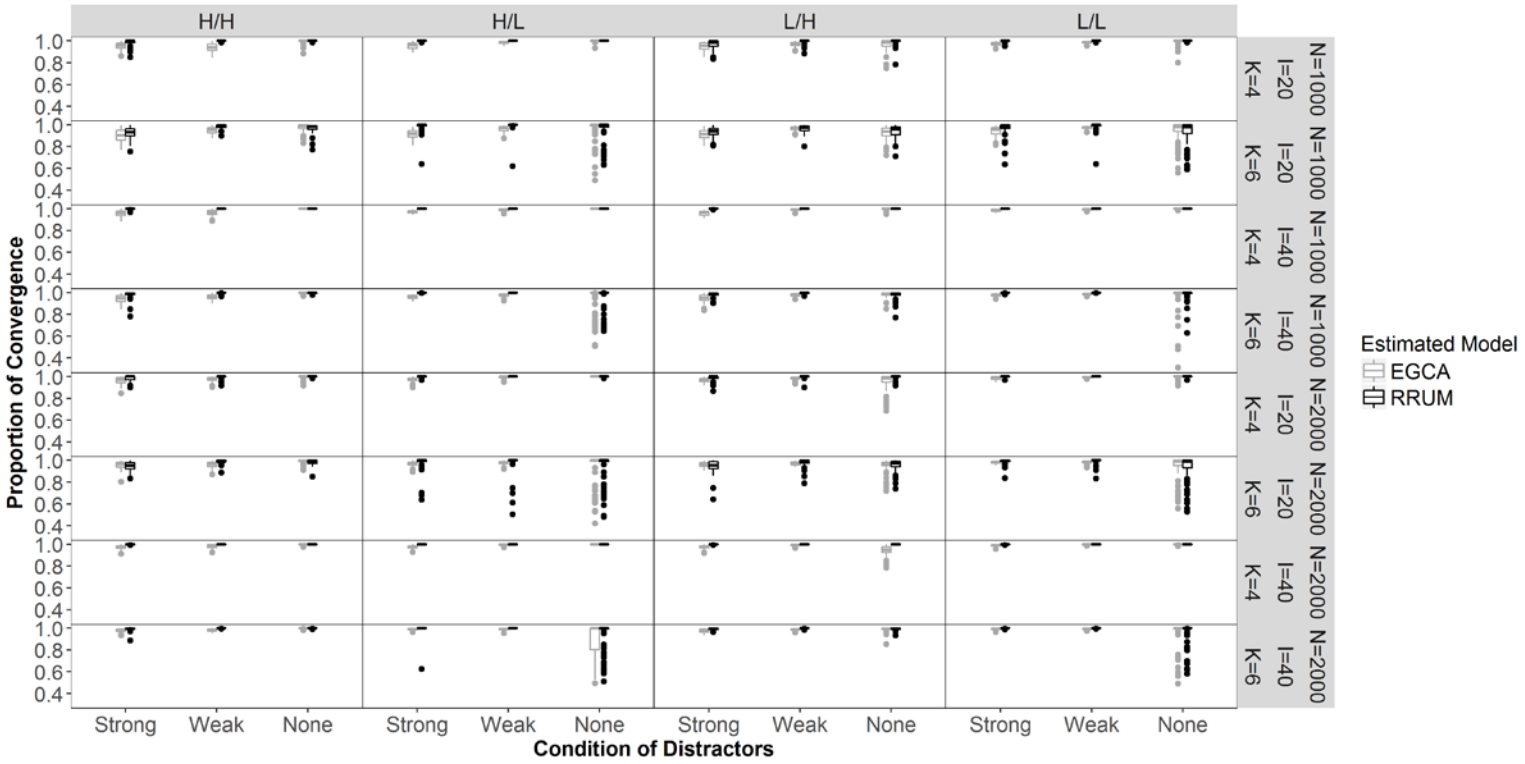

Figure 2. Proportion of Convergence Across All Conditions.
The proportion of convergence (PC) across different conditions is shown above. Each section of the graph represents three distractor conditions, and each color represents an estimated model. The columns represent the quality of the item while the rows represent a combination of sample sizes, attribute sizes, and test lengths.

Table 8. Median Proportion of Convergence Across All Conditions

N	*I*	*K*	*DQ*	H/H		H/L		L/H		L/L	
				RRUM	**EGCA**	**RRUM**	**EGCA**	**RRUM**	**EGCA**	**RRUM**	**EGCA**
1000	**20**	**4**	**None**	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00
			Weak	1.00	0.94	1.00	0.99	1.00	0.98	1.00	0.99
			Strong	1.00	0.96	1.00	0.96	0.98	0.95	1.00	0.97
		6	**None**	0.99	0.99	1.00	1.00	0.96	0.94	0.98	0.98
			Weak	0.99	0.95	1.00	0.97	0.97	0.96	1.00	0.98
			Strong	0.93	0.90	1.00	0.92	0.94	0.91	0.99	0.95
	40	**4**	**None**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			Weak	1.00	0.97	1.00	0.99	1.00	0.99	1.00	0.99
			Strong	1.00	0.96	1.00	0.97	1.00	0.96	1.00	0.99
		6	**None**	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00
			Weak	1.00	0.96	1.00	0.98	1.00	0.98	1.00	0.99
			Strong	0.99	0.95	1.00	0.96	0.99	0.95	1.00	0.98
2000	**20**	**4**	**None**	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00
			Weak	1.00	0.98	1.00	1.00	1.00	0.98	1.00	0.99
			Strong	1.00	0.97	1.00	0.97	1.00	0.97	1.00	0.99
		6	**None**	0.99	1.00	1.00	1.00	0.97	0.97	0.99	0.99
			Weak	0.99	0.96	1.00	0.98	0.99	0.97	1.00	0.99
			Strong	0.95	0.96	1.00	0.97	0.96	0.96	0.99	0.98
	40	**4**	**None**	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00
			Weak	1.00	0.98	1.00	1.00	1.00	0.99	1.00	1.00
			Strong	1.00	0.97	1.00	0.98	1.00	0.98	1.00	0.99
		6	**None**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			Weak	1.00	0.98	1.00	0.99	1.00	0.99	1.00	0.99
			Strong	0.99	0.98	1.00	0.99	0.99	0.98	1.00	0.99

Note: *N* is the sample size, *I* is the test length, *K* is the number of attributes, and *DQ* is the quality of the distractors. H/H is high π and high r condition, H/L is high π and low r condition, L/H is low π and high r condition, and L/L is low π and low r condition.

In general, the RRUM had equal or better convergence compared to the EGCA. When the RRUM was equivalent to the EGCA with respect to the total amount of information obtained from the complete responses (i.e., non-informative distractors), the median PC of the RRUM was nearly identical to the mean PC for the EGCA in most cases. However, when the polytomous responses using the EGCA provided more information than the data was dichotomized for in an estimation using the RRUM (i.e., strong- and weak-distractor conditions), the median PC was higher for the RRUM than for the EGCA, which could be due to the EGCA's larger number of parameters.

4.1.2 Parameters Recovery

In addition to exploring convergence, item-parameter estimation explored. Note that the quality of item parameter estimates directly influences CCRs, which is the primary focus of this study. The mean absolute difference and correlation between the estimated and true parameters for the EGCA across all conditions were presented. Lastly, the correct option parameters of the EGCA were compared to the parameters of the RRUM using the MAD and correlation.

4.1.2.1 Mean Absolute Difference Between the True and Estimated Parameters of EGCA

The MAD results between the estimated EGCA parameters and the true parameters are shown in Figure 3 and Table 9. In general, the MAD of the parameters was smaller than .18 for the most of conditions indicating that the parameters were recovered well for the model. The MAD of π (.01~.08) was smaller than the MAD of r (.02~.20), because there are more π than r in all conditions. In addition to the general difference between the MADs for estimation of the π and the estimation of r, the MAD

varied with respect to the number of attributes. Specifically, when using 6 attributes (.02~.20), the MAD was higher than the MAD of the 4-attribute (.01~.17) conditions. The results could be caused the added complexity (i.e., the number of parameters) relative to the lack of increase in sample size for the six attributes when compared to the four-attribute conditions. Furthermore, the MAD of the 40-item condition (Mean = .06) was lower than the MAD of the 20-item condition (Mean = .07), and the MAD of the 2,000 sample-size condition (Mean = .06) was lower than the MAD of the 1,000 sample-size condition (Mean = .08). The results show that the accuracy of estimation is positively associated with the sample size and test length.

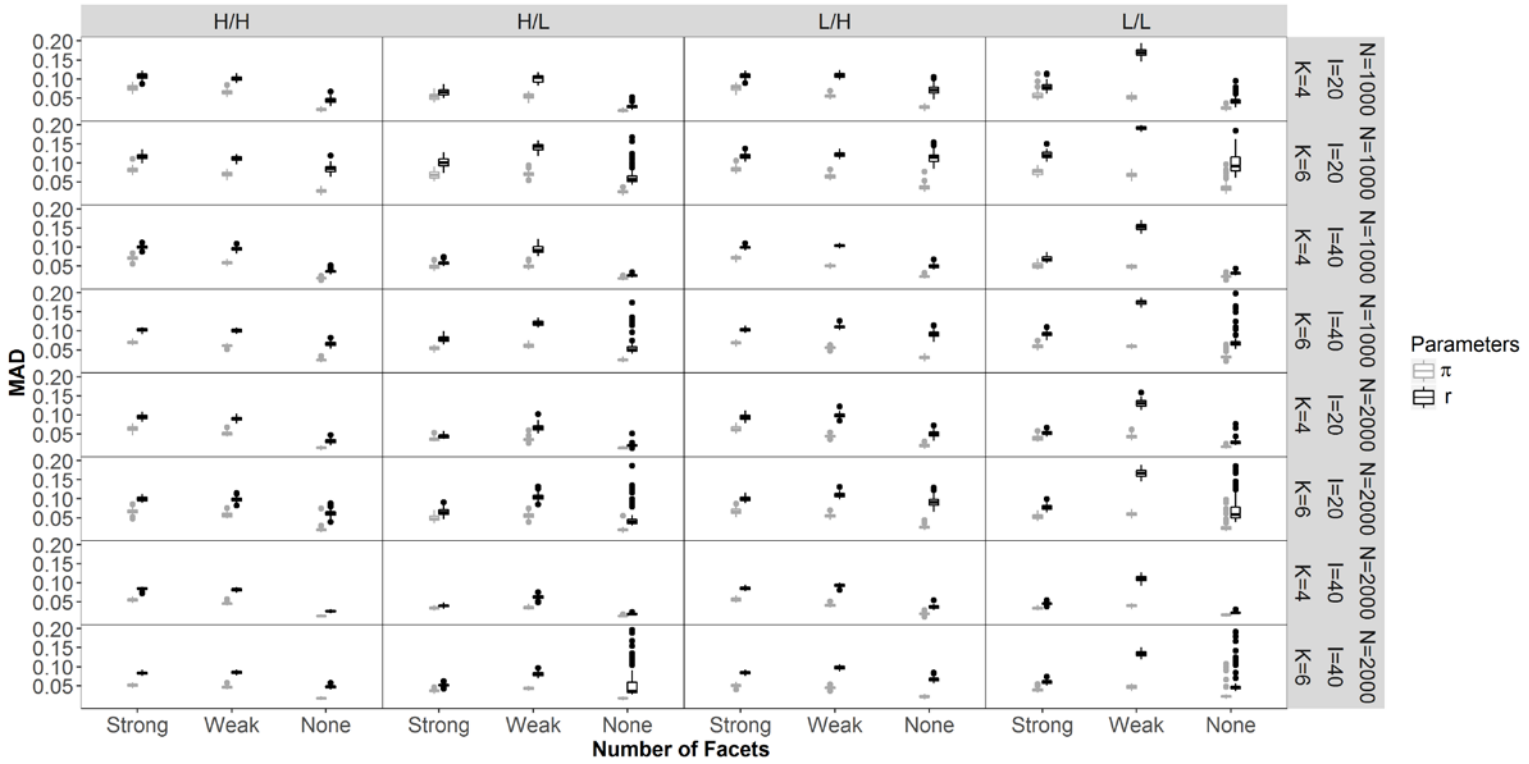

Figure 3. Mean Absolute Difference Between Estimated Parameters and True Parameters for the EGCA Across Various Conditions.

The mean absolute difference (MAD) between the true and estimated parameters across different conditions are shown above. Each section of the graph represents the MAD under strong, weak and non-informative distractors. The columns represent the quality of the item while the rows represent a combination of sample sizes, attribute sizes, and test lengths. Each color represents a kind of parameter (i.e., π and r).

