CONTINUOUS RESPONSE IN COGNITIVE DIAGNOSIS MODELS: RESPONSE TIME MODELING, COMPUTERIZED ADAPTIVE TESTING, AND Q-MATRIX VALIDATION

BY NATHAN DANIEL MINCHEN

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Education

Written under the direction of

Jimmy de la Torre

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2017

ABSTRACT OF THE DISSERTATION

Continuous Response in Cognitive Diagnosis Models: Response Time Modeling, Computerized Adaptive Testing, and Q-Matrix Validation

by Nathan Daniel Minchen

Dissertation Director: Jimmy de la Torre

At present, many cognitive diagnosis models (CDMs) have been developed for dichotomous response, several of which have been extended to handle polytomous response. CDMs to handle continuous response, however, have not been extensively explored beyond the recently proposed continuous deterministic inputs, noisy "and" gate (C-DINA) model and its generalized version. The studies that comprise this dissertation aim to extend model development in the context of continuous response and to address several key issues that arise from its use in CDM.

In the first study, a hierarchical framework is employed for using response time to improve examinee ability estimation and classification accuracy. Under this framework, response time and response accuracy are construed as arising from separate continuous, possibly correlated, unidimensional latent variables.

A higher-order attribute specification is used to link the general ability to the probability of mastering certain attributes. Results show that both examinee classifications and higher-order ability estimation can be improved by using response time. A real data example is included to demonstrate the viability of the method.

In the second study, a new item selection algorithm is presented for computerized adaptive testing applications that use continuous response CDMs. The algorithm uses the Jensen-Shannon divergence, which quantifies the total degree of dissimilarity in a set of two or more probability distributions, as an item selection algorithm. Results demonstrate that the method typically outperforms random item administration with respect to both classification accuracy and test efficiency. A real data example shows that an existing test could be shortened considerably while still producing a high level of classification agreement with the original.

In the final study, a new Q-matrix validation procedure is proposed for continuous response CDMs. The method presented is designed to work with a generalized continuous response model, and is based on a weighted least squares regression. The simulation study shows that the method performs increasingly well as item quality increases. The method was also applied to an existing dataset, with results confirming most of the entries in the existing Q-matrix.

Table of Contents

Abstract ii

List of Tables vi

List of Figures viii

1. Introduction 1

1.1. Cognitive Diagnosis Models 1

1.2. Continuous Responses 3

1.3. Objectives 5

References 8

2. Study I: Modeling Response Time in Cognitive Diagnosis . . . 10

2.1. Introduction 11

2.2. Continuous Response Measures 12

2.3. Cognitive Diagnosis Models 16

2.4. The LN+HO-DINA Model 19

2.5. Design and Analysis 25

2.6. Results 28

2.7. Real Data Example 35

2.8. Summary and Discussion 39

References 42

3. The Jensen-Shannon Divergence as an Item Selection Algorithm in CD-CAT 46

3.1. Introduction 47
3.2. Cognitive Diagnosis Computerized Adaptive Testing 49
3.3. Cognitive Diagnosis Models 54
3.4. The Jensen-Shannon Divergence 57
3.5. The JSD as an Item Selection Index 58
3.6. Design and Analysis 61
3.7. Results 65
3.8. Real Data Example 73
3.9. Discussion and Conclusion 75
References 78

4. Study III: A Q-Matrix Validation Method for Continuous Response CDMs 81
4.1. Introduction 82
4.2. Cognitive Diagnosis Models 84
4.3. Q-Matrix Validation 88
4.4. Proposed Method: Weighted Least Squares Q-Matrix Validation Procedure 90
4.5. Results 101
4.6. Real Data Example 105
4.7. Discussion and Summary 107
References 111

5. Conclusion 115
References 120

List of Tables

2.1. Simulation Study Q-Matrix 27

2.2. RMSEs for Model Parameters 30

2.3. Mean Classification Rates 31

2.4. RMSE and Bias for θ for Low Quality Items 32

2.5. RMSE and Bias for $\hat{\theta}$ for Medium Quality Items 33

2.6. RMSE and Bias for $\hat{\theta}$ for High Quality Items 33

2.7. Q-Matrix for the Balance Scale Data 35

3.1. JSD Example (1) 61

3.2. JSD Example (2) 61

3.3. Mean Classification Rates by Attribute Pattern: 5-Item Tests . . 66

3.4. Mean Classification Rates by Attribute Pattern: 10-Item Tests . . 67

3.5. Average Number of Items Administered by Attribute Pattern . . 68

3.6. Average Efficiency of the JSD CAT Algorithm by Attribute Pattern 68

3.7. Balance Scale Data CAT Results 74

4.1. Forming the Candidate Posterior Distribution with q-vector [110] 92

4.2. Simulation Study Q-matrix 97

4.3. C-DINA WLS Mean True Positive Rates 102

4.4. C-DINA WLS Mean False Positive Rates 102

4.5. C-DINA Max R^2 Mean True Positive Rates 103

4.6. C-G-DINA WLS Mean True Positive Rates 104

4.7. C-G-DINA WLS Mean False Positive Rates 105

4.8. Reduced Q-matrix for the Balance Scale Data 106

4.9. SSE From Real Data Example 107
4.10. Real Data True and False Positive Rates 108

List of Figures

2.1. Dichotomous- and Continuous-Response CDMs 18
2.2. Acyclic Diagram of the LN+HO-DINA Model 21
2.3. Higher-Order Ability Distribution by Latent Class 36
2.4. Relationship between Higher-Order Ability Estimates 37
3.1. Various Lognormal Distributions with Identical Means 53
3.2. Binary and Continuous CDMs 57
3.3. JSD Example: C-DINA Item 60
3.4. Overall Item Usage: C-DINA 71
3.5. Overall Item Usage: C-G-DINA 72
3.6. Number of Items Administered by Minimax Condition 74
4.1. Example Items of Low and High Discrimination Under Various CDMs 98

Chapter 1
Introduction

1.1 Cognitive Diagnosis Models

Cognitive diagnosis modeling, which is the statistical technique used to extract diagnostic information from cognitively diagnostic assessments (CDAs; de la Torre & Minchen, 2014), is a psychometric framework for formative assessment that stands in contrast to traditional testing frameworks such as item response theory (IRT) and classical test theory (CTT). Whereas the goal in the latter two frameworks is generally to provide summative feedback for the purpose of rank-ordering examinees, the goal of cognitive diagnosis models (CDMs) is to provide timely diagnostic feedback with respect to a set of discrete skills so that teachers are equipped with specific information about their students' knowledge states. Assessments with this purpose are in great demand (DiBello & Stout, 2007).

In contrast, IRT- and CTT-based tests typically provide information on a small number of broadly-defined abilities (Junker & Sijtsma, 2001) - many times just a single ability - such as mathematics or reading. Thus, it may be difficult to know how to improve the learning process. De la Torre (2012) and de la Torre and Minchen (2014) have discussed the use of item maps as tools to make diagnostic inferences in the context of traditional assessments. They concluded that doing so may result in drawing diagnostic inferences that are conflated with the nuances and idiosyncrasies of the questions themselves, thereby jeopardizing the validity of the inferences, largely because the questions have not been designed to provide information at that level.

CDMs aim to provide feedback on a multivariate set of discrete skills or attributes. The number of skills measured can be large, but is generally recommended to be less than 10 (Tatsuoka et al., 2016; DiBello, Roussos, & Stout, 2007). The levels of the attributes are generally conceptualized as skills being "present" or "not present." In educational settings, this typically translates into categories of mastery versus non-mastery. When CDMs are used in medical settings, these categories can be interpreted as the absence or presence of symptoms (Templin & Henson, 2006; de la Torre, van der Ark, & Rossi, 2015). For the level of measurement to be reduced to a simple dichotomy without losing meaningful information, it should be clear that this set of skills must be much finer-grained than the ability or abilities measured in traditional assessments.

For CDAs to have maximum potential, they should be intentionally designed from their very inception to be diagnostic (de la Torre & Minchen, 2014), a process that is lengthy and involves many steps. Tjoe and de la Torre (2013b) outline the steps involved in the attribute-validation process, which include a literature review, conferring with relevant experts (e.g., researchers, teachers, and psychometricians), and think-aloud problem solving sessions with students of various academic levels. Tjoe and de la Torre (2013a) also outline the item-writing process for such an assessment. Although it would be convenient to simply apply a CDM to an ordinary test in an effort to make diagnostic inferences, such efforts have had only limited success (de la Torre & Karelitz, 2009).

One of the distinguishing features of CDMs is the way that they model the interaction between examinees' skill patterns and the attributes measured by the item. The probability of a correct response for an examinee on a given item is a function of the examinee's attribute pattern and the attributes required for the item, with some cognitive process governing this function, which is defined by the specific CDM. For example, both the deterministic inputs, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) and the deterministic inputs, noisy

"or" gate (DINO; Templin & Henson, 2006) models partition examinees into two groups for each item, but the groups are defined differently. To respond correctly, the DINA requires that examinees possess all required attributes, wheres the DINO only requires that examinees posess at least one of the required attributes. Otherwise, examinees are expected to answer incorrectly. Generalized models exist as well (e.g., de la Torre, 2011; von Davier, 2005) that relax the assumptions of simplified models, of which the DINO and DINA are examples.

1.2 Continuous Responses

Many advancements in the CDM literature have assumed a dichotomous, or sometimes a polytomous, item response, but continuous response CDMs have not been explored extensively. Thus, the focus of this research is to advance continuous response modeling in CDM. Continuous response is increasingly common and offers great potential by expanding applicability of CDAs, but it also presents a wide range of new challenges for the technical components of cognitive diagnosis modeling. Before exploring the objectives of this research, we will first discuss some of examples of continuous response.

Perhaps one of the most abundant continuous responses available today is response time, which is easily captured in computer-based exam delivery. Response time could be used as the intended measurement, or it could be used as ancillary information that is obtained in the process of measuring a construct for which response accuracy is of primary interest. It may also provide additional insight when there are clearly-defined developmental steps that need to be executed in order to solve a problem (van der Maas & Jansen, 2003).

Another type of continuous response is to simply place a mark on a line segment. Such a response type can be viewed as a generalization of a graded-response as the number of categories becomes infinite (Samejima, 1973). Noel

and Davier (2007) introduced a model for analyzing this type of response. They applied their model to a real data example in which the responses were levels of agreement (0-100%) with a statement. For example, if the mark was placed one-quarter of the distance from the left end of the segment, then the response would be approximately 0.25, or 25% agreement. Noel (2014) extended this model to accommodate responses that have an unfolding nature.

Another type of continuous response is "probability testing," in which examinees report the probability that an answer is correct, with possible marks ranging from 0 to 100. Probability testing has the potential to reveal the following types of knowledge: "full knowledge," "partial knowledge," "partial misinformation," "full misinformation," and the "absence of information" (Ben-Simon, 1997, p. 69-70). In the full information setting, examinees report a probability of 1 for the answer that is correct, and 0 for all others, demonstrating complete knowledge. With partial information, examinees report a nonzero probability for the correct answer, but also report nonzero probabilities for other incorrect answers. Assigning a probability of zero to the correct answer represents full misinformation, whereas assigning equal probabilities to all answers reflects the absence of information. Other variations on this method exist and are discussed or cited in Ben-Simon (1997).