Table 9. Mean MAD of EGCA

N	*I*	*K*	*DQ*	H/H		H/L		L/H		L/L	
				π	r	π	r	π	r	π	r
1000	20	4	None	0.02	0.04	0.02	0.03	0.03	0.07	0.02	0.04
			Weak	0.07	0.10	0.05	0.10	0.06	0.11	0.05	0.17
			Strong	0.08	0.11	0.05	0.07	0.08	0.11	0.06	0.09
		6	None	0.03	0.08	0.03	0.07	0.04	0.11	0.04	0.10
			Weak	0.07	0.11	0.07	0.14	0.07	0.12	0.07	0.20
			Strong	0.08	0.12	0.07	0.10	0.08	0.12	0.08	0.12
	40	4	None	0.02	0.04	0.02	0.03	0.02	0.05	0.02	0.03
			Weak	0.06	0.09	0.05	0.09	0.05	0.10	0.05	0.15
			Strong	0.07	0.10	0.05	0.06	0.07	0.10	0.05	0.07
		6	None	0.02	0.07	0.02	0.06	0.03	0.09	0.03	0.07
			Weak	0.06	0.10	0.06	0.12	0.06	0.11	0.06	0.18
			Strong	0.07	0.10	0.05	0.08	0.07	0.10	0.06	0.09
2000	20	4	None	0.01	0.03	0.01	0.02	0.02	0.05	0.02	0.03
			Weak	0.05	0.09	0.04	0.07	0.04	0.10	0.04	0.13
			Strong	0.06	0.09	0.04	0.04	0.06	0.09	0.04	0.05
		6	None	0.02	0.06	0.02	0.05	0.03	0.09	0.03	0.07
			Weak	0.06	0.10	0.06	0.11	0.06	0.11	0.06	0.17
			Strong	0.07	0.10	0.05	0.07	0.07	0.10	0.05	0.08
	40	4	None	0.01	0.03	0.01	0.02	0.02	0.04	0.02	0.02
			Weak	0.05	0.08	0.04	0.06	0.04	0.09	0.04	0.11
			Strong	0.05	0.08	0.03	0.04	0.06	0.09	0.03	0.05
		6	None	0.02	0.05	0.02	0.06	0.02	0.07	0.02	0.06
			Weak	0.05	0.09	0.04	0.08	0.05	0.10	0.05	0.13
			Strong	0.05	0.08	0.04	0.05	0.05	0.08	0.04	0.06

Note: *N* is the sample size, *I* is the test length, *K* is the number of attributes, and *DQ* is the quality of the distractors. H/H is high π and high r condition, H/L is high π and low r condition, L/H is low π and high r condition, and L/L is low π and low r condition.

In addition to the estimation of each parameter, there is a distinction between parameters related to the correct response and those associated with the distractors. Recall that, when using the EGCA, the correct option did not include guessing whereas the distractors require a guess to some degree. The MAD of π in the strongly informative distractor condition was between .03 and .08 and generally similar to the MAD of the estimation of π in the weak-distractor condition, which was between .04 and .07. In addition, the MAD of both conditions (strong and weak) was higher than the MAD of the non-informative distractors condition (.01~.04). Given the same test length and sample size, the strong and weak distractor π recovery was worse than the π recovery in the non-informative distractors because there were more π estimated. Recall that in the non-informative condition, π values were constrained to be equal to zero for the unlinked distractors.

However, the MAD of r did not follow the sequence, which was the MAD of r in a non-informative condition was smaller than the MAD of r in a strongly informative and weakly informative distractor condition. The MAD of r was the highest for the weak-distractor condition, which was between .06 and .20, followed by the strong-distractor condition (.04~.12). The non-informative distractor condition (.02~.11) has the lowest MAD compared to the other two conditions. The results could be because the values of r may be more difficult to recover when the values of π are small (i.e., a weak informative distractor condition) than the values of π are large (i.e., strongly informative and non-informative distractor condition). Only in the weakly informative distractor condition is

the π for the distractors half the value of the π for the correct option. With small π values, even a small value of r will have difficulty to recover.

The estimation of item parameters also depends on the item quality condition. Item parameter estimation tends to be the best (i.e., low MAD) for the good-quality item condition. The MAD of π was generally low under the high π and low r condition (Mean = .04), followed by the low π and low r (Mean =.04). The high π and high r condition (Mean = .05) and the low π and high r condition (Mean = .05) had the highest MAD of π. Similarly, The MAD of r was also generally low under the high π and low r condition (Mean = .07), followed by the high π and high r (Mean =.08). The low π and low r condition (Mean = .09) and the low π and high r condition (Mean = .09) had the highest MAD of r. The results indicate that the magnitude of the parameters directly influenced the recovery of the parameters. The results also suggest that the recovery of π and r are best when the item quality is high (i.e., high π and low r).

4.1.2.2 Correlation Between True and Estimated Parameters in EGCA

The results in Figure 4 and Table 10 shows the correlation between the true and estimated parameters (i.e., π and r). In general, increases in the number of attributes decreased the correlation between the estimates and the true parameters, while increases in the sample size and test length led to increased values in the correlation between the estimated and true parameters with their respective estimates. The correlation of the true values and the estimates for π (Mean = .85) was higher than the correlation for the estimates with truth for r(Mean = .50) because there are more r than π in all the

conditions. The correlation of π (excluding π parameters that are constrained to 0) with their corresponding estimates was higher under the non-informative distractor condition (Mean = .90) and weak-distractor condition (Mean = .93) than the strong- distractor condition (Mean = .73). The reason could be that π in the strong distractor condition was rescaled to be smaller than in other conditions (See Table A1) and it was not recovered well. The correlation of r parameters with the corresponding estimates were highest under the non-informative distractor condition (Mean = .67) followed by the strong-distractor condition (Mean = .46), with the weak-distractor condition having the lowest correlation (Mean = .35). The reason that r in the non-informative distractors condition was recovered better than the other two conditions is that the average π was relatively larger in the non-informative condition than in other two conditions, and r is difficult to recover under the small π condition.

The correlation between the true and estimated π were also related to item quality. The correlation was the highest for the high π and low *r* condition ($\bar{r}$ = .92) and lowest for the low π and high *r* condition ($\bar{r}$ = .82). The low π and low *r* condition ($\bar{r}$ = .84) and the high π and high *r* condition ($\bar{r}$ = .83) are in between. The results showed that recovery of the π was influenced by the combination of the value of π and r, specifically, a good quality of item could result in a better recovery of the π. Similarly, the correlation of r was the highest for the high π and high *r* condition ($\bar{r}$ = .61) and lowest for the low π and low *r* condition ($\bar{r}$ = .38). The low π and high *r* condition ($\bar{r}$= .54) and the high π and low *r* condition ($\bar{r}$ = .45) are in between. The results show that a higher value of r in

combination with a higher value of π could result in a higher recovery of r. In cases in which the π values are low, it might be difficult to recover r.

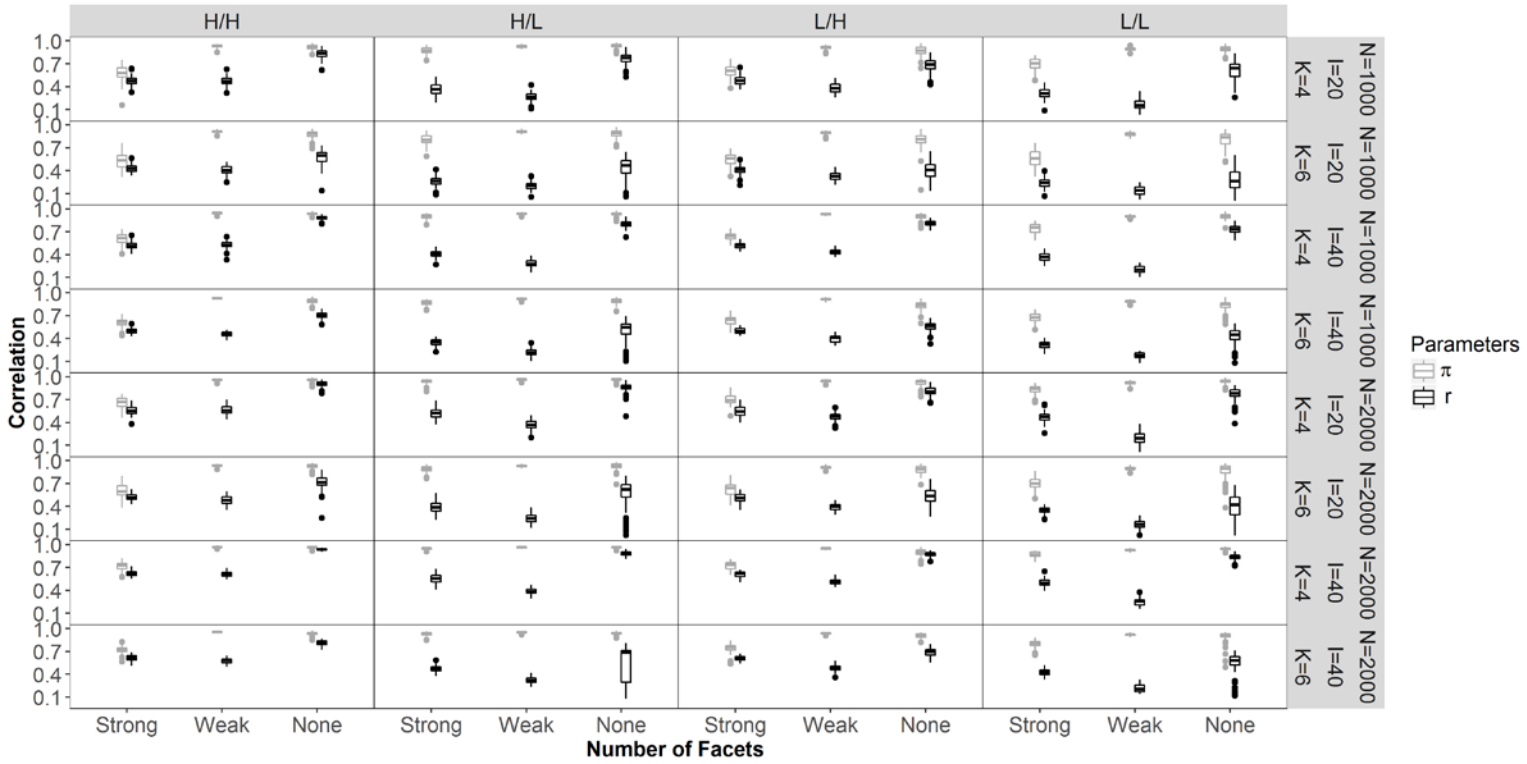

Figure 4. Correlation Between Estimated and True Parameters for the EGCA Across Various Conditions
The correlation between EGCA true parameters and estimated parameters across different conditions is shown above. Each section of the graph represents the three distractor conditions. The columns represent the quality of the item while the rows represent the combination of sample sizes, attribute sizes, and test lengths. Each color represents a kind of parameter (i.e., π and r)

Table 10. Mean Correlation of EGCA Estimated and True Parameters

N	*I*	*K*	*DQ*	H/H		H/L		L/H		L/L	
				π	r	π	r	π	r	π	r
1000	20	4	None	0.91	0.83	0.93	0.76	0.87	0.68	0.89	0.62
			Weak	0.93	0.47	0.93	0.26	0.91	0.38	0.89	0.16
			Strong	0.58	0.47	0.87	0.37	0.60	0.48	0.67	0.31
		6	None	0.87	0.58	0.88	0.43	0.80	0.40	0.80	0.27
			Weak	0.91	0.41	0.91	0.20	0.89	0.33	0.88	0.14
			Strong	0.52	0.43	0.80	0.26	0.55	0.41	0.56	0.24
	40	4	None	0.93	0.88	0.93	0.80	0.89	0.81	0.90	0.73
			Weak	0.94	0.52	0.94	0.28	0.93	0.43	0.90	0.20
			Strong	0.61	0.52	0.90	0.41	0.64	0.52	0.74	0.37
		6	None	0.89	0.70	0.89	0.49	0.83	0.56	0.83	0.43
			Weak	0.93	0.46	0.92	0.22	0.91	0.40	0.89	0.18
			Strong	0.61	0.49	0.87	0.35	0.63	0.50	0.67	0.31
2000	20	4	None	0.96	0.90	0.96	0.86	0.92	0.80	0.94	0.77
			Weak	0.96	0.57	0.96	0.36	0.94	0.47	0.92	0.20
			Strong	0.66	0.55	0.93	0.52	0.70	0.54	0.83	0.46
		6	None	0.92	0.71	0.93	0.57	0.88	0.53	0.87	0.38
			Weak	0.93	0.48	0.93	0.24	0.91	0.39	0.89	0.16
			Strong	0.61	0.52	0.89	0.38	0.62	0.51	0.70	0.35
	40	4	None	0.96	0.94	0.96	0.88	0.90	0.87	0.94	0.83
			Weak	0.97	0.61	0.96	0.39	0.95	0.51	0.93	0.25
			Strong	0.71	0.62	0.95	0.55	0.73	0.61	0.87	0.50
		6	None	0.94	0.81	0.94	0.56	0.90	0.69	0.89	0.53
			Weak	0.95	0.58	0.95	0.32	0.94	0.48	0.92	0.21
			Strong	0.72	0.62	0.93	0.47	0.74	0.61	0.80	0.43

Note: *N* is the sample size, *I* is the test length, *K* is the number of attributes, and *DQ* is the quality of the distractors. H/H is high π and high r condition, H/L is high π and low r condition, L/H is low π and high r condition, and L/L is low π and low r condition.

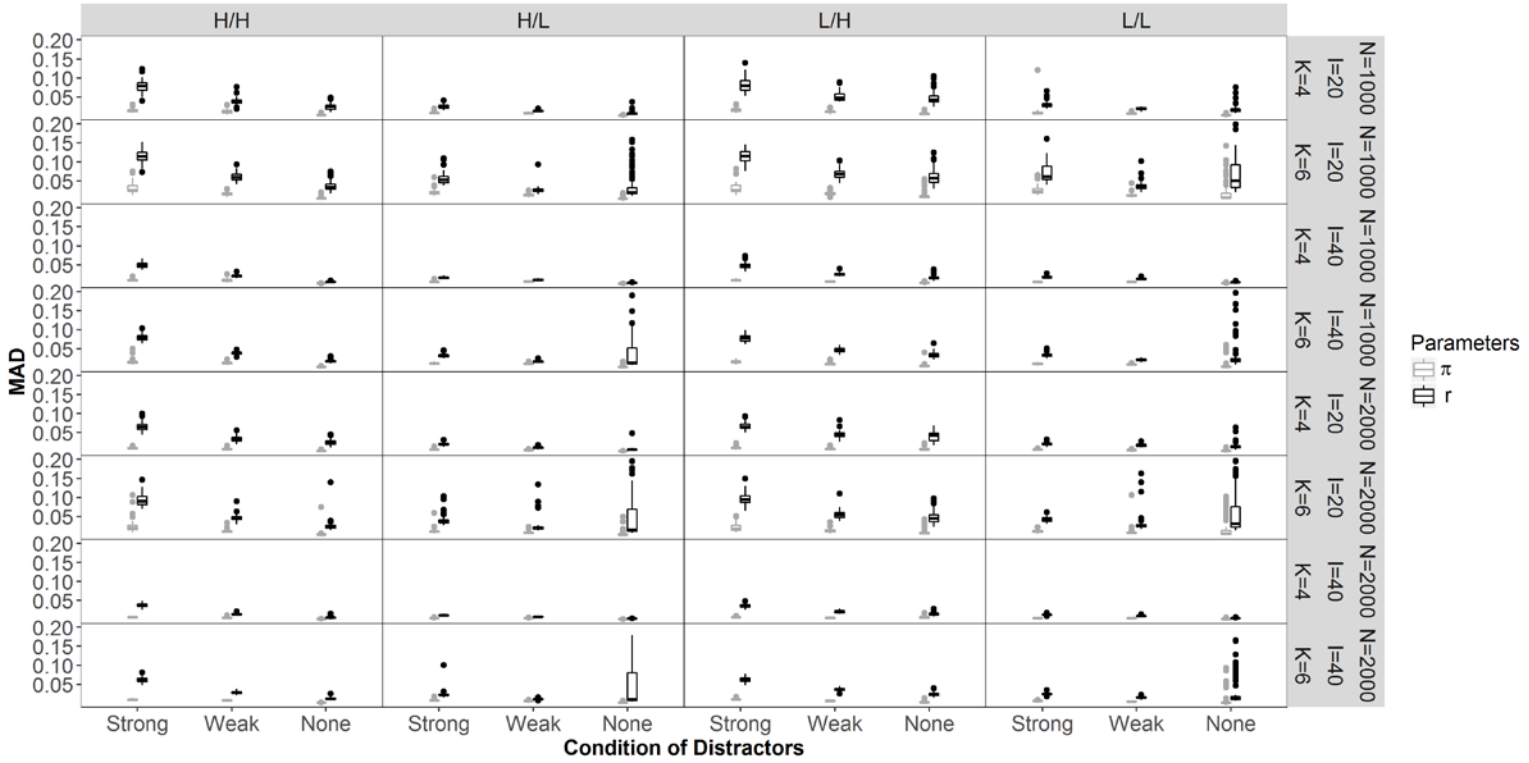

Figure 5. Mean Absolute Difference Between the Correct Option Parameters of the EGCA and the Parameters of the RRUM

The mean absolute difference (MAD) between the EGCA correct-option parameters and the RRUM parameters is shown above. Each section of the graph represents the three distractor conditions. The columns represent the quality of the item while the rows represent a combination of sample sizes, attribute sizes, and test lengths. Each color represents a kind of parameter (i.e., π and r).

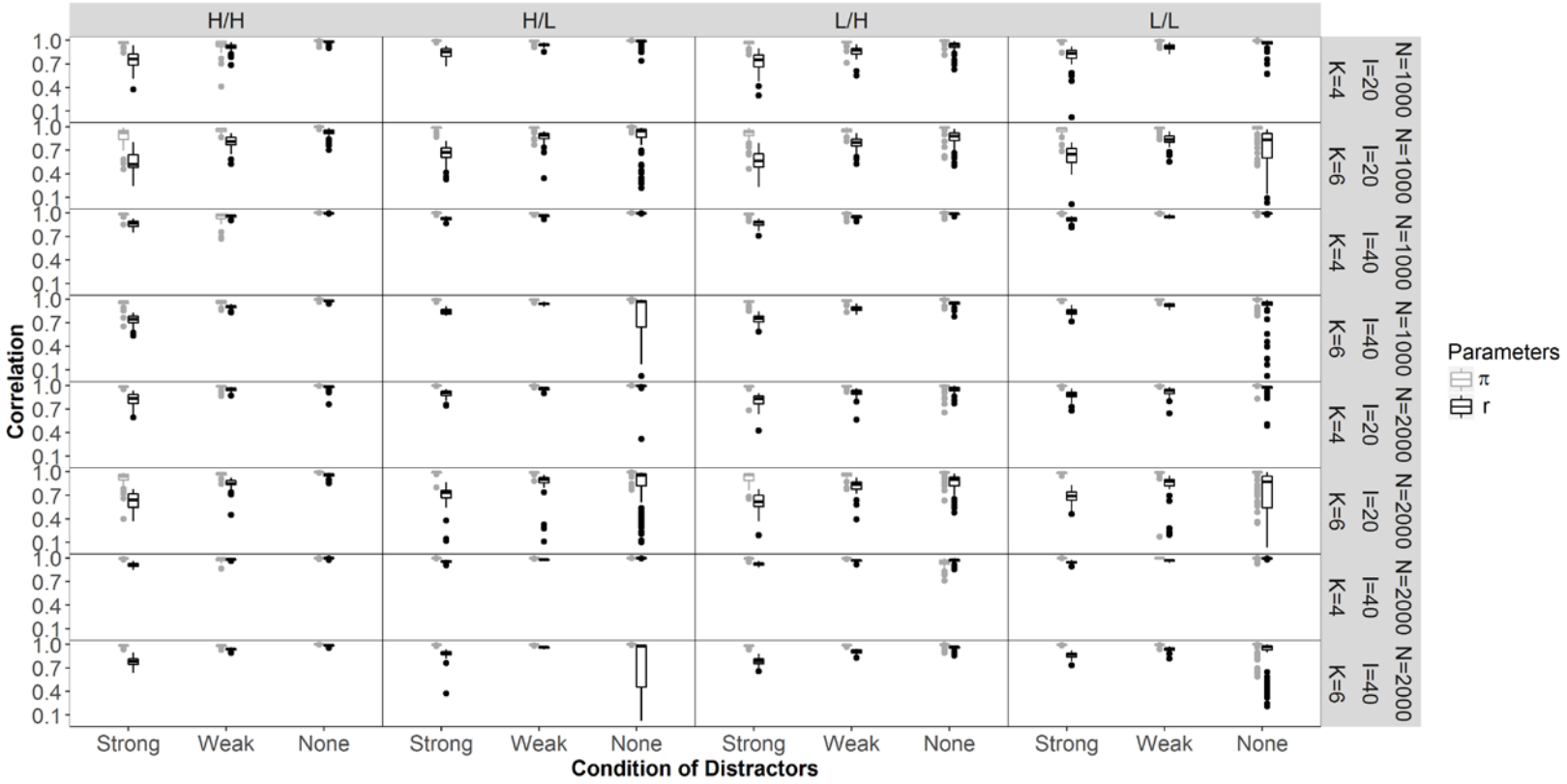

Figure 6. Correlation Between the Correct Option Parameters of the EGCA and the Parameters of the RRUM
The correlation between the EGCA correct-option parameters and the RRUM parameters is shown above. Each section of the graph represents the three distractor conditions. The columns represent the quality of the item while the rows represent a combination of sample sizes, attribute sizes, and test lengths. Each color represents a kind of parameter (i.e., π and r).

4.1.2.3 MAD Between the Correct Option EGCA and the RRUM

Furthermore, as a proof of concept, it was shown that the correct option parameters of the EGCA were equivalent to the parameters of RRUM. The MAD for the estimated EGCA and RRUM parameters (i.e., π, r) is shown in Figure 5. In general, the MAD was low across all conditions, indicating that the correct option parameters of the EGCA and the RRUM parameters were equivalent. In addition, similar to estimation accuracy in general, The MAD of π (Mean = .005) was lower than the MAD of r (Mean = .027). The range of the MAD of r tended to be higher than the range of the MAD of π. The MAD of π between the two models was the lowest when the item quality was high. Specifically, the MAD of π was lowest under a high π and a low r condition, and highest under a low π and a high r condition. Similarly, the MAD of r was lowest under a high π and a high r and a high π and low r condition, and highest under a low π and a high r condition. The results showed that the MAD of r between the two models was low when the π was large because the low π may lead to difficulty of recovering r.

Moreover, the MAD between the two models was the lowest for r when the distractor was non-informative. As the distractor became more informative, the MAD of r increased as well, and the MAD of π was similar across all distractor conditions. The result could be that, for strongly and weakly informative distractor conditions, the π of the distractors were small, which could lead to difficulty of recovering r.

Additionally, increases in the number of attributes measured by the assessment led to an increase in MAD between the two models, but increases in the sample size and

test length both led to decreases in all MAD estimates. When the test length or the sample size was small, or the attribute number measured by the assessment was large, the EGCA parameters were not recovered well, and thus, the MAD between the EGCA and RRUM would almost necessarily increase. Despite that, the average MAD between the two models was small (below .1), which was sufficient to demonstrate the fact that the item parameters of the two models were equivalent under the non-informative distractor condition.

4.1.2.4 Correlation Between the Correct Option EGCA and the RRUM

The correlations between the estimated parameters (i.e., π and r) when the EGCA was estimated versus the estimates obtained from the RRUM across different conditions are shown in Figure 6. In general, the results of the correlation between these estimates was consistent with the results when computing the MAD. Increasing the sample size and test length or decreasing the number of attribute measured an increase in the correlation between the correct option parameters of the EGCA and the RRUM parameters.

As in the MAD between the two models, the correlation between the two models was the highest for r when the distractor was non-informative and, as the distractor became more informative, the correlation of r decreased as well. The correlation of π was similar across all distractor conditions. The π ($\bar{r}$ = .99) between the two models was very highly correlated across all the conditions, and r ($\bar{r}$ = .92) had a very high correlation overall. The r was the lowest in the good-quality condition (High π and Low r condition; $\bar{r}$ = .89) and moderate-quality condition (Low π and Low *r* condition; $\bar{r}$ =

.88) compared to the high π and high r condition ($\bar{r}$ = .97) and the low π and high r condition ($\bar{r}$ = .93). Such results could be due to convergence. Recall that the proportion of converged parameters was lower for the EGCA under the non-informative condition for the high π and low r and the low π and low r conditions when the attribute size was 6. However, the difficulty of obtaining a converged solution for all parameters in these conditions did not affect the estimates of the attributes (i.e., correct classification rates did not appear to be affected).

Table 11. Ranges of Rescaled π (Mean Min~ Mean Max) Across Different Conditions

N	*I*	*K*	*DQ*	H/H	H/L	L/H	L/L
1000	20	4	None	0.71~0.89	0.71~0.89	0.51~0.69	0.51~0.69
			Weak	0.28~0.82	0.34~0.89	0.25~0.69	0.25~0.69
			Strong	0.35~0.62	0.40~0.85	0.33~0.63	0.38~0.70
		6	None	0.71~0.89	0.71~0.89	0.51~0.69	0.51~0.69
			Weak	0.26~0.81	0.28~0.89	0.24~0.69	0.25~0.69
			Strong	0.35~0.63	0.40~0.85	0.34~0.64	0.38~0.69
	40	4	None	0.70~0.90	0.71~0.90	0.51~0.69	0.51~0.69
			Weak	0.28~0.82	0.33~0.89	0.25~0.70	0.25~0.70
			Strong	0.34~0.63	0.39~0.86	0.32~0.65	0.37~0.70
		6	None	0.71~0.90	0.71~0.89	0.50~0.70	0.51~0.69
			Weak	0.25~0.82	0.27~0.89	0.24~0.69	0.25~0.69
			Strong	0.34~0.64	0.39~0.87	0.33~0.65	0.37~0.69
2000	20	4	None	0.71~0.89	0.71~0.89	0.51~0.69	0.51~0.69
			Weak	0.29~0.82	0.34~0.89	0.25~0.69	0.25~0.69
			Strong	0.35~0.61	0.40~0.86	0.33~0.62	0.38~0.70
		6	None	0.71~0.89	0.71~0.89	0.51~0.69	0.51~0.69
			Weak	0.26~0.81	0.28~0.88	0.25~0.69	0.25~0.69
			Strong	0.35~0.63	0.40~0.85	0.34~0.64	0.38~0.69
	40	4	None	0.71~0.90	0.70~0.89	0.51~0.70	0.50~0.70
			Weak	0.28~0.82	0.33~0.89	0.25~0.69	0.25~0.70
			Strong	0.34~0.63	0.39~0.86	0.32~0.64	0.37~0.70
		6	None	0.70~0.90	0.71~0.89	0.51~0.69	0.51~0.70
			Weak	0.25~0.82	0.28~0.89	0.24~0.69	0.25~0.69
			Strong	0.34~0.66	0.39~0.86	0.33~0.66	0.37~0.70

Note: *N* is the sample size, *I* is the test length, *K* is the number of attributes, and *DQ* is the quality of the distractors. H/H is high π and high r condition, H/L is high π and low r condition, L/H is low π and high r condition, and L/L is low π and low r condition.