More generally, continuous responses have numerous applications in measurement models, and the CDA/CDM assessment framework has much to offer in its diagnostic potential. Thus, it is prudent to further the body of research that explores the viability of continuous responses in the context of CDM. To that end, the objectives of this research are discussed next.

1.3 Objectives

Many of the current cognitive diagnosis models (CDMs) have been developed for dichotomous response, several of which have been extended to handle polytomous response (e.g., de la Torre, 2009; Ma & de la Torre, 2016). CDMs to handle continuous response, however, have not been extensively explored beyond the work of Minchen, de la Torre, and Liu (in press), in which the continuous deterministic inputs, noisy "and" gate (C-DINA) model is developed, and Minchen and de la Torre (2016), in which the C-DINA model was generalized.

To expand the body of research in this area, this dissertation aims to extend model development and methodology as they pertain to continuous response in CDM. To that end, the objectives of the research presented herein are three-fold: (1) to adapt a hierarchical framework (van der Linden, 2007), which has been used in IRT for jointly modeling response time and response accuracy, for use in CDM, (2) to propose a new item selection index for use in cognitive diagnosis computerized adaptive tests (CD-CAT) in the context of a generalized continuous response model (Minchen & de la Torre, 2016), and (3) to introduce a statistical Q-matrix validation procedure for verifying the attributes that are assumed to be measured by each item in the context of continuous response models. Additional details for each objective are now discussed in turn.

In the first study, a hierarchical framework (van der Linden, 2007) is employed for using response time to improve both examinee classifications and ability estimates. Using this framework, response time and response accuracy arise from separate, but correlated, latent variables. On the response time side of the model, a lognormal model (van der Linden, 2006, 2007) is used, and a continuous, unidimensional latent variable governs the speed at which examinees work. On the response accuracy side of the model, a higher-order attribute formulation (de la Torre & Douglas, 2004) is used to link the general ability to the probability of

mastering each attribute. A simulation study was carried out, and factors investigated in the simulation study were item quality, sample size, and the relationship between the higher-order ability and the attributes. Classification accuracy of attributes patterns and estimation accuracy of model parameters and higher-order abilities were analyzed. Finally, the method was applied to a real data example.

In the second study, a new item selection algorithm is presented for computerized adaptive testing applications that use continuous response CDMs. In a simple continuous response model such as the C-DINA (Minchen, de la Torre, & Liu, in press), it may be possible to adapt existing selection algorithms. Doing so, however, presents certain challenges, which may be serious, and are discussed. When using the continuous generalized DINA (C-G-DINA; Minchen & de la Torre, 2016), adaptations from the dichotomous response models can result in a loss of information. Thus, the Jensen-Shannon divergence (JSD; Lin, 1991), which quantifies the total degree of dissimilarity in a weighted set of two or more probability distributions, is proposed for use as an item selection algorithm. The weights used in the item selection algorithm are examinees' current posterior distribution probabilities. Both fixed- and variable-length tests were administered in the simulation study, the latter of which used the level of certainty in the posterior distribution as the stopping rule. Performance of the algorithm in fixed-length tests were evaluated using classification accuracy will be examined, whereas the number of items administered will be used to evaluate variable-length tests. Item usage was be examined, and the algorithm was applied to a real data example.

In the final study, a new Q-matrix validation procedure is proposed. Although an array of such techniques exist, they have largely been developed in the context of a dichotomous item responses. For a variety of mathematical reasons, which will be discussed, the existing methods may not be appropriate for continuous response data. The method presented in this study is designed to work with the C-G-DINA model, and is based on a weighted least squares regression technique.

The method is exhaustive in the sense that each q-vector is evaluated for each candidate item. The posterior distributions under the candidate q-vector are obtained and used as weights to partition the responses into the latent groups created under that q-vector. The coefficient of determination, R^2, is used to describe the proportion of explained variance in the responses under each grouping. The fully-specified q-vector will produce the highest R^2, and the q-vector with the smallest number of specifications that does not have a significantly different R^2 is chosen. In the simulation study, the level of misspecification in the Q-matrix, item discrimination, generating model were manipulated. To evaluate the method, the number of times a correction is made to an incorrect q-vector (true positive) was tabulated, as well as the number of times a correct q-vector was made wrong (false positive). Finally, the method was also applied to a real data example for which the Q-matrix has been established.

References

Ben-Simon, A., Budescu, D. V., & Nevo, B. A. (1997). Comparative study of measures of partial knowledge in multiple choice tests. *Applied Psychological Measurement, 21*, 65-88.

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

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

de la Torre, J. (2012). Application of the DINA Model Framework to Enhance Assessment and Learning. In M. Mok (Ed.), *Self-directed learning oriented assessments in the Asia-Pacific* (pp. 92-110). New York: Springer.

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

de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement, 46*, 450-469.

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

de la Torre, J., van der Ark, L. A., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, 1-16.

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

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of statistics, 26*, 979-1030.

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

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

Lin, J. (1991). Divergence measures based on the Shannon entropy. *Information Theory, IEEE Transactions on, 37*(1), 145-151.

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

Minchen, N. D., & de la Torre, J. (2016, July). *The continuous G-DINA model and the Jensen-Shannon divergence.* Paper presented at the International Meeting of the Psychometric Society, Asheville, NC.

Minchen, N. D., de la Torre, J., & Liu, Y. (in press). A cognitive diagnosis model for continuous response. *Journal of Educational and Behavioral Statistics.*

Noel, Y., & Davier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement, 31*, 47-73.

Noel, Y. (2014). A beta unfolding model for continuous bounded responses. *Psychometrika, 79*, 647-674.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*, 203-219.

Tatsuoka, C., Clements, D. H., Sarama, J., Izsak, A., Orril, C. H., de la Torre, J., & Khasanova, E. (2016). Developing workable attributes for psychometric models based on the Q-matrix. *Psychometric Methods in Mathematics Education: Opportunities, Challenges, and Interdisciplinary Collaborations*, 73-96.

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

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*, 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287-308.

van der Maas, H. L. J., & Jansen, B. R. J. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology, 85*, 141-177.

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

Chapter 2

Study I: Modeling Response Time in Cognitive Diagnosis

Abstract

At present, examinees classifications in cognitive diagnosis are typically based solely on response accuracy. However, response time is another source of information that contains potentially valuable information, and may be easily obtained in computer-based testing settings. Although a continuous response cognitive diagnosis model that can be used in response time modeling was recently proposed, examinee classifications are still made on the basis of a single response type. This paper proposes a framework in which both response accuracy and response time can be used jointly to classify examinees. Specifically, a hierarchical framework is used in conjunction with a higher-order deterministic inputs, noisy "and" gate model. In this framework, latent variables governing speed and attribute mastery are assumed to be related, thus providing a bridge for response time to assist in the estimation of both the higher-order general abilities and the attribute patterns.

Keywords: cognitive diagnosis models, continuous response, response time, DINA model, lognormal

2.1 Introduction

Although response time is a readily-available source of information, its applications in cognitive diagnosis modeling have been limited. The recently proposed continuous deterministic inputs, noisy "and" gate (C-DINA; Minchen, de la Torre, & Liu, in press) model allows for examinee classifications to be made on the basis of a continuous response, of which response time is an example; however, this model, like many others, only uses a single response type to classify examinees. Also, Finkelman et al. (2014) present a method of incorporating response time into a computerized adaptive testing (CAT) selection algorithm for cognitive diagnosis. Their model, however, does not explicitly use response time to improve the estimation of ability, θ. To the extent that classification is improved, it is through the choice of items that are administered, not because the response time improves estimation of θ.

The goal of this study is to examine whether the estimation of examinees' general abilities and attribute patterns can be improved *directly* by using response time to assist in the estimation of θ. To this end, a new methodology for simultaneously modeling both response time (RT) and response accuracy (RA) within the context of CDM is presented. The proposed model, which was originally presented by Minchen and de la Torre (2016), rests on van der Linden's (2007) hierarchical framework. In this application, a cognitive diagnosis model (CDM) with a higher-order attribute structure (de la Torre & Douglas, 2004) is used as the model for the RA, and a lognormal model (van der Linden, 2006, 2007) is used as the model for the RT. The higher-order attribute distribution assumes that the mastery of attributes depends probabilistically on a higher-level general ability.

2.2 Continuous Response Measures

Many psychometric models are designed to work with multiple-choice items, and thus responses are frequently conceptualized as either dichotomous or polytomous in nature. Polytomous models (e.g., de la Torre, 2009; Ma & de la Torre, 2016) can be applied to multiple-choice questions, but they also can be applied to constructed response questions. Polytomous psychometric models can also allow for partial credit-scoring (e.g., de la Torre, 2010)

One way of viewing the nature of a continuous response is that such a format allows for infinitely many possible answer choices, and in that way is the generalization of a polytomous format as the number of response categories tends toward infinity (Samejima, 1973). There are a variety of types of tasks that may be best captured by a response that is continuous in nature. One type of response is simply the placement of a mark at some point along a continuum, indicating some degree of endorsement. A second type of continuous response is probability testing, in which respondents report what they believe to be the probability of their chosen response being correct. Another form of probability testing is for respondents to report what they believe to be the probability of each response alternative being correct. The third, and perhaps most popular use of continuous response, is response time.

With many testing programs utilizing a computer-based format, some of which are also adaptive, response time has become effortless to capture, and is essentially free ancillary information. Response times may also be of interest in and of themselves in applications that measure constructs such as reaction time, or in settings in which response time may be a function of developmental steps (e.g., van der Maas & Jansen, 2003). To the extent that response time can provide profitable information about examinees, its use could potentially improve a variety of aspects of testing.

Before discussing the various applications of response time in computerized testing, however, it is prudent to introduce van der Linden's (2007) hierarchical framework, which has been a monumental advancement in this niche, and on which many other applications, including this research, are based. Briefly, his framework posits separate latent variables to account for the students' abilities and work speeds. The ability side of the model is flexible and can accommodate any standard item response model that assumes a unidimensional ability variable, θ. The response time side of the model employs a lognormal model (van der Linden, 2006), which is a function of the item-specific structural parameters and the latent variable τ, which represents the speed intensity at which students work. The joint distribution of θ and τ is bivariate normal. This model also assumes that both θ and τ remain constant throughout the test. Van der Linden and Glas (2010) extended this model to allow for violations of conditional independence in the responses or response times. They found that parameter and ability estimates improved by using their model when conditional independence was violated.

A number of researchers have developed models that use response time as a way to detect responses that differ from what would be expected based on an examinee's performance on the remainder of the test (van der Linden & Guo, 2008; van der Linden & van Krimpen-Stoop, 2003). Such responses are often referred to in the literature as *aberrant responses*. Van der Linden and van Krimpen-Stoop (2003) discuss the use of residuals to detect such aberrant responses. They use residuals from both response accuracy and response time models and find that using response time is superior to using response accuracy for detecting aberrance. However, they do not explicitly model a relationship between examinees' speeds and abilities. Van der Linden and Guo (2008) show that residuals based on actual responses will be large for items whose difficulties are similar to the examinee's ability, which occur frequently in a computerized adaptive test, but contend that response times are not subject to the same relationship with difficulty. They

also use a hierarchical model (van der Linden, 2007) to account for the joint relationship between speed and accuracy.