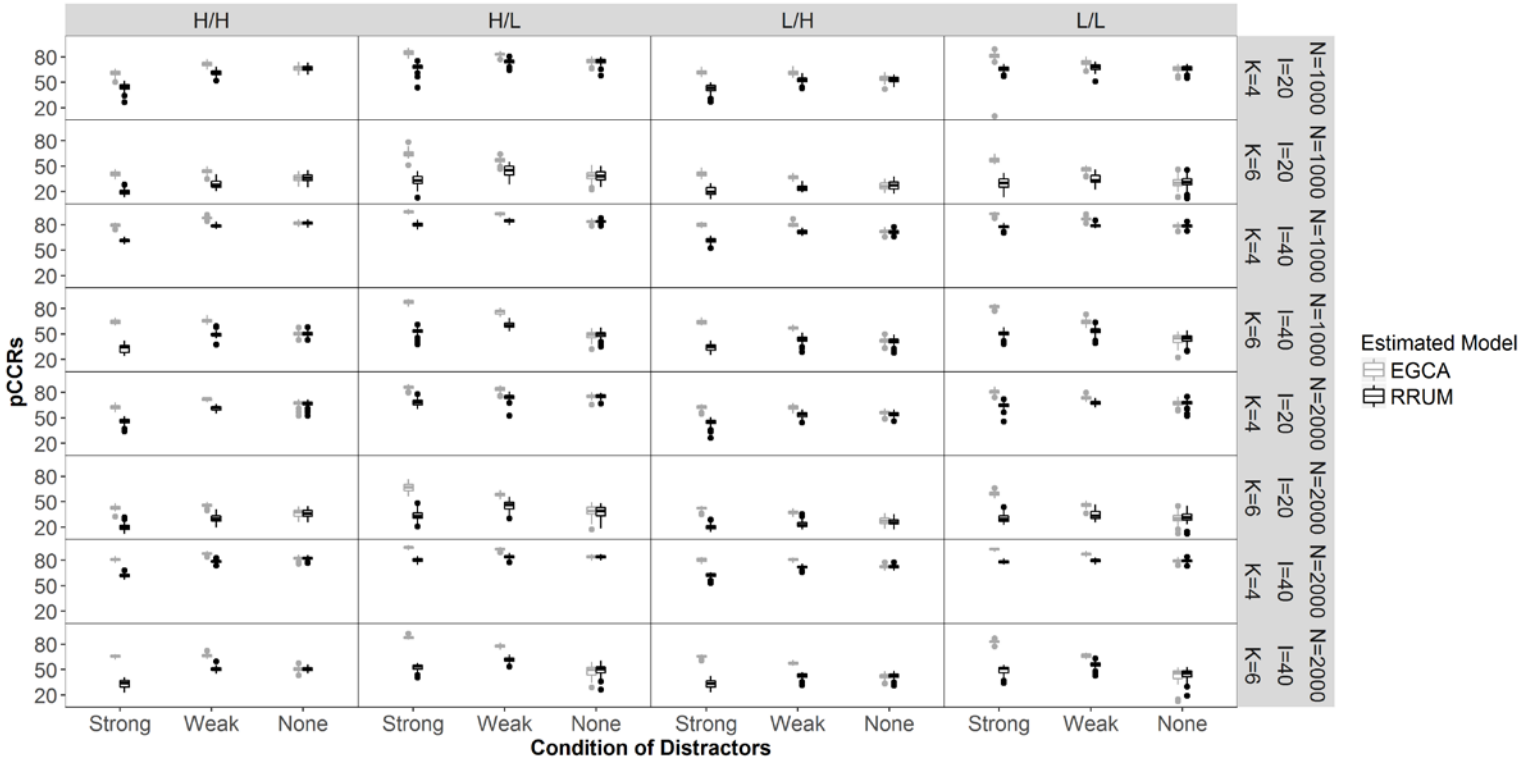

Figure 7. pCCRs Across Different Conditions

The profile correct classification rates (pCCRs) across different conditions are shown above. Each section of the graph represents three distractor conditions (i.e., strong, weak, and none) in which the EGCA and RRUM were used. The columns represent the item quality while the rows represent a combination of sample sizes, attribute sizes, and test lengths. Colors represent models.

Table 12. Mean pCCRs Across Different Conditions

N	I	K	DQ	H/H		H/L		L/H		L/L	
				RRUM	EGCA	RRUM	EGCA	RRUM	EGCA	RRUM	EGCA
1000	20	4	None	66.48	66.82	74.79	74.80	52.96	54.43	66.43	65.88
			Weak	61.30	71.44	74.42	82.94	52.37	61.43	67.60	73.61
			Strong	43.86	60.69	67.57	85.08	42.13	61.42	65.44	79.72
		6	None	35.72	35.93	38.66	38.31	27.21	26.45	31.05	30.65
			Weak	28.65	43.99	43.75	57.43	24.43	37.06	34.35	45.67
			Strong	19.17	40.67	33.52	64.62	20.43	40.87	29.86	57.42
	40	4	None	81.97	81.98	83.41	83.41	71.64	71.83	78.61	78.53
			Weak	78.77	87.80	84.45	92.76	71.19	79.84	79.01	86.60
			Strong	61.23	79.33	79.98	95.04	61.28	79.35	77.37	92.64
		6	None	50.39	50.38	49.03	48.35	41.56	41.84	44.50	43.65
			Weak	49.31	65.75	60.19	75.80	43.15	57.15	53.19	63.99
			Strong	32.78	64.46	52.80	87.62	33.89	63.98	50.23	81.88
2000	20	4	None	66.51	67.25	75.53	75.52	54.16	55.92	67.49	67.28
			Weak	61.51	72.05	74.35	83.65	53.60	62.21	67.58	73.48
			Strong	45.89	62.36	68.06	85.71	44.71	62.35	64.63	80.83
		6	None	35.98	36.04	37.96	38.80	26.40	27.61	30.39	30.08
			Weak	30.35	45.84	44.94	58.52	23.61	37.56	33.71	46.33
			Strong	18.49	42.48	33.74	66.66	19.62	42.14	30.45	59.61
	40	4	None	82.41	82.32	84.00	84.03	72.39	72.33	79.22	79.18
			Weak	78.74	87.89	83.92	92.92	71.80	80.51	79.44	87.18
			Strong	62.20	80.78	80.06	95.18	62.55	80.32	78.03	93.02
		6	None	50.57	50.49	48.92	47.94	42.14	41.86	44.32	43.77
			Weak	50.49	66.56	61.60	77.66	42.66	57.40	55.35	66.25
			Strong	33.32	65.78	52.54	87.79	32.95	65.24	49.18	82.73

Note: *N* is the sample size, *I* is the test length, *K* is the number of attributes, and *DQ* is the quality of the distractors. H/H is high π and high r condition, H/L is high π and low r condition, L/H is low π and high r condition, and L/L is low π and low r condition.

4.1.3 Effect of Item Parameters Rescaling

Rescaling was used to ensure the simulated item parameters satisfy the constraint for the EGCA, which is $S_\alpha < 1$. As mentioned in the Methods section, the rescaling only happened in π, not in r. The ranges of simulated π after rescaling across all conditions is shown in Table 11. There were two conditions of π: high and low. When the distractors were not informative, the parameters were not rescaled. However, when the distractors were weakly or strongly informative, the mean largest possible π decreased for all conditions expect for low π and low r condition. For the high π and high r condition, the mean largest π decreased drastically compared to the other distractor conditions.

Additionally, the largest π in the strong distractor condition appears to be smaller than the largest π in the weak- and non-informative distractor conditions, which can lead to better item quality for the weak-distractor condition when π and r were both high. Lastly, the same size, test length, and attributes did not appear to affect the range of the π, and the r was not affected by rescaling (See Table A1 in Appendix A)

4.1.4 Correct Classification Rates for the Profile (pCCRs)

Given the good item parameters in general, the results of pCCRs can be interpreted with confidence. Figure 7 and Table 12 show the results of the mean pCCRs across different conditions. Recall that the pCCR is the proportion of times that an examinee's estimated profile perfectly matched the true (simulated) attribute profile. In general, as the sample size and test lengths increased, the mean pCCRs increased, and as the number of attributes measured by a test increased, the mean pCCRs decreased. The

mean pCCRs ranged from 18% to 84% for the RRUM and 26% to 95% for the EGCA, indicating some conditions might have acceptable classifications and others might not have acceptable classification rates of a complete profile.

In terms of the average effect of item quality, the high π and low *r* condition had the highest pCCRs for the RRUM (Mean = 51.08%) and the EGCA (Mean = 74.28%), the low π and low r for the RRUM (Mean = 56.57%) and the EGCA (Mean = 67.09%) condition and the high π and high r for the RRUM (Mean = 51.08%) and the EGCA (Mean = 62.91%) had the medium pCCRs, and the low π and high *r* condition had the lowest pCCRs for the RRUM (Mean = 45.34%) and the EGCA (Mean = 56.75%). Again, as previously discussed, although the rescaling did have some marginal effect, these results suggest that the rescaling of items to satisfy the requirements of the EGCA did not have an effect. Specifically, because the π of all options were rescaled if S_α was greater than 1, the results suggest that the rescaling did not change the order of the different combination of π and r. For example, the high π and low r had the highest pCCRs in a previous study (Oksana et al. 2016) and it still had the highest pCCRs. The results also confirmed that the quality of the item was influenced by the value of π and *r* of the options. The EGCA (Mean = 56.03%) and RRUM (Mean = 55.92%) had essentially equal (within a small margin of error) pCCRs across most conditions when there was no information in the distractor. In cases where the convergence of parameters was an issue for the EGCA, specifically, under the condition of high π and low *r*—that is, 6 attributes, 40 items with 1,000 or 2,000 simulees—the RRUM (Mean = 48.06 %) had similar

pCCRs than the EGCA (Mean = 47.73%). The results in this particular instance indicate that CCRs were not much affected by the convergence of item parameters.

However, for the EGCA, when there was information in the distractors, with the exception of the high π and high r condition, there was a relationship between the amount of information in the distractors and pCCRs. Specifically, a strong-distractors condition (Mean = 74.69%) always had the highest pCCRs, which was followed by pCCRs when using weak distractors (Mean = 68.29%). The non-informative distractor condition had the lowest pCCRs (Mean = 55.14%). As for the high π and high r condition, it is believed that the different effect could have contributed to the need for rescaling the parameters. Specifically, for the high π and high r condition, as previously mentioned, the maximum rescaled π was higher for the weak- distractor condition than for the strong-distractor condition, and the maximum rescaled π was higher for the non-informative distractor condition than for the weak-distractor condition. The results could be because pCCRs were lower for the strong-distractor condition than for the weak-distractor condition and lower for the weak-distractor condition than for the non-informative distractor condition when the π and r were high.

Furthermore, the EGCA outperformed the RRUM across all conditions (e.g., different conditions of π and r, number of attributes, number of items) when the options were informative. The differences between the EGCA and the RRUM were even more prominent for the strong distractor conditions (14~35%) than for the weak distractor condition (6~17%). In other words, the advantage of the EGCA over the RRUM with respect to pCCRs was higher when the distractors were strongly discriminated (Mean =

23.02%) compared to conditions when the distractors were weakly discriminated (Mean = 11.18%). The results indicated that the EGCA could utilize the information provided in the distractors to provide a more accurate profile and attributes classification

One finding that was not quite consistent with expectations was that pCCRs changed for the RRUM across distractor conditions when all other conditions were held constant. Note that, based on the definition of the EGCA, the RRUM is defined such that the item parameters are identical to the parameters of the EGCA correct option. Within conditions when manipulating the quality of the distractors, the overall quality of the correct response should not change and, as a result, the item quality when using the RRUM should not change. However, these results could be caused by a combination of several factors. Due to the effect or to rescaling, the largest π decreased all item quality conditions except for the low π and low r condition, which explain the similarities of pCCRs for the low π and low r condition and a decrease in pCCRs for other item quality conditions for the RRUM.

4.1.5 Marginal Correct Classification Rates for an Attribute (aCCRs)

The results of aCCRs are discussed in this section. Figure 8 and Table 13 show the results of aCCRs across different conditions respectively. The mean aCCRs ranged from 75% to 98% for the EGCA and 71% to 95% for the RRUM. As the sample size and test length increased, or the number of attributes measured by a test decreased, the aCCRs increased, which is similar to the trends observed with the pCCR. The high π and low r condition had the highest aCCRs for the RRUM (Mean = 88.50%) and the EGCA (Mean = 98.06%), and the low π and high r condition had the lowest aCCRs for the RRUM

(Mean = 82.35%) and the EGCA (Mean = 94.54%). The aCCRs of the other π and r conditions were somewhere between the two conditions. The EGCA (Mean = 86.22%) and RRUM (Mean = 86.10%) had equal aCCRs across most of the conditions when there was no information in the distractors. In cases where the convergence of parameters was an issue for the EGCA (i.e., high π and low r, 6 attributes, 40 items) in the non-informative distractor condition, the RRUM and the EGCA still had the same aCCRs. For the EGCA, except for the high π and high r condition, the strong-distractor condition (Mean = 93.66%) always had higher aCCRs than the non-informative distractor condition (Mean = 85.80%), and the aCCRs of the weak distractor condition (Mean = 91.50%) was between the aCCRs of the strong distractor and non-informative distractor condition. The results may be also due to the rescaling of the π as discussed in the previous section. As with the results of pCCRs, the EGCA outperformed the RRUM across all conditions under 3L. The advantage of the EGCA over the RRUM was more prominent for the strong-distractor condition (Mean = 9.38%) than for the weak-distractor condition (Mean = 4.67%).

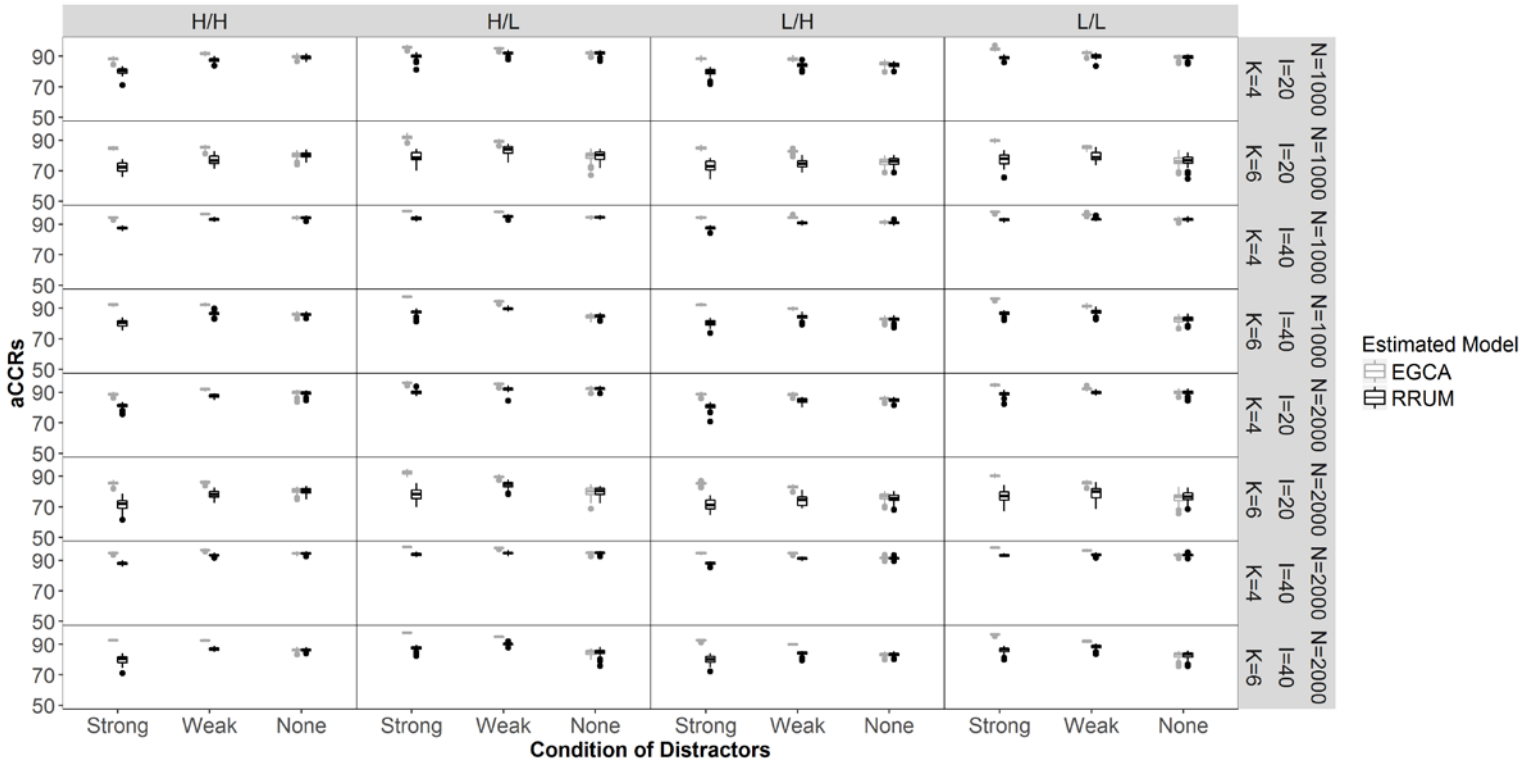

Figure 8. aCCRs Across Different Conditions.
The profile correct classification rates (aCCRs) across different conditions are shown above. Each section of the graph represents three distractor conditions (i.e., strong, weak, and none) where the EGCA and RRUM were used. The columns represent the quality of the item while the rows represent a combination of sample sizes, attribute sizes, and test lengths. Colors represent models.

Table 13. Mean aCCRs Across Different Conditions

N	*I*	*K*	*DQ*	H/H		H/L		L/H		L/L	
				RRUM	**EGCA**	**RRUM**	**EGCA**	**RRUM**	**EGCA**	**RRUM**	**EGCA**
1000	**20**	**4**	**None**	89.69	89.24	92.15	92.07	85.21	84.21	89.39	89.42
			Weak	91.67	87.43	95.10	91.99	88.14	84.03	92.22	89.96
			Strong	88.11	80.20	95.95	89.82	88.39	79.48	93.88	89.09
		6	**None**	80.12	79.99	79.42	79.66	75.78	75.98	76.43	76.69
			Weak	85.38	77.04	89.15	83.43	82.59	74.70	85.27	79.56
			Strong	84.79	72.19	91.80	79.06	84.84	73.08	89.72	77.30
	40	**4**	**None**	94.25	94.18	94.54	94.52	91.30	91.03	93.23	93.22
			Weak	96.67	93.19	98.04	95.03	94.32	90.86	96.26	93.38
			Strong	94.37	87.43	98.72	93.73	94.39	87.52	98.08	92.97
		6	**None**	86.00	86.04	84.57	84.78	83.04	82.90	82.67	82.83
			Weak	92.28	86.52	94.55	89.77	89.84	84.41	91.38	87.52
			Strong	92.34	80.07	97.56	87.16	92.22	80.40	96.25	86.33
2000	**20**	**4**	**None**	89.86	89.28	92.45	92.38	85.82	84.76	89.93	89.83
			Weak	91.93	87.64	95.36	91.96	88.48	84.52	92.17	89.85
			Strong	88.69	81.20	96.18	90.04	88.68	80.64	94.71	88.91
		6	**None**	80.08	80.24	79.93	79.75	76.53	75.65	75.82	76.69
			Weak	86.03	78.06	89.61	84.22	82.90	74.04	85.63	78.61
			Strong	85.46	71.65	92.35	78.26	85.35	71.36	90.30	77.00
	40	**4**	**None**	94.41	94.36	94.80	94.77	91.43	91.36	93.49	93.48
			Weak	96.72	93.33	98.10	94.76	94.53	91.10	96.41	93.57
			Strong	94.79	87.91	98.77	93.75	94.64	88.04	98.19	93.21
		6	**None**	86.25	86.25	84.54	84.74	83.18	83.25	82.85	82.86
			Weak	92.55	86.90	95.02	90.11	89.97	84.11	92.01	88.24
			Strong	92.66	79.92	97.57	87.35	92.55	79.99	96.45	86.17

Note: *N* is the sample size, *I* is the test length, *K* is the number of attributes, and *DQ* is the quality of the distractors. H/H is high π and high r condition, H/L is high π and low r condition, L/H is low π and high r condition, and L/L is low π and low r condition.

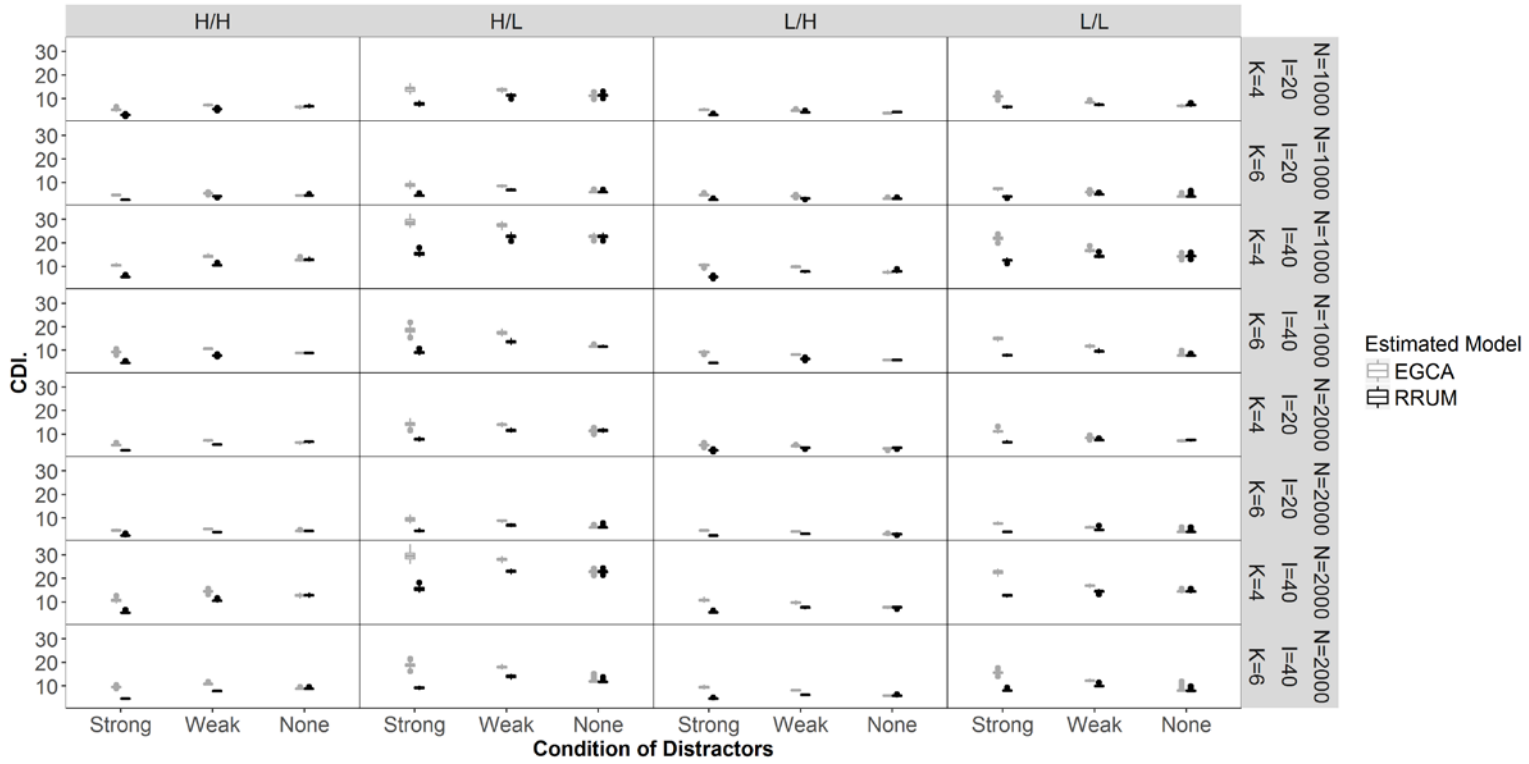

Figure 9. $CDI_.$ Across Different Conditions

The test-level cognitive discrimination index ($CDI_.$) across different conditions are shown above. Each section represents three distractor conditions (i.e., strong, weak, and none) where the EGCA and RRUM were used. The columns represent the quality of the item while the rows represent a combination of sample sizes, attribute sizes, and test lengths. Colors represent models.

Table 14. Mean $CDI_{\bullet}$ Across Different Conditions

N	*I*	*K*	*DQ*	H/H		H/L		L/H		L/L	
				RRUM	**EGCA**	**RRUM**	**EGCA**	**RRUM**	**EGCA**	**RRUM**	**EGCA**
1000	**20**	**4**	**None**	6.73	6.38	11.49	11.33	4.26	3.86	7.22	6.95
			Weak	5.60	7.20	11.40	13.73	4.13	4.87	7.36	8.34
			Strong	3.05	5.26	7.74	14.10	3.06	5.20	6.46	10.91
		6	**None**	4.58	4.59	5.88	5.89	3.15	3.15	4.08	4.08
			Weak	4.09	5.33	6.86	8.55	3.30	4.16	4.90	5.86
			Strong	2.60	4.67	4.49	8.95	2.62	4.69	4.05	7.31
	40	**4**	**None**	12.86	12.73	22.76	22.73	7.82	7.56	14.32	14.24
			Weak	10.51	14.39	22.63	27.53	7.74	9.78	14.32	16.74
			Strong	5.51	10.47	15.38	28.80	5.53	10.54	12.58	21.92
		6	**None**	8.81	8.82	11.63	11.67	5.89	5.89	7.76	7.84
			Weak	7.77	10.60	13.70	17.39	6.33	8.19	9.68	11.79
			Strong	4.61	9.32	9.01	18.53	4.61	9.30	7.91	15.02
2000	**20**	**4**	**None**	6.79	6.42	11.62	11.46	4.27	3.89	7.50	7.21
			Weak	5.63	7.26	11.59	14.05	4.16	4.94	7.42	8.46
			Strong	3.11	5.39	7.89	14.44	3.08	5.34	6.63	11.25
		6	**None**	4.54	4.55	6.05	5.99	3.12	3.12	4.26	4.19
			Weak	4.04	5.44	7.01	8.90	3.33	4.25	5.00	6.00
			Strong	2.61	4.80	4.64	9.47	2.55	4.70	4.15	7.78
	40	**4**	**None**	12.91	12.77	22.87	22.84	7.78	7.76	14.47	14.40
			Weak	10.53	14.48	22.98	28.17	7.70	9.79	14.43	16.96
			Strong	5.55	10.70	15.63	29.63	5.62	10.75	12.76	22.71
		6	**None**	8.75	8.76	11.84	12.01	5.86	5.87	7.94	8.06
			Weak	7.76	10.79	14.02	18.04	6.21	8.19	9.99	12.20
			Strong	4.61	9.53	9.14	18.81	4.59	9.42	8.03	15.56

Note: *N* is the sample size, *I* is the test length, *K* is the number of attributes, and *DQ* is the quality of the distractors. H/H is high π and high r condition, H/L is high π and low r condition, L/H is low π and high r condition, and L/L is low π and low r condition.