Response times have also been used extensively in CAT. Fan, Wang, Chang, and Douglas (2012) develop the *maximum information per time unit* selection algorithm, in which the ratio of the item's information at $\hat{\theta}$ and its time required to completion is maximized. Using their algorithm, highly informative items are desired, but longer expected completion times reduce the likelihood that they will be chosen (Fan et al., 2012). In an effort to control item exposure, which is not naturally controlled through maximum information criteria, they also propose an adjustment to the alpha-stratified with difficulty blocking algorithm (Chang, Qian, & Ying, 2001) by penalizing items based on how long they take to complete (Fan et al., 2012). Finkelman, Kim, Weissman, and Cook (2014) use response time in conjunction with response accuracy in a CDM to develop an item selection algorithm that is based on Fan et al.'s (2012) idea of maximizing information per time unit. However, they do not model the relationship between speed and ability. In a slightly different application, van der Linden (2008) shows that the inclusion of response times into a hierarchical framework (van der Linden, 2007) can improve item selection by virtue of the improvement in ability estimation. Finally, Sie, Finkelman, Riley, and Smits (2015) and van der Linden (2009) have employed stopping rules in computerized tests whereby the test is terminated if all remaining items have a probability greater than some threshold of causing the test to exceed its time limit.

Ferrando and Lorenzo-Siva (2007) introduce a framework for modeling response time that is based on the distance-difficulty (DD) hypothesis, which assumes that there is more uncertainty in binary personality item responses as the item location gets closer to the examinee location. In an item response theory (IRT) context, this uncertainty can be thought of as the variance of the response,

X_{ij}, which is computed as $Var(X_{ij}) = p_{ij}(1 - p_{ij})$, where p_{ij} represents the probability of a correct response for examinee i on item j. Assuming no guessing parameter, this quantity is maximized at $p = 0.5$, which occurs when the item's threshold and the examinee's latent trait are equal. In the context of personality testing, however, the distance between the person and item locations can be thought of as difficulty; here, distance and difficulty have an inverse relationship. This distance is the fundamental component of the response time model, but it does not have a monotonic relationship with the latent variable for speed, thus the monotonicity assumption on which van der Linden's (2007) hierarchical framework is built would be violated (Ranger & Kuhn, 2012).

In Ferrando and Lorenzo-Siva's (2007) model, the item response model and response time model parameters are estimated independently of one another. They are then treated as fixed and used in the response time portion of the model. Therefore, although the authors demonstrated that the model can improve the precision of ability estimation, the response times do not aid in the estimation of the parameters in the item response model (Ranger & Kuhn, 2012). Ranger and Kuhn (2012) developed a new estimation strategy whereby all parameters are jointly estimated. In doing so, they found that the estimation of parameter estimates, in addition to the ability estimates, were improved.

Response times clearly have a variety of applications in psychometrics. Much of the model development, however, has been in the context of IRT, with only limited developments in CDM (e.g., Minchen, de la Torre, & Liu, in press; Finkelman et al., 2014). Thus, more work needs to be done in continuing to develop CDMs that utilize continuous responses. Continued model development will also necessitate research in other areas, such as CAT, and Q-matrix validation, a model selection, which are important applications in CDM.

2.3 Cognitive Diagnosis Models

In contrast to the goal of traditional assessments, which are typically used to rank-order students, cognitively diagnostic assessments (CDAs; de la Torre & Minchen, 2014) aim to provide diagnostic information on examinees. Traditional assessments often use IRT or classical test theory (CTT) to estimate examinees' latent trait levels on a continuous spectrum, whereas CDAs use cognitive diagnosis models (CDMs) to estimate the class membership of each examinee. In educational settings, classes represent various combinations of skills that examinees have mastered. It should be noted that, in some settings, not all combinations will exist. For example, in the case where attribute A_2 requires the successful implementation of attribute A_1, classes in which examinees possess attribute A_2 but not A_1 will not exist. Such situations impose restrictions on the attribute space, and a variety of restrictions exist.

The Q-matrix (Tatsuoka, 1983) is frequently used in CDM applications to identify which skills are being measured in each item. The matrix is of dimension $J \times K$, where J represents the number of items on the test, and K represents the number of attributes being measured. For dichotomous attributes, entries in the Q-matrix are either 1 or 0, indicating that item j either requires or does not require the use of attribute k, respectively. Correspondingly, the examinee-level latent variable is a Boolean vector of length K, in which entries of 1 and 0 denote that examinee i possesses or does not posess the k^{th} attribute, respectively. Assuming an unrestricted attribute space, there will be $L = 2^K$ possible attribute patterns. The interaction of $\boldsymbol{\alpha}_i$ and $\boldsymbol{q}_j$, along with other quantities such as the characteristics of the item, govern the probability of a particular response.

This paper employs the deterministic inputs, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model, which has become commonplace in the CDM literature. Thus, only a brief review will be offered. At the core of the

DINA model is the latent response variable, η_{ij}, which takes on a value of 0 if the examinee is missing any of the item's required skills and 1 otherwise. With $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$, the probability of a correct response can be written as

$$P(X_{ij} = 1|\boldsymbol{\alpha}_i, s_j, g_j) = (1 - s_j)^{\eta_{ij}}(g_j)^{1-\eta_{ij}}, \tag{2.1}$$

where g_j and s_j are the guessing and slip item parameters, which are interpreted as the probability of $\eta = 0$ examinees providing a response of 1, and an $\eta = 1$ examinees providing a response of 0 to item j, respectively. These parameters capture aberrations from the model; the larger their values, the lower the quality of the item. De la Torre (2008) defined item quality for the DINA model as the sum of g and $1 - s$, where increasing values of the quantity represent decreasing item quality. Although the DINA model may be overly simple (de la Torre, 2011; Henson & Douglas, 2005), one of its key advantages is its parsimony and relative ease of interpretation.

The DINA model serves as a basis for understanding other CDMs. For example, the generalized-DINA (de la Torre, 2011) model relaxes the strict assumption that only examinees possessing all required attributes respond correctly. Instead, the G-DINA model can estimate unique probabilities for each latent group, where the latent group is defined as the subset of attributes that are required by a particular item.

The C-DINA (Minchen, de la Torre, & Liu, in press) and the continuous generalized (C-G-DINA; Minchen & de la Torre, 2016) are the continuous-response analogs to the DINA and G-DINA models, respectively. Figure 2.1 shows an example of each of these models for an item that requires two attributes. Note that the variances of the latent groups in the continuous-response models on the right side of the plot are similar, but this need not be the case.

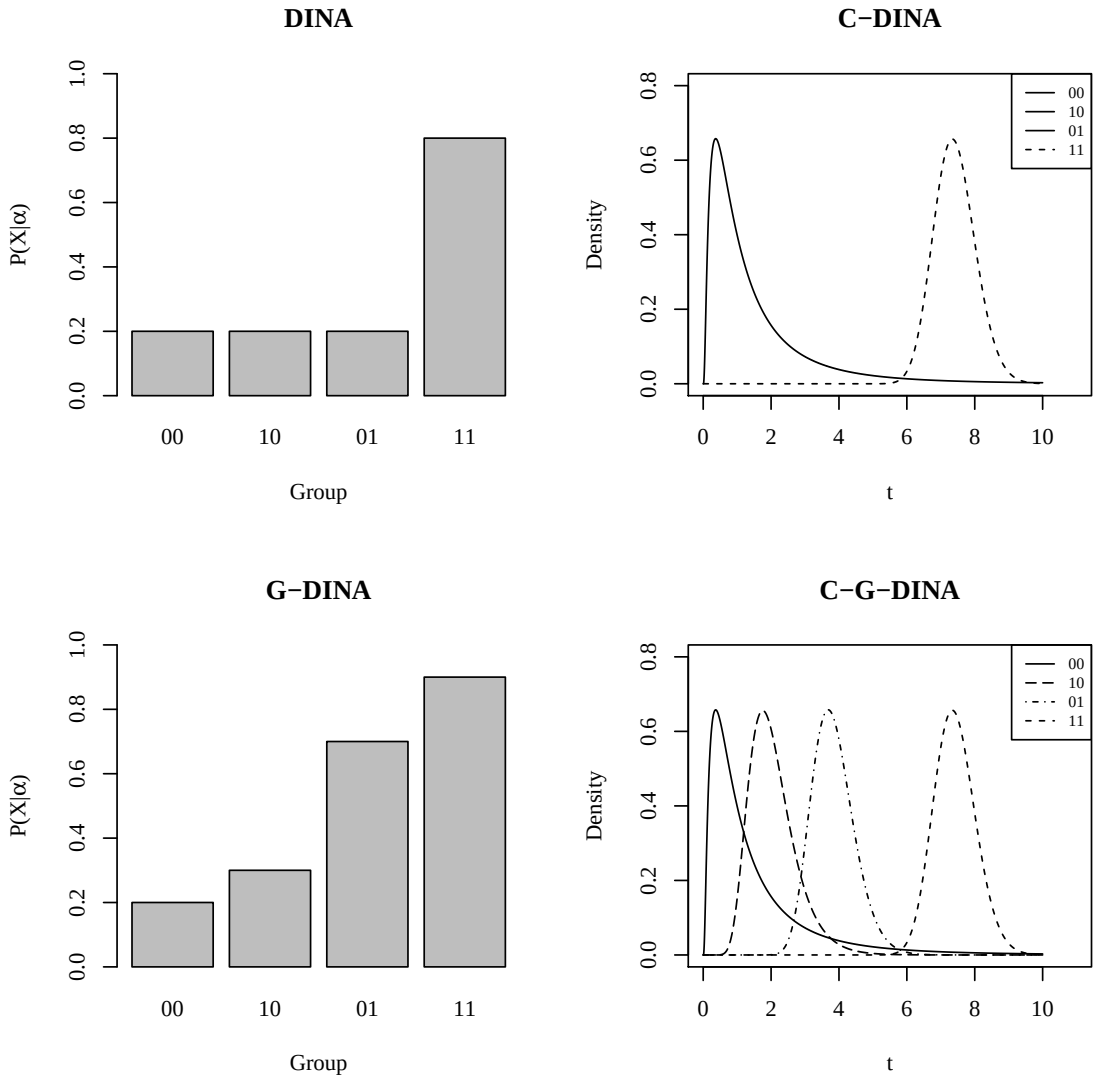

Figure 2.1: Dichotomous- and Continuous-Response CDMs

The Joint Distribution of the Attributes

De la Torre and Douglas (2004) discussed several possibilities for modeling the structure of the attribute space, which can be thought of as the joint distribution of attributes. Of those they discussed, the saturated and higher-order joint distributions were used in this paper. In the saturated model, all permutations of the K-length binary attribute vector, $\boldsymbol{\alpha}$, are permissible, requiring a total of

$2^K - 1$ parameters. In the higher-order model, a general ability is related to the likelihood of possessing certain attributes. Specifically, each $\boldsymbol{\alpha}_i$ is modeled as a probabilistic function of θ_i, both of which are latent variables, using an IRT model, where θ_i takes on the same meaning as it does in IRT and each of the attributes functions like an item. Other joint distributions exist, but will not be discussed here (de la Torre, Hong, & Deng, 2010; Leighton, Gierl, & Hunka, 2004).