4.1.6 $CDI_.$

The correlation between $CDI_.$ and the classification accuracy across all conditions was first examined. Because previous literature has shown that the correlation between $CDI_.$ and pCCRs are not linear (Henson et al. 2008), the Pearson correlation between the log of $CDI_.$ and pCCRs were used. The results show that the correlation between the log of $CDI_.$ and pCCRs was .82, and between the log of $CDI_.$ and aCCRs was .81 for the EGCA. They were .80 and .82 respectively for the RRUM. The strong positive correlation between the log of $CDI_.$ and CCRs for the different models indicates that $CDI_.$ can be used as an accuracy indicator for the classification of polytomous and dichotomous DCM models.

The results of the mean $CDI_.$ across different conditions were summarized in Figure 9 and Table 14. In general, the results of $CDI_.$ were similar to the results of pCCRs and aCCRs. The $CDI_.$ ranged from 2.56 and 22.98 for the EGCA and from 3.13 and 29.32 for the RRUM. Since $CDI_.$ is a function of test length, the mean $CDI_.$ doubled as the test lengths increased from 20 items to 40 items. The mean $CDI_.$ increased slightly as the number of examinees increased. The mean $CDI_.$ decreased as the number of attributes measured by a test increased. Additionally, $CDI_.$ values were impacted by different values of π and r. High π and low r had a higher $CDI_.$ (Mean = 15.92) than low π and low r (Mean = 11.05) and high π and high r (Mean =8.34), which were higher than low π and high r (Mean = 6.35) for the EGCA. The results confirmed that the item

quality was the highest when the item quality was high, and lowest when item quality was low.

When no information was measured by the distractors, the EGCA (Mean = 8.66) and RRUM (Mean = 8.75) had the same mean $CDI_{\bullet}$, even in conditions where the EGCA parameters had an issue with convergence. Similar to the results of classification, for the EGCA, the strong distractor condition (Mean =13.05) always had higher $CDI_{\bullet}$ values on average than the weak distractors condition, and the weak distractor (Mean = 11.53) condition had higher $CDI_{\bullet}$ than the non-informative distractors condition (Mean = 8.84). This trend was generally true, except for the High π and High *r* conditions, which may be due to the rescaling issue (See Table 12) that was also described with respect to CCRs. The results indicated that strongly informative distractors had a higher discrimination than the weakly informative distractors, which had higher discrimination than the non-informative distractor condition. $CDI_{\bullet}$ of the EGCA model (Mean = 12.29) was higher than $CDI_{\bullet}$ of the RRUM (Mean = 7.53) across all conditions when the distractors were informative. The differences between the EGCA and the RRUM were even larger when the distractors were strongly discriminated than when the distractors were weakly discriminated.

The highest mean $CDI_{\bullet}$ for the EGCA was when a test had 40 good-quality (i.e., high π and low r) items that contained strong-informative distractors, measured 4 attributes, and had 2,000 examinees; the mean $CDI_{\bullet}$ was 20.63 for the EGCA, and it dropped to 9.97 when the RRUM was used. The results indicate that when the distractors

had more useful information, the EGCA model would have a higher $CDI_{\bullet}$ than the RRUM model. Otherwise, a lot of information would be lost if the RRUM was used, which was also reflected in the classification accuracy.

4.2 Real Data

In addition to a simulation study, a real-world data analysis was performed to examine the difference between the EGCA and the RRUM. As for the convergence of item parameters, the PC is .92 for the EGCA, and 1.00 for the RRUM. However, it seemed that item parameters with high values of GRR converged in a large range (with a large standard error) after the MCMC chains were visually inspected. Therefore, the results regarding classification and *CDI* can be meaningfully interpreted.

4.2.1 Classification

In the real data analysis, although the true parameters are not known, there are specific results that can be compared. Absolute deviance (AD) between posterior probability and .5 were summed across all examinees for each attribute for the EGCA and the RRUM (Figure 10). The higher the sum of AD is associated with higher classification discrimination. The results show the EGCA had a higher AD for Attribute 1(Skill 1) and Attribute 3(Misconception 2) than the RRUM. However, the RRUM had higher AD for Attribute 2 (Misconception 1) than the EGCA. The results indicate that the posterior probabilities of mastery when analyzing the polytomous data using the EGCA is higher than the posterior probability of mastery when analyzing the corresponding dichotomous data using the RRUM.

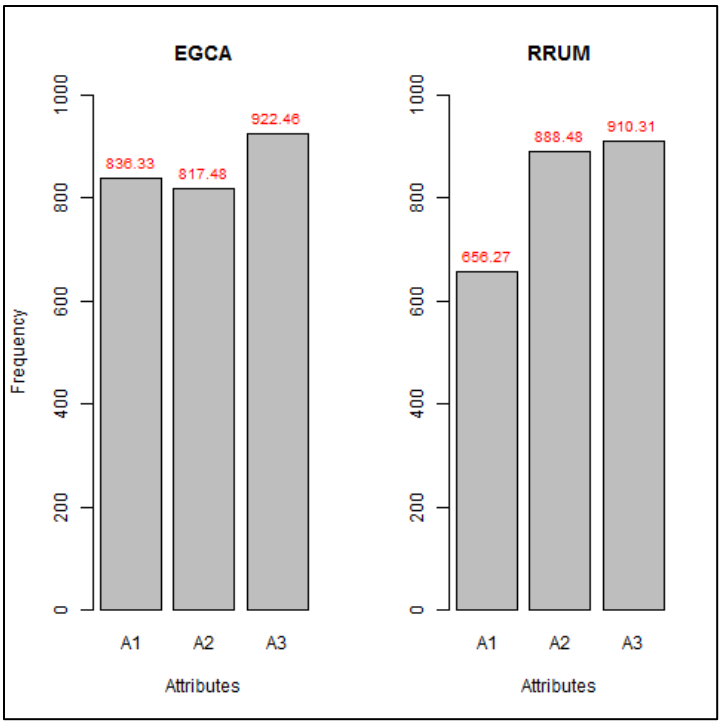

Figure 10. The Sum of the Absolute Deviance Between the Posterior Probability and .5 for Each Attribute Estimated Using EGCA and RRUM.
Note that: A1 is Skill 1, and A2 and A3 are Misconceptions 1 and 2. The A2 and A3 of the RRUM were recoded using 1- the posterior probability to make a fair comparison.

Figure 11 shows the attribute classification between the EGCA and RRUM analysis. The results show that the EGCA classified more examinees to have Skill 1 (62%) than the RRUM (39%). In addition, The EGCA classified fewer examinees (43%) to master Attribute 2 (Misconception 1) than the RRUM (68%). The EGCA also classified (26%) fewer examinees to master Attribute 3 (Misconception 2) than the RRUM (39%). There were big differences between the two models' classification. However, without knowing the true attributes the assessment measured, it is difficult to tell which model is closest to reality. However, the percentage of agreement of attribute

profiles between the two models was 48%, which indicates some similarity between the two models' classification.

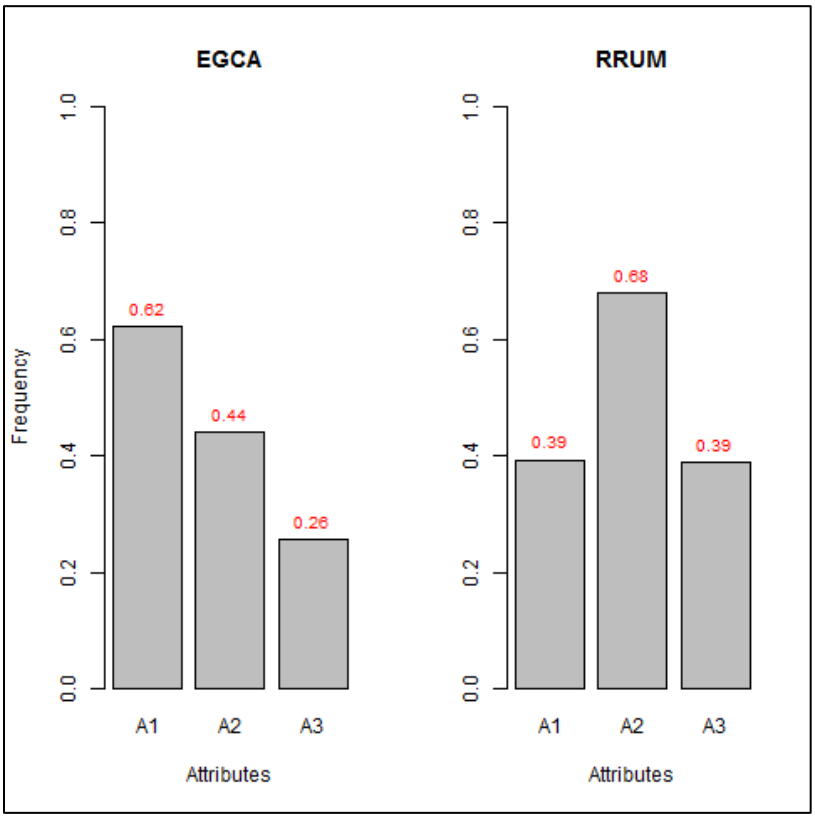

Figure 11. The Proportion of Examinees That Mastered Each Attribute Estimated Using EGCA and RRUM.
Note that: A1 is Skill 1, and A2 and A3 are Misconceptions 1 and 2. The A2 and A3 of the RRUM were recoded from 0 to 1 to make a fair comparison.

In addition to describing the general classification of examinees, it is possible to describe the association between an examinees' classification and his or her responses to the items. It is assumed that a better model would do a better job at predicting behavior. In the real data analysis, the correct answers measured Skill 1 or lack of Misconception 1 or the lack of Attribute 2. The misconception option in the real-world data analysis only

measures Misconception 1 or Misconception 2 (see Table 7 for the Q-matrix). It is expected that individuals matching the Q-matrix entry would be more likely to match that entry. This association is measured using a correlation between whether an individual matched the Q-matrix and the sum of the items with that option. Therefore, whether the examinees were classified as mastering the Skill 1 and lacking Misconception 1 was correlated with the number of correct options selected by examinees for items 1–5 that only measured Skill 1 and Misconception 1. The correlation was .33 for the EGCA and .03 for the RRUM. Additionally, whether the examinees were classified as mastering the Skill 1 and lacking Misconception 2 was correlated with the number of correct options they chose for items 6–12 that only measured Skill 1 and Misconception 2. The correlation was .36 for the EGCA and close to 0 (-.01) for the RRUM. The results show that the classification of examinees associated with the profile diagnosed by the correct options was moderately correlated with the correct options chosen by examinees using the EGCA. This was not the case for the RRUM.

The detailed classification corresponding to the number of the correct options chosen is shown in Table 14. The results illustrate that for examinees who did not select any correct options that measured mastery of Skill 1 and non-mastery of Misconception 1 were never classified as having mastered Skill 1 and not having mastered Misconception 2 when modeling the assessment using the EGCA. However, many examinees were classified to have mastered Skill 1 and not mastered Misconception 1 under the RRUM. In addition, the proportion of examinees mastering Skill 1 and not mastering Misconception 1 was directly related to the number of items answered correctly. When

more correct options were chosen, the both EGCA and RRUM analyses resulted in more examinees classified to have mastered Skill 1 and to lack Misconception 1. In particular, when the number of misconception options chosen was 4 or 5 by the examinees, both models classified the examinees to master Skill 1 and to lack Misconception 1.

Table 15. The Number of Selected Correct Options and the Corresponding Classification

Number of Correct Options	**Skill 1+ Misconception 1**				**Skill 1+Misconception 2**			
	EGCA		RRUM		EGCA		RRUM	
	0	1	0	1	0	1	0	1
0	478	0	332	146	834	0	801	33
1	345	4	158	191	342	0	274	68
2	258	144	117	285	209	12	104	117
3	46	299	37	308	103	93	42	154
4	1	237	1	237	7	115	8	114
5	0	199	0	199	0	116	0	116
6					0	98	0	98
7					0	82	0	82

Note: 0 means that not mastering the attribute profile, and 1 means mastering the attribute profile. For instance, if an examinee masters Skill 1 and lacks Misconception 1, then it is 1. If an examinee masters Skill 1 as well as Misconception 1, then it is 0.

Table 16. The Number of Selected Misconception Options and the Corresponding Classification

Number of Misconception Options	**Misconception 1**				**Misconception 2**			
	EGCA		RRUM		EGCA		RRUM	
	0	1	0	1	0	1	0	1
0	478	0	332	146	834	0	801	33
1	345	4	158	191	342	0	274	68
2	258	144	117	285	209	12	104	117
3	46	299	37	308	103	93	42	154
4	1	237	1	237	7	115	8	114
5	0	199	0	199	0	116	0	116
6					0	98	0	98
7					0	82	0	82

Note: 0 means lacking the misconception and 1 means mastering the misconception.

Having the attribute profile (mastering Skill 1 and lacking Misconception 2) was more aligned with the number of correct options selected by the examinees using the EGCA model than using the RRUM. For the other attribute profile (i.e., mastering Skill 1 and lacking Misconception 2), the results were similar. When 0 correct options were selected, the EGCA diagnosed no examinees to master Skill 1 and lack Misconception 2. However, this was not the case for the RRUM because some examinees were still classified to have that profile (i.e., mastering Skill 1 and lacking Misconception 2). Both models diagnosed the same number of the profiles when the number of correct options chosen was 5–7. The results were consistent with the findings of the correlation study, indicating that the EGCA was more realistic than the RRUM when used to model the skill and misconceptions of the correct option.

Moreover, for Misconception 1, the correlation between the examinees classified as mastering Misconception 1 and selecting the option related to that misconception was 0.81 for the EGCA, which was higher than 0.54 for the RRUM. For Misconception 2, the correlation for the EGCA was still higher (0.85) than the correlation for RRUM (0.76). The results show that the classification of the EGCA, which took into account the distinction between distractors that measured misconceptions, had a stronger correlation with the number of misconception options chosen than the RRUM.