De la Torre and Douglas (2004) present the higher-order DINA (HO-DINA) model, in which the higher-order IRT model and the item-level DINA model are combined. First, the joint distribution of the attributes is given by

$$P(\boldsymbol{\alpha}_i|\theta_i) = \prod_{k=1}^{K} \frac{\exp(\lambda_{0k} + \lambda_1\theta_i)}{1 + \exp(\lambda_{0k} + \lambda_1\theta_i)}, \tag{2.2}$$

where λ_{0k} are the intercept parameters associated with attribute k, and λ_1 is the slope parameter that governs the strength of the relationship between θ and $\boldsymbol{\alpha}$ in the same way that the discrimination parameter governs the strength of the relationship between θ and $\boldsymbol{X}$ in the Rasch family of IRT models. In this formulation, the attributes are expressed as a function of θ, where the function is a one-parameter logistic IRT model in slope-intercept form. The probability of a correct response is simply the DINA IRT given in Equation 4.1. One advantage of the HO-DINA model is that it provides both θ estimates and $\boldsymbol{\alpha}$ classifications (de la Torre & Douglas, 2004). It should also be noted that the item-level portion of the model can easily be replaced with any CDM.

2.4 The LN+HO-DINA Model

Van der Linden's (2007) framework, on which the proposed model (Minchen & de la Torre, 2016) is built, defines separate person parameters to represent ability and speed intensity, which are postulated to be jointly normal with a covariance

that is likely to be nonzero. The HO-DINA model (de la Torre & Douglas, 2004) is used on the RA side of the model. On the RT side of the model, a 2PL-like IRF that was used by van der Linden (2006, 2007) is used, which is given by

$$P(T_{ij}|\tau_i, a_j, b_j) = \frac{1}{t_{ij}\sqrt{2\pi a_j^2}} \exp\Big[-\frac{(\ln t_{ij} - (b_j - \tau_i))^2}{2a_j^2}\Big], \tag{2.3}$$

where τ_i is the speed intensity latent variable for examinee i, a_j is a parameter that reflects the variation in log-times for item j, and b_j is the time threshold parameter for item j, which describes the mean of the log-times. Figure 2.2 shows a diagram of the model.

Based on this paramaterization, examinees with larger values of τ_i will work more quickly. The a_j parameter describes the variance of the log-times of respondents at a particular level of τ_i. As this value decreases, examinees with a similar speed-intensity parameter will have more homogeneous response times, allowing them to be differentiated other examinees. Note that the interpretation of a_j in this parameterization is based on the variance of the log-times and that smaller values of a_j indicate larger discriminations. This model will be referred to as the lognormal HO-DINA (LN+HO-DINA) model (Minchen & de la Torre, 2016).

The assumptions of this model are the same as those discussed in van der Linden (2007), and are

$$f(\boldsymbol{\theta}, \boldsymbol{\tau}) \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.4}$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{bmatrix}, \tag{2.5}$$

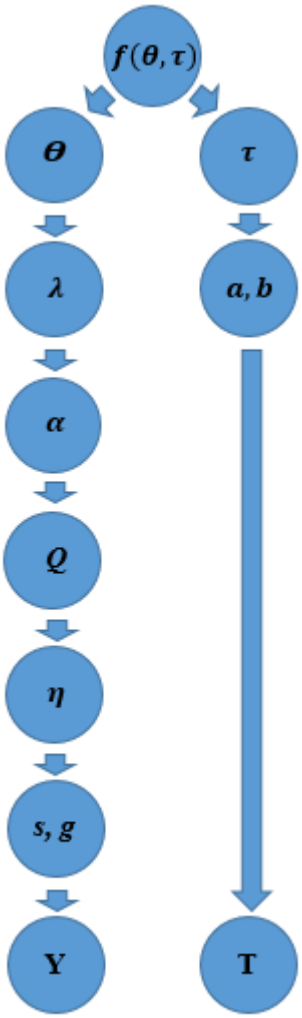

Figure 2.2: Acyclic Diagram of the LN+HO-DINA Model

$$\boldsymbol{\mu} = \mathbf{0}, \tag{2.6}$$

and

$$\sigma_\theta^2 = 1, \tag{2.7}$$

the latter two of which are for the sake of identifiability (van der Linden, 2007).

Note also that

$$\rho_{\theta\tau} = \frac{\sigma_{\theta\tau}}{\sigma_\theta \sigma_\tau}. \tag{2.8}$$

Other assumptions are the independence of the responses and the response times conditional on θ and τ, respectively (van der Linden, 2007), and the monotonicity of both θ and τ.

The critical component of the LN+HO-DINA model is the correlation between θ and τ. It is through this relationship that the estimation of θ is improved. Specifically, a nonzero correlation implies that θ and τ each provide some information about each other, thereby improving the estimation of both parameters. As a result of the improved estimation of θ, estimation of $\boldsymbol{\alpha}$ can also be improved. Higher magnitudes of this correlation are hypothesized to improve θ estimation to a greater degree, but it does not matter whether the correlation is negative or positive; in either case, it is the strength of the relationship that is crucial.

Parameter Estimation

The likelihood of the response accuracy data in the LN+HO-DINA can be written as

$$\begin{aligned} L(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}) &= \prod_{i=1}^{N}\prod_{j=1}^{J} P_{ij}^{Y_{ij}}(1-P_{ij})^{1-Y_{ij}} \\ &= \prod_{i=1}^{N}\prod_{j=1}^{J} [s_j^{1-Y_{ij}}(1-s_j)^{Y_{ij}}]^{\eta_{ij}} [g_j^{Y_{ij}}(1-g_j)^{1-Y_{ij}}]^{1-\eta_{ij}} \end{aligned} \tag{2.9}$$

where the IRF $P_{ij} = P(Y_{ij} = 1|\boldsymbol{\alpha}_i, s_j, g_j, \boldsymbol{\lambda}_1, \lambda_1) = P(Y_{ij} = 1|\boldsymbol{\alpha}_i, s_j, g_j)$ as is given in Equation 2.1. The likelihood of the response time portion of the model is given

by

$$L(\boldsymbol{T}|\boldsymbol{\tau}, \boldsymbol{a}, \boldsymbol{b}) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(T_{ij}|\tau_i, a_j, b_j), \tag{2.10}$$

where $P(T_{ij}|\tau_i, a_j, b_j)$ is defined in Equation 2.3. Thus, the likelihood of the complete data is

$$L(\boldsymbol{Y}, \boldsymbol{T}|\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{a}, \boldsymbol{b}) = L(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}) L(\boldsymbol{T}|\boldsymbol{\tau}, \boldsymbol{a}, \boldsymbol{b}). \tag{2.11}$$

Due to the complexity of this model, parameters were estimated using Markov chain Monte Carlo (MCMC). Under the MCMC approach, the goal is to sample from the joint posterior, which is given by

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\lambda_0}, \lambda_1, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\mu}, \boldsymbol{\sigma_\tau^2}, \boldsymbol{\rho_{\theta\tau}}|\boldsymbol{Y}, \boldsymbol{T}) \quad \propto \quad & L(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}) L(\boldsymbol{T}|\boldsymbol{\tau}, \boldsymbol{a}, \boldsymbol{b}) \\ & \times P(\boldsymbol{\alpha}|\boldsymbol{\lambda_0}, \lambda_1) P(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{\mu}, \boldsymbol{\sigma_\tau^2}, \boldsymbol{\rho_{\theta\tau}}) \\ & \times P(\boldsymbol{s}) P(\boldsymbol{g}) P(\boldsymbol{\mu}) P(\boldsymbol{\sigma_\tau^2}, \boldsymbol{\rho_{\theta\tau}}) \\ & \times P(\boldsymbol{\lambda_0}) P(\lambda_1) P(\boldsymbol{a}) P(\boldsymbol{b}). \end{aligned} \tag{2.12}$$

However, such sampling is difficult to do, so the set of full conditional distributions is obtained from the joint posterior and is used in the Gibbs sampler. The full

conditional distributions for the parameters are presented as follows:

$$
\begin{aligned}
P(\boldsymbol{\lambda_0}, \lambda_1|\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\sigma_\tau^2}, \boldsymbol{\rho_{\theta\tau}}, \boldsymbol{Y}, \boldsymbol{T}) &\propto P(\boldsymbol{\alpha}|\boldsymbol{\lambda_0}, \lambda_1)P(\boldsymbol{\lambda_0})P(\lambda_1), && (2.13)\\
P(\boldsymbol{s}, \boldsymbol{g}|\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\sigma_\tau^2}, \boldsymbol{\rho_{\theta\tau}}, \boldsymbol{Y}, \boldsymbol{T}) &\propto L(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}) && \\
&\quad \times P(\boldsymbol{s})P(\boldsymbol{g}), && (2.14)\\
P(\boldsymbol{\alpha}|\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\lambda_0}, \lambda_1, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\sigma_\tau^2}, \boldsymbol{\rho_{\theta\tau}}, \boldsymbol{Y}, \boldsymbol{T}) &\propto L(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}) && \\
&\quad \times P(\boldsymbol{\alpha}|\boldsymbol{\lambda_0}, \lambda_1), && (2.15)\\
P(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\sigma_\tau^2}, \boldsymbol{\rho_{\theta\tau}}, \boldsymbol{Y}, \boldsymbol{T}) &\propto L(\boldsymbol{T}|\boldsymbol{\tau}, \boldsymbol{a}, \boldsymbol{b})P(\boldsymbol{a})P(\boldsymbol{b}), && (2.16)\\
P(\boldsymbol{\theta}|\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\sigma_\tau^2}, \boldsymbol{\rho_{\theta\tau}}, \boldsymbol{Y}, \boldsymbol{T}) &\propto L(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}) && \\
&\quad \times P(\boldsymbol{\theta}|\boldsymbol{\tau}), && (2.17)\\
P(\boldsymbol{\tau}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\sigma_\tau^2}, \boldsymbol{\rho_{\theta\tau}}, \boldsymbol{Y}, \boldsymbol{T}) &\propto L(\boldsymbol{T}|\boldsymbol{\tau}, \boldsymbol{a}, \boldsymbol{b})P(\boldsymbol{\tau}|\boldsymbol{\theta}), && (2.18)\\
P(\boldsymbol{\sigma_\tau^2}|\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\rho_{\theta\tau}}, \boldsymbol{Y}, \boldsymbol{T}) &\propto P(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{\sigma_\tau^2}, \boldsymbol{\rho_{\theta\tau}})P(\boldsymbol{\sigma_\tau^2}), \text{ and} && (2.19)\\
P(\boldsymbol{\rho_{\theta\tau}}|\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\sigma_\tau^2}, \boldsymbol{Y}, \boldsymbol{T}) &\propto P(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{\sigma_\tau^2}, \boldsymbol{\rho_{\theta\tau}})P(\boldsymbol{\rho_{\theta\tau}}). && (2.20)
\end{aligned}
$$

Finally, the prior distributions in the full conditional distributions above are chosen as follows:

$$
\begin{aligned}
\boldsymbol{s} &\sim \text{Uniform}(a_s, b_s) && (2.21)\\
\boldsymbol{g} &\sim \text{Uniform}(a_g, b_g) && (2.22)\\
\boldsymbol{\lambda_0} &\sim \text{Normal}(\mu_{\lambda_0}, \sigma_{\lambda_0}) && (2.23)\\
\boldsymbol{\lambda_1} &\sim \text{Lognormal}(\mu_{\lambda_1}, \sigma_{\lambda_1}) && (2.24)\\
\boldsymbol{a} &\sim \text{Lognormal}(\mu_a, \sigma_a) && (2.25)\\
\boldsymbol{b} &\sim \text{Normal}(\mu_b, \sigma_b) && (2.26)\\
f(\boldsymbol{\theta}, \boldsymbol{\tau}) &\sim \text{MVN}(\boldsymbol{\mu^*}, \boldsymbol{\Sigma^*}) && (2.27)\\
\boldsymbol{\sigma_\tau^2} &\sim \text{Uniform}(a_{\sigma_\tau}, b_{\sigma_\tau}) && (2.28)\\
\boldsymbol{\rho_{\theta\tau}} &\sim \text{Uniform}(-1, 1). && (2.29)
\end{aligned}
$$

Next, the Metropolis-Hastings within Gibbs sampler (e.g., Gelman, Carlin, Stern, & Rubin, 2004) is used to conduct sampling from the full conditional distributions given above. Parameter estimates were computed after removing the burn-in samples, and were based on expected a posteriori (EAP). The number of draws required to reach convergence was evaluated using the potential scale-reduction factor (PSRF; Gelman & Rubin, 1992), $\hat{R}$, for a single replication of all low-quality item conditions. Because low-quality conditions contain more noise than medium- or high-quality conditions, they were expected to take longer to converge. The required chain lengths for the low-quality conditions were also used for their medium and high item quality conditions.

To determine the requisite chain length, five parallel chains with different starting values were used in the convergence analysis. The number of draws required for $\hat{R} \leq 1.2$ (de la Torre & Douglas, 2004) for each of six conditions - one for each of the model conditions for both values of λ_1 - was obtained. The value of $\hat{R}$ for all structural parameters was computed at each multiple of 2,500 draws. To be conservative, the first half of the draws were discarded as burn-in samples at each check. For example, when evaluating the $\hat{R}$ after 7,500 draws, the first 3,750 were discarded. Chain lengths satisfying $\hat{R} \leq 1.2$ were obtained for all low item quality conditions with one small exception. For the low λ_1, $\rho_{\theta\tau} = 0.8$ condition, the $\hat{R}$ for one parameter was slightly above the threshold ($\hat{R} \approx 1.25$) after a total of 25,000 draws.

2.5 Design and Analysis

A simulation study was conducted to evaluate the performance of the proposed model; the primary goal of the study was to determine the extent to which the inclusion of response times improved examinee classifications and higher-order ability estimates, and the relationships between this improvement and the conditions

of the simulation study. Data generation, parameter estimation, and subsequent computations and graphics were performed in R (R Core Team, 2015).

The quality of parameter estimation, attribute classification, and higher-order ability estimation was determined for a range of conditions. The critical component through which the response times were expected to aid in the estimation of examinee abilities was through the correlation between θ and τ and the strength of the relationship between θ and $\boldsymbol{\alpha}$. The estimation of τ was expected to improve the estimation of θ, thereby increasing the precision of $\boldsymbol{\alpha}$ estimation.

Factors in the simulation study were the correlation between θ and τ ($\rho_{\theta\tau} = 0, .8$), item quality (low, medium, and high), and the strength of the association between θ and $\boldsymbol{\alpha}$ ($\lambda_1 = 1.25, 2.5$). For all conditions, $K = 5$ and $\lambda_0 = [-1, -.5, 0, .5, 1]$. Conditions were fully crossed and replicated 100 times. The slip and guessing parameters were distributed as $s, g \sim U(0, .1)$, $s, g \sim U(.1, .2)$, and $s, g \sim U(.2, .3)$, representing items of high, medium, and low quality, respectively. It should be noted that these descriptors are relative and are not meant to be general judgments on item quality.

The $\rho_{\theta\tau} = .8$ condition displayed the overall effect of the collateral information; whereas the $\rho_{\theta\tau} = 0$ condition, when compared to the RA only model (HO-DINA) displayed the cost of simultaneous estimation when the inclusion of response time was not expected to be beneficial.

The effect of including response time was evaluated by comparing the quality of the parameter and ability estimates of the LN+HO-DINA ($\rho_{\theta\tau} = .8$) model to those for the HO-DINA model. Conditions in which the test was less informative (i.e., low quality items and low values of λ_1) and in which the attribute distribution was strongly related to the higher-order ability (i.e., the high λ_1 condition) were expected to offer the greatest opportunity for improvement to the classification accuracy and higher-order ability estimation from the inclusion of response time.

The estimation quality of the item parameters was assessed by using the root

Table 2.1: Simulation Study Q-Matrix

	Attribute				
Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	1	0	0	0
7	0	1	1	0	0
8	0	0	1	1	0
9	0	0	0	1	1
10	1	0	0	0	1
11	1	1	1	0	0
12	0	1	1	1	0
13	0	0	1	1	1
14	1	1	0	0	1
15	1	0	0	1	1

mean squared error (RMSE), computed as

$$RMSE = \sqrt{\frac{\sum_{r=1}^{Reps}\sum_{h=1}^{H}(\hat{\beta}_{hr} - \beta_{hr})^2}{Reps \times H}}, \tag{2.30}$$

where $Reps = 100$ replications, β is the parameter of interest, and H is the number of parameters per replications. For example, H is equal to N for θ, J for the item parameters, K for λ_0, and 1 for ρ and λ_1. Item parameters (i.e., s, g, a, and b) were not fixed across replications. Mean bias was computed for $\hat{\theta}$ as

$$BIAS = \frac{\sum_{r=1}^{Reps}\sum_{i=1}^{N}(\hat{\theta_{ir}} - \beta_{ir})}{Reps}. \tag{2.31}$$

To evaluate the quality of $\boldsymbol{\alpha}$ estimation, the correct attribute classification (CAC) and correct vector classification (CVC) were computed as

$$CAC = \frac{\sum_{r=1}^{Reps}\sum_{i=1}^{N}\sum_{k=1}^{K} I[\boldsymbol{\alpha}_{ik}^{(r)} = \hat{\boldsymbol{\alpha}}_{ik}^{(r)}]}{Reps \times N \times K},$$

and

$$CVC = \frac{\sum_{r}^{Reps} \sum_{i=1}^{N} \prod_{k=1}^{K} I[\boldsymbol{\alpha}_{ik}^{(r)} = \hat{\boldsymbol{\alpha}}_{ik}^{(r)}]}{Reps \times N}, \tag{2.32}$$

respectively, with the vector-level metric being the more stringent criterion of the two.

Lastly, because ancillary information is usually more beneficial in shorter tests, the number of items used in this study was fixed to 15. The Q-matrix used is given in Table 2.1. Although this Q-matrix does not include every possible item type, is balanced in the sense that each attribute is measured six times, and it is also *complete* (Chiu, Douglas, & Li, 2009), meaning that all attribute patterns can be statistically distinguished from one other. A Q-matrix that contains items that measures each attribute by itself, as does the one presented in Table 2.1, guarantees completeness, although it may not be necessary under certain CDMs (Köhn & Chiu, 2017).

2.6 Results

Model Parameters

Before discussing the classification accuracy of the examinees, we first discuss the estimation quality of all other parameters for all conditions, for which the RMSEs are shown in Table 2.2. For the sake of readability, the HO-DINA model will generally be referred to as simply the RA model, and the zero- and high-correlation LN+HO-DINA models will be referred to as the zero- and high-correlation RA+RT model.

Examining the $\lambda_1^{(L)}$ first, the most notable result in Table 2.2 is the vast reduction in the RMSE of $\hat{\theta}$ when using the high-correlation RA+RT model compared with the other two models. Even when the item quality was high, in which case

the expected improvement in estimation was lower due to the inherently higher level of informativeness of the test, there still was a substantial reduction in the RMSE of $\hat{\theta}$ from .678 (RA) and .674 (zero-correlation RA+RT) to .522 (high correlation RA+RT model). Also notable was the fact that that the estimation accuracy of ρ improved for medium and high item quality conditions.

The cost associated with including response time when it was not beneficial primarily plagued the estimation of λ_1, but the effect was not dramatic. Its RMSE increased slightly under all $\lambda^{(L)}$ conditions when comparing the RA model to the zero-correlation RA+RT model. Even when comparing the high-correlation model with the RA only model, the RMSE of λ_1 was similar or worse except for the high item-quality condition. The RMSEs for other parameters were quite similar regardless of the model.

As with the results under $\lambda_1^{(L)}$, the reduction in RMSE of $\hat{\theta}$ was also the most notable result for $\lambda_1^{(H)}$ conditions. Additionally, estimation quality of ρ improved for all item quality conditions. The decreases in RMSE were slightly smaller under $\lambda_1^{(H)}$, but this was because raising λ_1 benefited all models in the form of decreased RMSE, including the RA only model. Most importantly, the resulting RMSEs for $\hat{\theta}$ were lowest under $\lambda_1^{(H)}$ for a given item quality condition. Also, the RMSEs for parameters in the high-correlation model often were slightly lower than for those with the RA-only model for $\lambda_1^{(H)}$, but there were several exceptions (s for medium quality and λ_0 for high quality items). Clearly, the most important finding is the reduction in the RMSE of $\hat{\theta}$ under both levels of λ_1.

Classification Accuracy

Mean classification rates for all conditions are presented in Table 2.3, in which several trends were noteworthy. First, as with the model parameters, it was again evident that there was very little cost associated with estimating the more complex model when there was no correlation between the ability and speed parameters.

Table 2.2: RMSEs for Model Parameters

		$\lambda_1^{(L)}$			$\lambda_1^{(H)}$		
Item Quality	Parameter	-	$\rho = 0$	$\rho = .8$	-	$\rho = 0$	$\rho = .8$
Low	θ	0.805	0.805	0.585	0.688	0.696	0.522
	λ_0	0.243	0.230	0.235	0.161	0.164	0.142
	λ_1	0.156	0.168	0.198	0.374	0.409	0.324
	a	-	0.024	0.023	-	0.025	0.024
	b	-	0.045	0.046	-	0.055	0.046
	s	0.044	0.044	0.041	0.036	0.037	0.035
	g	0.038	0.037	0.034	0.039	0.039	0.032
	ρ	-	0.056	0.084	-	0.043	0.027
Medium	θ	0.729	0.730	0.545	0.584	0.586	0.472
	λ_0	0.118	0.117	0.117	0.089	0.093	0.088
	λ_1	0.090	0.093	0.090	0.204	0.186	0.196
	a	-	0.023	0.024	-	0.024	0.024
	b	-	0.050	0.043	-	0.053	0.050
	s	0.029	0.028	0.029	0.025	0.026	0.025
	g	0.024	0.025	0.023	0.026	0.029	0.024
	ρ	-	0.047	0.038	-	0.040	0.024
High	θ	0.678	0.674	0.522	0.521	0.521	0.435
	λ_0	0.084	0.082	0.082	0.059	0.061	0.063
	λ_1	0.075	0.082	0.070	0.146	0.126	0.145
	a	-	0.024	0.024	-	0.024	0.024
	b	-	0.048	0.046	-	0.057	0.051
	s	0.016	0.015	0.015	0.015	0.014	0.014
	g	0.013	0.013	0.012	0.016	0.017	0.016
	ρ	-	0.044	0.028	-	0.039	0.021

For low quality items, these classification rates were different by at most .005, and the differences were inconsistent in their direction. As a result, the HO-DINA model can be compared directly with the high-correlation LN+HO-DINA model and, with the exception of a small amount of noise, any differences in the model performance can be attributed to the inclusion of response time.