The detailed classification of the misconception corresponding to the number of misconception options chosen is shown in Table 15. The results show that none of the examinees who chose 0 of the misconception options were diagnosed to have the corresponding misconception under the EGCA. However, some examinees were

diagnosed to have the same misconception when analyzing the data using the dichotomized data and the RRUM. When more misconception options were chosen, both the EGCA analysis and the RRUM analysis classified more examinees to master the misconceptions. In particular, when the number of misconception options chosen by the examinees was 5–7, both models classified the examinees to have misconceptions. The results indicate that the EGCA is more predictive than the RRUM in diagnosing the misconceptions.

4.2.2 Item/Test Quality (CDI_i / $CDI_\bullet$)

The $CDI_\bullet$ was 8.81 for the EGCA and 7.51 for the RRUM, which indicates that if the model fits the data, the EGCA results in item characteristics that would be predicted as more discriminating than the RRUM results. Table 17 shows the CDI_i for each item by the EGCA analysis and the RRUM analysis estimates. The CDI_i differences between the EGCA and RRUM are shown in Figure 12. For the first 10 items, the CDI_i was higher for the EGCA than the RRUM and for the last two items the CDI_i was higher for the RRUM than for the EGCA. This was not expected because the EGCA should always have higher or equal CDI_i estimates than the same value computed using the parameters estimates from the RRUM analysis. A comparison of the item parameters between the two models for items 11 and 12 did not show any major problem (See Appendices B1 and B2). This result could be due to the misspecification of the Q-matrix. In general, the results support the finding that the EGCA was more discriminating than the RRUM in diagnosing examinees' attributes.

Moreover, because items 1–5 measured Skill 1 and Misconception 1, and items 7–12 measured the same skill and Misconception 2, the sum differences between the EGCA and the RRUM for items 1–5 and items 6–12 were summarized (see Table 16). The results show that differences in $CDI_{\bullet}$ were .86 for the first 5 items and .43 for the last 7 items. Combined with strong positive correlations between the CCRs and the $CDI_{\bullet}$ in the simulation study, the results indicate that the EGCA most likely has an advantage over the RRUM in discriminating the attributes for the first 5 items than for the last 7 items and, as a result, would be predicted to have a CCR. The findings are consistent with previous classification results, which showed that the correlation between the misconception and the number of the options measuring the misconception chosen by the examinees was 0.81 for the EGCA and 0.54 for the RRUM in the first five items compared to 0.85 for the EGCA and 0.76 for the RRUM for the last seven items. The results suggest that the last seven items that were used to measure Skill 1 and Misconception 2 may contain poor distractors, which could lead to a smaller difference between the EGCA and the RRUM concerning $CDI_{\bullet}$.

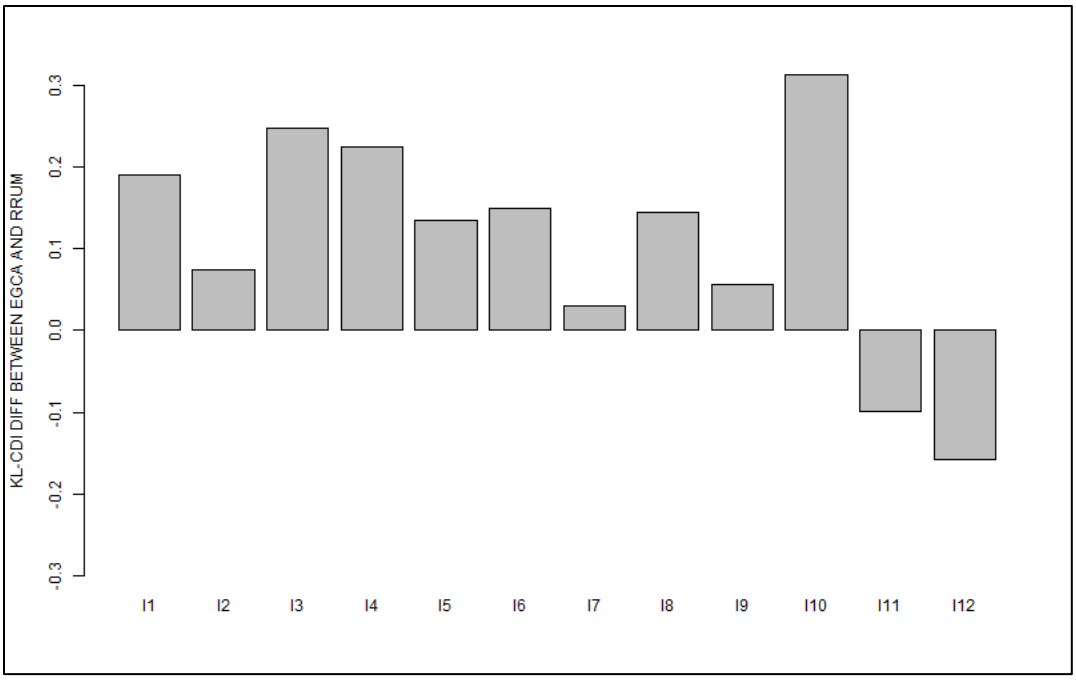

Figure 12. CDI_i Difference Between the EGCA and RRUM by Item.
Note: Each bar indicates the difference (EGCA minus RRUM) of the CDI_i

Table 17. CDI_i Between the EGCA and RRUM by Item

Item No.	EGCA	RRUM	Difference Between Sum of Items 1–5	Difference Between Sum of Items 6–12
1	0.63	0.44		
2	0.84	0.76		
3	0.69	0.44		
4	1.12	0.9		
5	0.49	0.36	.86	
6	0.98	0.83		
7	0.85	0.82		
8	0.62	0.48		
9	0.34	0.28		
10	0.59	0.28		
11	0.67	0.77		
12	0.99	1.15		.43
Sum	8.81	7.51		

CHAPTER V

DISCUSSION

The study investigated the performance of a polytomously scored data analysis versus a dichotomously scored analysis with the same data. The EGCA, a submodel of the GDCM-MC, was first introduced as a polytomous data analysis that could be directly compared to a dichotomous data analysis (the RRUM) without model-induced bias due to misfit. A simulation study was designed to compare the two approaches under a broad set of conditions that included the informativeness of distractor, item quality, sample size, test length, and attribute size. Within the simulation study, parameter-convergence estimations of both the EGCA and RRUM were evaluated to ensure any differences in CCRs between approaches could be attributed to the model and not to the estimation. The results showed that the item parameters of both the RRUM and EGCA did converge in most of the conditions. Even in conditions where there was a small portion of non-convergence, the CCRs and $CDI_{\bullet}$ were still acceptable. Therefore, the results could be meaningfully interpreted.

The study also hypothesized that the EGCA analysis using polytomous data would be better than the RRUM analysis with dichotomous data when using multiple-choice items with informative distractors, although the two approaches are expected to behave similarly when the distractors are not informative. This difference is expected because the EGCA directly models the polytomous responses as opposed to only

right/wrong responses. As a result, the EGCA shows not just that an examinee misses an item, but also how the item was missed. In contrast, the RRUM only models the correct option versus the incorrect option.

The results show that when the item distractors are not informative, both models produce equivalent CCRs (i.e., pCCRs and aCCRs) and item parameters. However, when the distractors are informative, the EGCA produced higher CCRs than the RRUM across all conditions. Furthermore, this difference in CCRs was larger as the distractors were more informative. However, the CCRs of the EGCA increased as the distractors became more informative across all item quality conditions with the exception of the high π and high r condition. Recall that in this condition, rescaling changed the item quality. The largest π values in the strong distractor condition were much smaller than the largest rescaled π values in the weak and non-informative distractor condition because the π values had to be scaled to ensure that the requirement $S_\alpha < 1$ was maintained. Therefore, the CCRs of the RRUM were not the same for the diffferent distractors condition when the values of π and r were high. Despite that, the results confirmed the second research aim, that is, the importance of using the EGCA to model multiple-choice items when the distractors were strongly informative. Additionally, the results showed that that the CCRs of both models would increase when the sample size, test length, and item quality increased and decreased when the number of attributes increased. The CCRs in the good-quality item condition (i.e., high π and low r) were better estimated than those in the low-quality item condition.

As for the test-level discrimination index $CDI_{\bullet}$, the results were consistent with the third research aim. The correlation between a measure of test quality (log of $CDI_{\bullet}$) and the CCRs was relatively high for both the EGCA and the RRUM estimates ($\bar{r}$ = .81 and $\bar{r}$ = .80, respectively). Note that the relationship was also shown to be strong when using the aCCRs. The results were consistent with previous literature, which showed a strong correlation between classification accuracy and the $CDI_{\bullet}$ (Henson & Douglas, 2005). As the number of attributes decreased or the sample size increased, the mean $CDI_{\bullet}$ increased. The effect could be because the test could better differentiate fewer attributes than more attributes given the same amount of information, and that the increase in sample size could increase the test level of discrimination. As the test length increased from 20 items to 40 items, the $CDI_{\bullet}$ doubled because it was a function of test length. In addition, in a simulated condition with high-item quality (i.e. high values of π and low values of r) $CDI_{\bullet}$, values were highest, whereas for poor-quality item condition (low values of π and high values of r), the $CDI_{\bullet}$ was lowest. Also, because the EGCA is essentially equal to the RRUM when using noninformative distractors, it was demonstrated that the EGCA and RRUM had an equal $CDI_{\bullet}$ when there were no informative distractors. Noticeably, when the distractors were informative, the EGCA produced much higher $CDI_{\bullet}$ than the RRUM. The EGCA advantage was even more prominent when distractors were strongly discriminating when compared to weakly discriminating distractors.

The simulation study results could provide useful guidelines in designing diagnostic assessments based on multiple-choice items for assessment developers. When distractors are not informative, both models can be used interchangeably. When distractors have embedded information regarding skills or misconceptions and item quality is high, the EGCA is preferred over the RRUM. In order to have the best classification, it is preferred to have a larger sample size and a longer test with good quality items and good distractors. Previous literature has suggested that using partial correct answers that may not contain all required skills could create good distractors (Ali et al., 2016). Additionally, the distractor-driven assessment used in this study incorporated the student misconceptions in the concept of distractors. However, future research is needed to explore more ways to create strongly informative and less-informative distractors.

A real distractor-driven assessment with informative distractors was used to compare the two models. The profile classification-agreement rate between the two models was 48%, and the classification of each attribute was different. The EGCA classified more students as masters of the first skill and fewer students to master the two measured misconceptions when compared to student results using the RRUM. This could be because the EGCA used the information provided in the distractors, which was ignored by the RRUM. Further exploration of the association between the option classification and selection showed stronger predictive validity evidence of using the EGCA instead of the RRUM. The profile diagnoses associated with the correct option were more correlated with the number of correct options chosen using the EGCA analysis

when compared to the RRUM analysis. The classification of misconceptions was more likely to be associated with misconception options chosen by examinees using the EGCA than the RRUM. The results indicate that the EGCA provides more meaningful classification than the RRUM because the classifications were more aligned with the type of options selected by the examinees. In addition, the EGCA estimation of item parameters produced a higher $CDI_{\bullet}$ than the RRUM analysis. Most of the items have higher a CDI_i for the EGCA than for the RRUM, which indicates that if the model is appropriate, the EGCA is most likely more discriminating than the same assessment when using the RRUM with dichotomous data as long as the additional information was measured by the distractors. The results provide evidence that the use of the EGCA for modeling the distractor-driven assessment would be more helpful in examinee classification when compared with a dichotomized RRUM analysis.

Distractors of multiple-choice items have been shown to influence the item quality. Previous literature shows that the quality of distractors has been correlated with item difficulty and item discrimination. When non-informative distractors were replaced by informative distractors, item difficulty increased even to the level of difficulty of free responses (Ali et al., 2016). If non-informative distractors are added to an item, the item difficulty and discrimination are not affected (Cizek & O'Day, 1994). Similarly, the results of this study showed that non-informative distractors did not add discrimination(i.e., *CDI*) to items using either the dichotomously scored model (RRUM) or the polytomously scored model (EGCA), and including informative distractors in

items that increased the discrimination of the items when the polytomously scored model (EGCA) was used.

Moreover, a previous study showed that data simulated from a polytomous DCM and analyzed by the polytomous model has better classification accuracy than that diagnosed by the corresponding dichotomous DCM when options of test items measured certain attributes (de la Torre, 2009a). While it is expected that modeling the distractors and capitalizing on such additional information will increase the CCRs, this study has a possible confound. Specifically, the study did not consider the dichotomous model misfit, and as a result, it is at least possible that the improved CCRs are partially due to a lack of fit for the dichotomous model. In this study, the EGCA was introduced that could fit both dichotomous and polytomous data when the distractors have no information. The study showed that the correct EGCA option parameters were equivalent to the RRUM item parameters that only used the correct option information. In addition, the study showed that the EGCA was equivalent to the RRUM with respect to the CCRs when no skills or misconceptions are measured by the incorrect options. Furthermore, the results of the study showed that the EGCA provided better classification and test discrimination than the RRUM when the distractors were informative. The previous study showed that the profile CCRs of the polytomous DCM (i.e., MC-DINA) was 20% higher than the profile CCRs of the DINA (de la Torre, 2009a), while the current study demonstrated that, if distractors were constructed intelligently, the EGCA could capitalize on the additional information in a way that the RRUM model ignores. This additional information can increase the profile CCRs by as much as 35% in certain situations.