Next, comparing the RA only model with the high-correlation RA+RT model, the results indicated that the RA+RT model either outperformed or performed equivalently to the RA only model for every condition for both the CAC and CVC rates, a finding that was more consistent, and thus somewhat different, than what

Table 2.3: Mean Classification Rates

Item Quality	Model	ρ	Classification	$\lambda_1^{(L)}$	$\lambda_1^{(H)}$
Low	HO-DINA	-		0.813	0.858
	LN+HO-DINA	0	CAC	0.814	0.857
	LN+HO-DINA	.8		0.827	0.874
	HO-DINA	-		0.400	0.510
	LN+HO-DINA	0	CVC	0.402	0.506
	LN+HO-DINA	.8		0.428	0.542
Medium	HO-DINA	-		0.904	0.928
	LN+HO-DINA	0	CAC	0.906	0.928
	LN+HO-DINA	.8		0.910	0.933
	HO-DINA	-		0.638	0.716
	LN+HO-DINA	0	CVC	0.643	0.715
	LN+HO-DINA	.8		0.656	0.730
High	HO-DINA	-		0.976	0.981
	LN+HO-DINA	0	CAC	0.976	0.981
	LN+HO-DINA	.8		0.976	0.982
	HO-DINA	-		0.892	0.912
	LN+HO-DINA	0	CVC	0.892	0.913
	LN+HO-DINA	.8		0.895	0.918

was found for the structural parameters. Improvements tended to increase as the item quality decreased, and tended to be larger at the vector level.

In the low item quality, $\lambda_1^{(H)}$ condition, the improvements at the attribute- and vector-levels were .016 and .032, respectively. The corresponding improvements for the $\lambda_1^{(L)}$ were .014 and .028, respectively. Although these improvements were smaller than anticipated, they still could result in the correct classification of many more test takers, depending on the number of examinees. These rates for the medium quality items were .005 and .014 for $\lambda_1^{(H)}$ and .006 and .018 for $\lambda_1^{(L)}$, respectively.

For the high item quality conditions, the inclusion of response time still led to improvements, with the exception of the $\lambda_1^{(L)}$ condition at the attribute level, but they were more modest. At the vector level for the $\lambda_1^{(L)}$, the improvement was .003. At the attribute- and vector-levels for the $\lambda_1^{(H)}$, the improvements were .001 and .006, respectively.

Higher-Order Ability Estimation

Despite the fact that improvements in classification accuracy were modest at best for high item quality conditions, Table 2.2 shows that the estimation of the higher-order ability parameter, θ, still improved substantially. This finding suggests that there was still a benefit to including response time when the item quality was high, a conclusion which would not have been seen by examining classification accuracy alone.

Table 2.4: RMSE and Bias for θ for Low Quality Items

			Range of θ					
λ_1	Index	Model	$(-\infty,-2]$	$(-2,-1]$	$(-1,0]$	$(0,1]$	$(1,2]$	$(2,\infty)$
$\lambda_1^{(L)}$	RMSE	H	1.905	1.037	0.534	0.624	0.992	1.603
		L1	1.895	1.041	0.536	0.626	0.987	1.594
		L2	0.949	0.668	0.518	0.526	0.644	0.898
	Bias	HO	1.840	0.944	0.237	-0.310	-0.820	-1.512
		L1	1.835	0.943	0.243	-0.310	-0.817	-1.503
		L2	0.790	0.433	0.122	-0.141	-0.402	-0.743
$\lambda_1^{(H)}$	RMSE	H	1.695	0.850	0.490	0.584	0.729	1.331
		L1	1.733	0.854	0.493	0.593	0.732	1.299
		L2	0.991	0.606	0.453	0.470	0.537	0.840
	Bias	H	1.622	0.721	0.108	-0.216	-0.507	-1.238
		L1	1.652	0.724	0.106	-0.217	-0.501	-1.206
		L2	0.862	0.401	0.071	-0.121	-0.297	-0.718

Note. H: HO-DINA Model. M2: LN+HO-DINA Model (ρ=0) M3: LN+HO-DINA Model (ρ=.8)

To examine its behavior more closely, the RMSE and mean bias of $\hat{\theta}$ were analyzed at various regions across the domain of θ, as shown in Tables 2.4, 2.5, and 2.6. We begin by analyzing the low item quality results, which showed that there does not appear to be a consistent substantial difference in performance between the RA model and the zero correlation RA+RT model. These results again confirmed that the inclusion of response time, when it was not expected to be helpful, had little to no negative effect on examinee ability estimation.

By comparing the high correlation RA+RT model to the RA model, quite a different finding emerged. First, increasing the level of λ_1 always resulted in

Table 2.5: RMSE and Bias for $\hat{\theta}$ for Medium Quality Items

			Range of θ					
λ_1	Index	Model	$(-\infty,-2]$	$(-2,-1]$	$(-1,0]$	$(0,1]$	$(1,2]$	$(2,\infty)$
$\lambda_1^{(L)}$	RMSE	H	1.662	0.887	0.535	0.611	0.844	1.358
		L1	1.653	0.883	0.532	0.612	0.842	1.389
		L2	0.939	0.620	0.480	0.487	0.594	0.872
	Bias	H	1.600	0.771	0.178	-0.242	-0.656	-1.282
		L1	1.591	0.766	0.174	-0.244	-0.654	-1.307
		L2	0.808	0.407	0.112	-0.123	-0.368	-0.736
$\lambda_1^{(H)}$	RMSE	H	1.445	0.649	0.460	0.528	0.557	1.099
		L1	1.433	0.653	0.463	0.526	0.559	1.100
		L2	0.914	0.528	0.418	0.428	0.471	0.771
	Bias	H	1.370	0.501	0.039	-0.126	-0.333	-1.031
		L1	1.358	0.501	0.034	-0.124	-0.332	-1.033
		L2	0.786	0.314	0.042	-0.089	-0.238	-0.656

Note. H: HO-DINA Model. M2: LN+HO-DINA Model (ρ=0) M3: LN+HO-DINA Model (ρ=.8)

Table 2.6: RMSE and Bias for $\hat{\theta}$ for High Quality Items

			Range of θ					
λ_1	Index	Model	$(-\infty,-2]$	$(-2,-1]$	$(-1,0]$	$(0,1]$	$(1,2]$	$(2,\infty)$
$\lambda_1^{(L)}$	RMSE	H	1.416	0.776	0.547	0.569	0.762	1.326
		L1	1.436	0.770	0.543	0.573	0.756	1.305
		L2	0.878	0.587	0.463	0.467	0.566	0.845
	Bias	H	1.347	0.628	0.156	-0.182	-0.582	-1.251
		L1	1.364	0.622	0.153	-0.180	-0.579	-1.235
		L2	0.749	0.377	0.102	-0.109	-0.352	-0.720
$\lambda_1^{(H)}$	RMSE	H	1.213	0.525	0.448	0.466	0.489	1.062
		L1	1.203	0.527	0.448	0.461	0.491	1.070
		L2	0.832	0.463	0.394	0.392	0.437	0.751
	Bias	H	1.143	0.349	0.029	-0.067	-0.276	-0.997
		L1	1.134	0.349	0.032	-0.059	-0.277	-1.006
		L2	0.720	0.246	0.039	-0.052	-0.204	-0.637

Note. H: HO-DINA Model. M2: LN+HO-DINA Model (ρ=0) M3: LN+HO-DINA Model (ρ=.8)

a reduction in the bias and RMSE of $\hat{\theta}$ for all three models, although these reductions were more substantial for more extreme θ. It also always resulted in a smaller reduction in the benefit of including response time with the exception of the RMSE for $-1 < \theta \leq 1$.

In the central regions of θ - those within one unit of zero - inclusion of response time offered a more modest improvement to the estimation of θ, with respect

to both the bias and the RMSE, than was seen at more extreme regions of θ. The intervals in which $\mid \theta \mid > 2$, the reduction in RMSE and bias as a result of including response time was very large. For the negative interval, the reduction in RMSE was .956 and .704 for the low and high λ_1 conditions, respectively. The corresponding reductions in bias for the same interval were 1.05 and .76. For the positive side of the interval, the RMSE reductions were .705 and .491, and the bias reductions were .769 and .52 for the low and high λ_1 conditions, respectively. The intermediate intervals demonstrate the same trend, but to a lesser degree. Collectively, these findings demonstrate that the ability of the RA model to estimate θ degraded as the true values moved away from zero, where the estimates showed strong shrinkage. Although the bias and RMSE results for the RA+RT model were also somewhat large for these intervals, they were a vast improvement over the performance of the RA model.

Examination of Tables 2.5 and 2.6 revealed that the benefit of including response time when the item quality was higher continued to provide a benefit in the estimation of θ, a result that was somewhat different than the corresponding findings for the estimation of $\boldsymbol{\alpha}$. Only under two conditions (medium and high item quality for high λ_1) and for one range of θ ($-1 < \theta \leq 0$), did the bias increase slightly when including response time. For these conditions, increases in the mean bias were small, at .003 and .01 for the medium and high item quality conditions, respectively. Conversely, the RMSE still declined in these conditions. Reductions in the RMSE and bias of $\hat{\theta}$ were still large, though, even for high item quality conditions for more extreme values of of θ.

These findings demonstrated that the RA only model tended to restrict the range of $\hat{\theta}$ in a way that may have corresponded to the threshold values of the attributes, showing a strong inward bias, indicating that the model may generally not be able to estimate θ precisely, except in the center of the ability distribution. The lack of the RA only model's ability to estimate θ precisely corroborates de la

Torre and Douglas' (2004) findings, who found RMSE values for $\hat{\theta}$ in the range of approximately .6 to .7, as well as shrinkage in $\hat{\theta}$ relative to the IRT estimates of θ. To the extent that values of θ are of interest, which is one practical advantage of using a higher-order attribute distribution with CDMs (de la Torre & Douglas, 2004), ancillary information such as response time may prove to be beneficial.

2.7 Real Data Example

Data Description

Van der Maas and Jansen (2003) collected data on both response time and response accuracy on a set balance scale questions. In each question, one or more weights were situated at various distances from the center of a balance scale. The task of each question was to determine whether or not the scale was balanced on the fulcrum. If it was not balanced, examinees were to determine which direction it would lean. To solve each question, examinees needed to take into account either the number of weights on both sides of the fulcrum, the location of the weights, or both pieces of information.

Table 2.7: Q-Matrix for the Balance Scale Data

			Attribute	
Item Type	# Items	Description	Distance	Torque
I	10	Simple-distance	1	0
II	10	Conflict-balance B	1	1
III	10	Conflict-distance	1	1
IV	10	Conflict-balance A	1	1

For questions in which the numbers of weights on either side of the scale were equal but the locations differed, the examinee would simply use the distance of the weights to determine which way the scale would tip. If, however, the numbers and the distances of the weights differed, then the examinee would need to apply

the torque formula to determine the direction the scale would tip. These skills can be viewed as hierarchical (Siegler, 1976, 1981); students are not expected to be able to apply the torque rule without command of the distance rule.

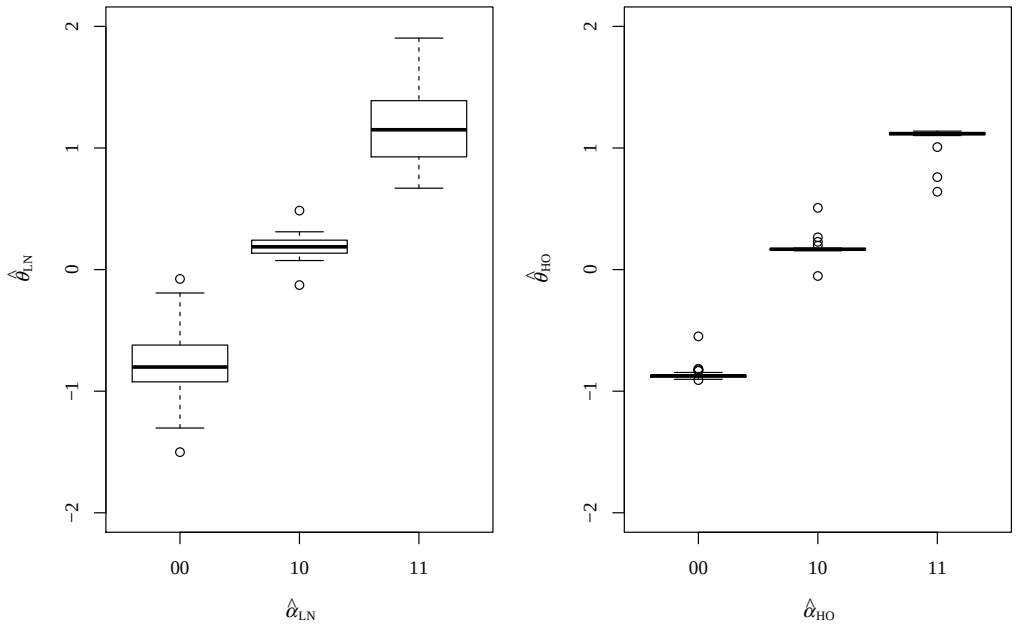

Figure 2.3: Higher-Order Ability Distribution by Latent Class

Minchen, de la Torre, and Liu (in press) analyzed a subset of the data, which included responses for both accuracy and time on 40 questions for 146 examinees, with both the DINA model and the C-DINA model. The Q-matrix used in their study, shown in Table 2.7, was also used in this example and gives item descriptions, which provide insight into the nature of the question and its answer. For example, conflict-balance means that the distance and number of weights give conflicting information that requires the use of the torque rule to solve. Solving the problem correctly will result in the determination that the scale will remain balanced. The Conflict-balance A and B problems differed in that the type A problems can also be solved with a different strategy (van der Maas & Jansen, 2003).

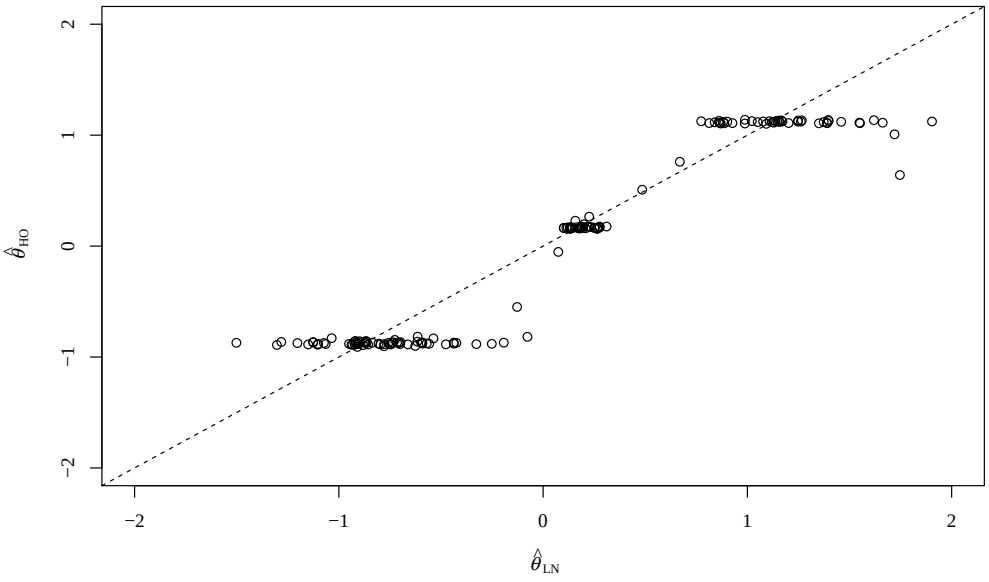

Figure 2.4: Relationship between Higher-Order Ability Estimates

Regardless of whether classifications were made using the DINA or C-DINA model, the results generally showed a similar pattern (Minchen, de la Torre, & Liu, in press). Examinees who had all attributes required by the problem had longer and more variable response times compared to examinees who did not have all required attributes. Furthermore, this difference was larger for problems with more required attributes.

Analysis and Results

In this example, the data were fitted with three different models: the DINA, the HO-DINA, and the LN+HO-DINA. First, the classifications obtained using the HO-DINA and LN+HO-DINA were nearly identical to those obtained by using the standard DINA model. However, because the true classifications were not known, an evaluation of classification accuracy was not possible. Nonetheless,

such high classification agreement suggested, at a minimum, that including response time did not have a negative affect on classification accuracy relative to the RA only model.

Next, the higher-order ability estimates for the HO-DINA and LN+HO-DINA models, $\hat{\theta}_{HO}$ and $\hat{\theta}_{LN}$, respectively, were compared to each other. Figure 2.3 shows the distributions of the higher-order ability estimates obtained under each of the higher-order models as a function of the associated latent classes, and Figure 2.4 shows the scatterplot of $\hat{\theta}_{HO}$ and $\hat{\theta}_{LN}$. Although the medians for the higher-order abilities for each latent class were similar for both models, and $cor(\hat{\theta}_{HO}, \hat{\theta}_{LN}) = .95$, the variances in the estimates were much larger when response time was included than when it was ignored, as seen in the left and right panels of Figure 2.3, respectively. It appeared that, without the use of response time, the higher-order ability estimates within a latent class were nearly identical.

The final analysis was to determine the extent to which the inclusion of response time could improve classification accuracy. To accomplish this, the results from the full-length test were used as the baseline measure. Next, the test length was halved, and both the HO-DINA and LN+HO-DINA model were fit to the data, and their results compared. To shorten the test, half of the items were removed at random such that 25% of the items measured only the first attribute, and the remaining items measured both attributes, which was also the case with the full-length Q-matrix. Classifications and higher-order ability estimates were averaged across 100 replications for each model.

The attribute- and vector-wise agreement rates between the shortened test fitted with the HO-DINA model and the full-length test were .98 and .95, respectively. The same rates for the shortened test fitted with the LN+HO-DINA model were .98 and .97. The correlations and RMSEs between the higher-order abilities for the two HO-DINA models was .94 and .29, respectively, whereas the same quantities between the full-length HO-DINA model and the LN+HO-DINA

model and were .98 and .18, respectively. In both cases, and by all three measures, the use of response time in the shorter test resulted in improved classification and ability estimates.

2.8 Summary and Discussion

This paper presented a method of including response time to assist in the estimation of person-level parameters in cognitive diagnosis modeling. To this end, a model was presented that adapted van der Linden's (2007) hierarchical framework for use in CDM, which assumes a bivariate normal correlation between speed and ability parameters. To adapt this framework to suit CDMs, de la Torre and Douglas's (2004) higher-order attribute distribution was used and applied to the DINA model.

In general, the inclusion of response time improved classification accuracy on average, provided that the latent variables for speed and response time had a nonzero correlation. Greater improvements were found at the vector level. The effect of including response time when it was not expected to be beneficial (i.e., the zero-correlation RA+RT model) was generally negligible. Interestingly, including response time did not result in a substantial improvement in the model parameters, with the exception of the higher-order ability parameter, θ. Because of this finding, improvements in classification accuracy were only attributable to improvements in the estimation of θ. The RMSE and bias of $\hat{\theta}$ also improved substantially for most ranges of θ and under most conditions when including response time. As the item quality increased, the benefit of response time lessened, but it still was generally beneficial in improving the estimation of θ.

The differing results on the effect of the inclusion of response time on classification accuracy and higher-order ability estimation accuracy is worth noting. The mechanism by which classification accuracy was improved was through the

reduced variation in the prior distribution of θ that comes as a result of the ancillary information provided by $\hat{\tau}$. Because the inclusion of response time did not have a substantial effect on the other model parameters, however, improvement in classification accuracy was nearly entirely dependent on the improvement in the estimation of θ. However, the amount of improvement in the estimation of θ does not correspond in a linear fashion to improvement of $\hat{\boldsymbol{\alpha}}$, as can be seen in the higher quality item conditions. This finding indicates two things. First, even with high attribute pattern classification rates, $\hat{\theta}$ can still be improved, and second, related to the first finding, attempting to improve $\hat{\boldsymbol{\alpha}}$ by improving $\hat{\theta}$ may only be beneficial under certain circumstances.

The proposed model was also applied to an existing data set. Results from the analysis were generally consistent with the results of the simulation study. Specifically, examinee classifications under the LN+HO-DINA model were very similar to those obtained under the HO-DINA and DINA models when using the complete test. However, the higher-order ability estimates, when using response time, had substantially more variation within the latent classes. When response time was not used, higher-order ability estimates had very little variation within classes. This was consistent with the behavior of $\hat{\theta}$ in the simulation, in which shrinkage was partially mitigated by including response time. Finally, when using a shortened version of the test, the inclusion of the response time resulted in improved classifications and higher-order ability estimates when the full-length HO-DINA model was used as the baseline.

One critical factor that was not investigated in this study is the effect that K has on the effect of response time with respect to improving classification accuracy. It is likely that larger values of K would result in improved estimation of θ in the same way that additional items result in improved estimation of θ in a standard IRT context. However, it may not be feasible to have values of K large enough to yield the desired precision of θ; thus, response time, or other ancillary

variables that are modeled in a way similar to how response time is modeled in this article, may prove to be beneficial.

Finally, using response time as a source of information to estimate abilities should be done with caution. One consideration has to do with the meaning of the response times (van der Linden, 2006), and whether or not their use as ancillary estimation is valid. In the example we provided, there is some theoretical basis to use response times (van der Maas & Jansen, 2003); however, in other settings this may not be the case. In particular, if high-stakes decisions are being made with test scores and the response times are being used to estimate ability, then examinees with the same response accuracy pattern could have different ability estimates, which may be problematic. However, in diagnostic applications, response times may simply provide additional insight into the response processes of examinees.