Previous research has also shown that polytomous scoring that takes into account the partial ability measured by distractors results in a different ability distribution when compared to the ability distribution that results from a dichotomous scoring analysis (Jiao et al., 2012). Although the general location may change for examinees in these cases, it was also found that the normality of the distribution may not be impacted (Grunert et al., 2013). Whether the differences in ability estimation stem from model differences or information from the distractors is unknown.

The real-world data analysis in this study explored the effect on the ability distribution when comparing polytomously scored items versus dichotomously scored items. Although this study focused on a diagnostic model as opposed to a continuous ability model, the results of the real-data study suggest that the polytomous DCM has stronger validity evidence than the dichotomous DCM in modeling a distractor-driven assessment. Specifically, the student profiles associated with the profile-measured correct option was more correlated with the actual selection of the correct answers using the EGCA than the RRUM. Furthermore, the students' diagnosed misconceptions were more correlated with the selected options that measured misconceptions using the EGCA than the RRUM. The findings of this study indicate that classification from the EGCA has more predictability than the classification from the RRUM for distractor-driven assessments.

This study also has certain limitations. There are more non-convergence cases for the EGCA model when the distractors are non-informative than when the distractors are informative. One possibility of having non-convergence cases in the non-informative

distractor cases may be due to the large attribute size because the non-convergence occurred more often in the six attributes condition than the four attributes condition, and it is difficult to estimate a large number of parameters with little information under the non-informative distractor condition. However, the mean CCRs and $CDI_{\bullet}$ of this condition did not appear to be affected by the non-convergence, because the two models have the same values of mean CCRs and $CDI_{\bullet}$ across all conditions. The possible explanation for the good recovery of the CCRs—given the potential convergence issues with the item parameters—could be related to the EGCA. The EGCA is a relatively complicated model, and the identification condition of the Q-matrix is not well understood. It is possible for the method that was used to generate Q-matrices to such matrices that result in a nonidentified item for the EGCA. In these instances, the item parameters may not be uniquely identified (i.e., more than one set of item parameters results in the same predicted probability for examinees). In this case, although convergence of the item parameters would directly be affected, actual examinee classification would not be effected.

In the simulation study, the correct option Q-vectors in the EGCA were used in the RRUM. The misconceptions Q-matrix entries 0s of the EGCA were recoded as 1s (i.e., "lack of the misconception" attribute) and used for the RRUM. However, the skills Q-matrix entries of the EGCA were the same as those of the RRUM. The simulation study only examined the classification of attributes as a whole, not separately. The skills classification may be different from the misconception classification for the EGCA and the RRUM because misconceptions were coded differently in the RRUM Q-matrix,

which may lead to a CCRs' difference in skills and misconceptions between the two models. Further study can examine the classification of skills and misconceptions separately for both simulation and real-data studies.

Additionally, as this study only used the GDCM-MC ERUM kernel as the base model, a future study could also compare other versions of the GDCM-MC (e.g., EDINA, EDINO) and their corresponding dichotomous DCMs (e.g., DINA, DINO). Lastly, the assessment used in this study is limited in that it only measured one skill and two misconceptions and contained five true distractor-embedded, multiple-choice items. Seven out of 12 items used in this study were the selected responses items, and they were restructured to be multiple-choice items, which could affect the generalizability of the results. Future research could also explore different assessments with more items that measure more skills and misconceptions.

In past decades, as large-scale summative educational assessments were required to provide multiple achievement levels (e.g., basic, proficient, or advanced) rather than just pass/fail levels (NCLB, 2001), it became more important to differentiate the information assessed by the test so that detailed feedback could be provided to various stakeholders (e.g., students, teachers). Formative assessment using DCMs can offer richer diagnostic information about student strengths and weaknesses that summative assessment cannot (Cizek, 2001). Therefore, the demand of implementing formative assessment using a DCM framework may be promising. However, longer tests and large sample sizes are needed for dichotomous DCMs to provide accurate diagnoses of students' attributes, because only the correct option attributes are considered in those

cases. The results of the current study show that a polytomous DCM can take into account distractor information and provide as much diagnostic accuracy as the dichotomous DCM with shorter tests and smaller sample sizes. For instance, a 20-item test with strong distractors condition had similar CCRs compared to a 40-item test for the RRUM. The results indicate that a polytomous DCM can be more useful than its analogous DCM informative educational testing.

In summary, formative educational assessments are an essential tool to assess students attributes. However, longer tests may be needed to assess accurate student-attribute information using dichotomous DCMs, which can hinder DCM application. Polytomous DCMs can obtain more accurate diagnoses of students' attributes with shorter tests because they utilize information provided in the distractors. The motivation of this study is not new; polytomous models have been compared to dichotomous models in the past (de la Torre, 2009a; Jiao et al., 2012). Although previous studies have shown that the polytomous models can provide better information about the examinees' attribute(s), results were confounded with model misfits. This study is the first to introduce a submodel of the GDCM-MC, the EGCA. Out of all the item options modeled by the EGCA, the correct option parameters are equivalent to parameters of its analogous dichotomous DCM, which only considers attributes measured by the correct option of the multiple choice items. The EGCA can model and measure any additional distractor information measured by the item, but this is not possible with the analogous DCM. Furthermore, assuming good distractors can be provided, the advantages (i.e., CCRs, and $CDI_{\bullet}$) of the EGCA-ERUM over the RRUM are even higher. Finally, this study provides

a real-world example in which the EGCA-ERUM and the RRUM are compared. The results suggest that the EGCA-ERUM classification has more predictability of correct or misconception options selected by examinees than the RRUM.

REFERENCES

Agresti, A. (2007). *An Introduction to Categorical Data Analysis Second Edition*. Hoboken, NJ: JohnWiley & Sons, Inc.

Andersen, E. B. (1983). Latent trait models. *Journal of Econometrics,* 22. 215-227

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573.

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

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.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability In F.M. Lord & M.R. Novick, (Eds.), Statistical theories of mental test scores (Part 5, pp, 397-497). Reading, MA: Addison-Wesley.

Cai, L., & Angeles, L. (2015). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335.

Chen, J. (2017). Advancing the Bayesian Approach for Multidimensional Polytomous and Nominal IRT Models: Model Formulations and Fit Measures. *Applied Psychological Measurement, 41*(1), 3–16.

Choppin, B. (1983). *A two-parameter latent trait model.* (CSE Report No. 197). Los Angeles, CA: University of California, Center for the Study of Evaluation, Graduate School of Education.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement Issues and Practice, 20,* 19–27.

Cizek, G. J., & O'Day, D. M. (1994). Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement, 54*(4), 861–872.

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

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

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

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.

Fu, Y., Rollins, J., & Henson, R. A. (2016). *Conditions impacting parameter and profile recovery of the NIDA model.* Paper presented at National Council of Measurement in Education Meeting, Washington D.C.

Gelman, A.& Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistics Science, 7,* 457–511.

Grunert, M. L., Raker, J. R., Murphy, K. L., & Holme, T. A. (2013). Polytomous versus dichotomous scoring on multiple-choice examinations: Development of a rubric for rating partial credit. *Journal of Chemical Education, 90*(10), 1310–1315.

Jiao, H., Liu, J., Haynie, K., Woo, A., & Gorham, J. (2012). Comparison between dichotomous and Polytomous Scoring of Innovative Items in a Large-Scale Computerized Adaptive Test. *Educational and Psychological Measurement, 72*(3), 493–509.

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.

Klute, M., Apthorp, H., Harlacher, J., & Reale, M. (2017). Formative assessment and elementary school student academic achievement: A review of the evidence.

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment*, *18*(1), 111–115.

Leighton, J. P., Gokiert, R. J., Cor, M. K., Heffernan, C., Leighton, J. P., Gokiert, R. J., … Heffernan, C. (2010). Teacher beliefs about the cognitive diagnostic information of classroom versus large-scale tests: implications for assessment literacy. Assessment in Education: Principles, Policy & Practice, 17(Feb), 7–21.

Luecht, R. M. (2007). Using information from multiple-choice distractors to enhance cognitive-diagnostic score reporting. In J. P. Leighton, M. J. Gierl, J. P. Leighton, M. J. Gierl (Eds.) , *Cognitive diagnostic assessment for education: Theory and applications* (pp. 319-340). New York, NY, US: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using mulitple sequences. *Statistics Science, 7*(4), 457–511.

Han, K. T., & Paek, I. (2014). A review of commercial software packages for multidimensional irt modeling. *Applied Psychological Measurement*, *38*(6), 486–498.

Haris, A.S., Carr, P. A., & Ruit, K. G. (2016). Validity and reliability of scores obtained on multiple-choice questions: why functioning distractors matter. *Journal of the Scholarship of Teaching and Learning Journal of the Scholarship of Teaching and Learning Josotl.Indiana*, *16*(1), 1–14.

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

Henson, R. A., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*(4), 262–277.

Henson, R. A., Rousso, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement*, *32*(4), 275-288.

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

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

No Child Left Behind Act (2001). Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Naumenko, O., Fu, Y., Henson, R. A., Stout, W., & DiBello, L. (2016). *Generalized DCMs for option-based scoring*. Paper presented at National Council of Measurement in Education Meeting, Washington D.C.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied Psychological Measurement*, *16*(2), 159.

Ozaki, K. (2015). DINA models for multiple-choice items with few parameters: Considering incorrect answers. *Applied Psychological Measurement*. Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, *33*(1), 36–48.

Pachai, M. V, DiBattista, D., & Kim, J. A. (2010). The effect of “None of the Above” on multiple choice questions in a first-year classroom. *The Canadian Journal for the Scholarship of Teaching and Learning*, *6*(3), 1–14.

Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, *33*(1), 36–48.

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

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer-Verlag.

Samejima, F. (1969). Estimation of Latent Ability Using a Response Patern of Graded Scores.

Samejima, F. (1979). *A New Family of Models for the Multiple-Choice Item*.

Sideridis, G., Tsaousis, I., & Al Harbi, K. (2017). Improving measures via examining the behavior of distractors in multiple-choice Tests. *Educational and Psychological Measurement*, *77*(1), 82–103.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference (pp. 82-98).* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Templin, J., & Bradshaw, L. (2013). The comparative reliability of diagnostic model examinee estimates. *Journal of Classification, 30*(2), 251–275.

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

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*(4), 501–519.

Urry, V. W. (1971). Approximation methods for the item parameters of mental test models.

von Davier, M. (2005). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287–308.

APPENDIX A

RANGE OF RESCALED r ACROSS DIFFERENT CONDITIONS

N	*I*	*K*	*DQ*	H/H	H/L	L/H	L/L
1000	**20**	**4**	**None**	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Weak	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Strong	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
		6	**None**	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Weak	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Strong	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
	40	**4**	**None**	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Weak	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Strong	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
		6	**None**	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Weak	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Strong	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
2000	**20**	**4**	**None**	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Weak	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Strong	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
		6	**None**	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Weak	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Strong	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
	40	**4**	**None**	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Weak	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Strong	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
		6	**None**	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Weak	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)
			Strong	(0.20~0.50)	(0.05~0.20)	(0.20~0.50)	(0.05~0.20)

Note: *N* is the sample size, *I* is the test length, *K* is the number of attributes, and *DQ* is the quality of the distractors. H/H is high π and high r condition, H/L is high π and low r condition, L/H is low π and high r condition, and L/L is low π and low r condition.

APPENDIX B1

ESTIMATED PARAMETERS USING THE EGCA

Item No.	Option No.	π	r_1	r_2	r_3
1	1	0.564	0.073	0.045	-
1	2	0.65	-	0.142	-
1	3	0.615	0.159	-	-
2	1	0.926	0.284	0.136	-
2	2	0.281	-	0.047	-
2	3	0.507	0.127	-	-
3	1	0.683	0.222	0.039	-
3	2	0.887	-	0.214	-
3	3	0.277	0.13	-	-
4	1	0.944	0.102	0.071	-
4	2	0.559	-	0.006	-
4	3	0.518	0.31	-	-
5	1	0.554	0.228	0.054	-
5	2	0.643	-	0.155	-
5	3	0.381	0.065	-	-
6	1	0.985	0.711	-	0.174
6	2	0.818	-	-	0.012
6	3	0.009	0.249	-	-
7	1	0.862	0.404	-	0.12
7	2	0.675	-	-	0.01
7	3	0.257	0.458	-	-
8	1	0.703	0.26	-	0.114
8	2	0.785	-	-	0.011
8	3	0.137	0.869	-	-
9	1	0.562	0.277	-	0.17
9	2	0.362	-	-	0.097
9	3	0.53	0.482	-	-
10	1	0.616	0.225	-	0.195
10	2	0.504	-	-	0.012
10	3	0.778	0.456	-	-
11	1	0.921	0.448	-	0.16
11	2	0.463	-	-	0.105
11	3	0.085	0.138	-	-
12	1	0.961	0.415	-	0.097
12	2	0.641	-	-	0.03
12	3	0.21	0.031	-	-

APPENDIX B2

ESTIMATED PARAMETERS USING THE RRUM

Item No.	$\boldsymbol{\pi}$	$\boldsymbol{r_1}$	$\boldsymbol{r_2}$	$\boldsymbol{r_3}$
1	0.665	0.301	0.073	-
2	0.962	0.752	0.153	-
3	0.765	0.452	0.128	-
4	0.922	0.655	0.052	-
5	0.657	0.397	0.123	-
6	0.989	0.938	-	0.246
7	0.980	0.473	-	0.192
8	0.812	0.344	-	0.162
9	0.587	0.574	-	0.139
10	0.609	0.573	-	0.158
11	0.982	0.628	-	0.209
12	0.990	0.746	-	0.083