The main finding in this research was that, on average, attribute classifications can be improved slightly when including response times, but higher-order ability estimates can be improved substantially. To that end, an important future direction would be to explore the relationship between ability estimates obtained using the LN+HO-DINA method and a standard IRT model. One thing that would be interesting to explore is whether data generated by the LN+HO-DINA model could be better estimated with a 2-parameter logistic IRT model rather than the LN+HO-DINA model. Another potential direction would be to explore the framework presented in this model with other CDMs.

References

Ben-Simon, A., Budescu, D. V., & Nevo, B. A. (1997). Comparative study of measures of partial knowledge in multiple choice tests. *Applied Psychological Measurement, 21*, 65-88.

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

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

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*, 343-362.

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

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

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

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

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

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics, 37*, 655-670.

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement, 31*, 525-543.

Finkelman, M., Kim, W., Weissman, A., & Cook, R. (2014). Cognitive diagnostic models and computerized adaptive testing: Two new item-selection methods that incorporate response times. *Journal of Computerized Adaptive Testing, 2*, 59-76.

Fosdick, B. K., & Raftery, A. E. (2012). Estimating the correlation in bivariate normal data with known variances and small sample sizes (Technical Report No. 591). Department of Statistics, University of Washington.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis, 2nd Ed.* Boca Raton: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 457-472.

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

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

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

Köhn, H. F., & Chiu, C. Y. (2017). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika, 82*, 1-21.

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

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

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica: Journal of the Econometric Society*, 647-663.

Minchen, N. D., & de la Torre, J. (2016, April). *Using response time in cognitive diagnosis models.* Poster presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Minchen, N., de la Torre, J., & Liu, Y. (in press). A cognitive diagnosis model for continuous response. *Journal of Educational and Behavioral Statistics.*

Noel, Y., & Davier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement, 31*, 47-73.

Noel, Y. (2014). A beta unfolding model for continuous bounded responses. *Psychometrika, 79*, 647-674.

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

Ranger, J., & Kuhn, J.-T. (2012). Improving item response theory model calibration by considering response times in psychological tests. *Applied Psychological Measurement, 36*, 214-231.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*, 203-219.

Sie, H., Finkelman, M. D., Riley, B., & Smits, N. (2015). Utilizing response times in computerized classification testing. *Applied Psychological Measurement, 39* 389-405.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8,* 481-520.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child development, 46*(2), 1-84.

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

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*, 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287-308.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*, 5-20.

van der Linden, W. J. (2009). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement, 33*, 25-41.

van der Linden, W. J., & Glas, C. A. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika, 75*, 120-139.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*, 365-384.

van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika, 68*, 251-265.

van der Maas, H. L. J., & Jansen, B. R. J. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology, 85*, 141-177.

Xu, X., Chang, H., & Douglas, J. (April 2003). *Computerized adaptive testing strategies for cognitive diagnosis.* Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.

Chapter 3

The Jensen-Shannon Divergence as an Item Selection Algorithm in CD-CAT

Abstract

Item selection algorithms for computerized adaptive testing (CAT) have been proposed within the context of both item response theory (IRT) and cognitive diagnosis models (CDMs). Although the literature in CDM has recently expanded to include continuous response modeling, CAT algorithms for CDM are currently limited to dichotomous response. For various reasons, existing algorithms may not be applicable in their present forms, if at all, for continuous response models, particularly one that is saturated. This article proposes a new algorithm developed in the context of a generalized continuous response CDM. The algorithm selects the item that maximizes the posterior-weighed Jensen-Shannon divergence, which is a proposed measure of item discrimination in continuous response CDMs. Results show that the algorithm provides a substantial improvement over random item administration. The method's viability is also demonstrated in a brief real data example.

Keywords: cognitive diagnosis models, continuous response, response time, Jensen-Shannon divergence, DINA model, G-DINA model, C-DINA model, C-G-DINA model

THE JENSEN-SHANNON DIVERGENCE AS AN ITEM SELECTION ALGORITHM IN CD-CAT

3.1 Introduction

Traditional assessments typically aim to estimate examinees' ability levels on a unidimensional, broadly-defined trait. In contrast, cognitively diagnostic assessments (CDAs; de la Torre & Minchen, 2014) and their associated statistical models, cognitive diagnosis models (CDMs), offer an alternative assessment paradigm in which the latent variable is generally conceptualized as a multidimensional set of discrete attributes. In educational testing applications, these attributes are usually thought of as skills. The latent variable in CDMs is represented as a binary vector that denotes the presence or absence of these attributes. As such, CDMs generally yield a profile of skills rather than a single value representing one's location on a continuum, as is done in traditional assessment. Such models are a relatively recent development in psychometrics, and research pertaining to these models has expanded rapidly in the last two decades.

Recent publications have proposed item selection algorithms for cognitive diagnosis computerize adaptive testing (CD-CAT; e.g., Xu, Chang, & Douglas; Kaplan, de la Torre, & Barrada, 2015) that will be discussed in detail later; however, these algorithms have been developed in the context of a dichotomous item response. Perhaps not coincidentally, most CDMs developed to date are designed to handle dichotomous responses (e.g., Haertel, 1989; Junker & Sijtsma, 2001; Templin & Henson, 2006; de la Torre, 2011), with only a few designed to handle polytomous responses (e.g., de la Torre, 2009a; Ma & de la Torre, 2016). However, response types of a continuous nature also exist and have been studied in the context of item response theory (IRT; e.g., Noel & Davier, 2007; Noel, 2014) and CDM (Minchen & de la Torre, 2016; Minchen, de la Torre, & Liu, in press).

One readily-available example of a continuous response is *response time*. Although it may be difficult to record item response time in traditional testing, it is much easier, and essentially free, to record it in computer-based testing formats. Because it is an additional source of information provided by examinees, methods should be developed to make use of response time. Although response time may be the most obvious example of a continuous response, other continuous response types exist as well. One such response type is "probability testing," in which examinees estimate the probabilities that various alternatives are correct. Typical multiple-choice questions without partial credit can be viewed as a special case of probability testing, in which the chosen answer is assigned a 100% probability of being correct. Probability testing can reveal more information per question (de Finetti, 1965). See Dressel and Schmidt (1953), Ben-Simon, Budescu, and Nevo (1997), and Minchen, de la Torre, and Liu (in press) for more information about probability testing. Another type of continuous response measure is simply to place a mark on a continuum, which can be seen as the continuous version of a Likert scale, for which Noel and Dauvier (2007) and Noel (2014) have developed IRT models.

It is important to differentiate the fact that response time is a continuous response that is separate from the response accuracy, whereas probability testing and the use of a mark on a line are continuous responses for the response accuracy. Therefore, if both the response time and response accuracy were to be recorded and analyzed, one or both measures could be continuous. Response time in and of itself, however, may be of interest to some researchers. For example, Minchen, de la Torre, and Liu (in press) analyzed response times for balance-scale data using their continuous response CDM.

Although continuous response CDMs are being studied, there is no CD-CAT algorithm designed specifically for this response type. In addition, for reasons that will be explored in detail later, adapting current CD-CAT algorithms to

continuous-response CDMs may not be straightforward, if possible at all. Thus, this paper proposes a CD-CAT algorithm based on the Jensen-Shannon divergence (JSD; Lin, 1991) that is developed in the context of a generalized continuous response CDM (Minchen & de la Torre, 2016). First, however, the developments in CD-CAT are reviewed, followed by brief overview of the CDMs relevant to this work.

3.2 Cognitive Diagnosis Computerized Adaptive Testing

CAT typically improves the efficiency and the accuracy of measurement (Xu, Wang, & Shang, 2016). In CAT, examinees are administered items that are tailored to their ability levels, which are re-estimated after each question (or set of questions, as in multi-stage testing) with greater precision as the test proceeds. Thompson and Weiss (2011) enumerate the following critical components of a CAT: 1) the item bank, 2) the ability level to which examinees are assigned prior to the observation of any responses, 3) the item selection algorithm, 4) the scoring method, and 5) the stopping criterion (or criteria). Most developments in CAT have been in the framework of traditional testing rather than CDA. However, some research has been conducted recently on CD-CAT. The critical elements of a CAT outlined above are retained in the context of CD-CAT with appropriate modifications.

In particular, the item selection algorithms used in CD-CAT are different than those used in IRT-based CATs due to the multivariate, discrete nature of the latent variable, and thus will be the primary focus herein. Specifically, it is common (Barrada, Olea, Ponsada, & Abad 2009; Xu, Chang, & Douglas, 2003) to choose as the next item the one that maximizes the Fisher information (Lord, 1980).

In IRT, the *progressive* method (Reveulta, 1995, as cited in Revuelta & Ponsada, 1998; Revuelta & Ponsada, 1996, as cited in Revuelta & Ponsada, 1998) is based on the Fisher information, whereas Chang and Ying's (1999) *alpha-stratified* method blocks items by their discriminations and the computes the magnitude of the linear difference between the ability estimate and the item's difficulty. In the case of information-based statistics, Xu, Chang, and Douglas (2003) point out that a requirement of using the Fisher information function is that the likelihood function be twice differentiable, but due to the discrete nature of attribute vectors, $\boldsymbol{\alpha}$, in CDM, no such derivative for $\boldsymbol{\alpha}$ will exist. In the case of the progressive method, the distance between the item and person locations cannot be measured in a simple unidimensional way.

Because of these challenges, CAT selection indices have been either developed or adapted from other types of IRT-based CAT selection indices. For example, the Kullback-Leibler (KL; Cover & Thomas, 1991) information, which is another way to measure information, was first used in CAT in the context of IRT (Chang & Ying, 1996). The KL computes the distance between two distributions; in the case of its application to CAT, these are the response probabilities, summed over all possible responses, to the next item for pairs of latent trait or class values. Xu, Chang, and Douglas (2003) compared this method with another selection algorithm based on the Shannon Entropy (Shannon, 2001), referred to as the SHE method (Tatsuoka, 2002). Briefly, the Shannon Entropy quantifies the dispersion of a single probability distribution, whereas the KL quantifies the degree of divergence between two distributions. The SHE method chooses the item that minimizes the sum of entropies of the updated posterior distributions under each possible response. In their study, Xu et al. (2003) found the SHE algorithm to be superior to the KL algorithm.

Cheng (2009) notes that under the KL selection algorithm, the prior distribution is not systematically updated in accordance with the responses. Thus,

Cheng (2009) proposed the posterior-weighted KL (PWKL) algorithm, in which both informative priors and updated priors may be used throughout the test. In this way, the divergence measure is weighted by the current probability that an examinee resides in that class. This index essentially summarizes the distance between the current estimate and all other possible values of the attribute vector. They found that the PWKL was superior to the SHE method. Kaplan, de la Torre, and Barrada (2015) proposed an improved version of the PWKL, referred to as the modified PWKL (MPWKL), in which the estimate of the attribute vector is replaced by the entire posterior distribution, thus avoiding the uncertainty associated with a classification. Kaplan et al.'s (2015) method outperformed the PWKL under most conditions.

Kaplan et al. (2015) also used the generalized discrimination index (GDI), which was originally introduced in the context of Q-matrix validation (de la Torre & Chiu, 2016), as an item selection index. The GDI chooses the item from the bank that maximizes the variance of the probabilities of success, which are assumed to be known for each item due to prior calibration, weighted by the examinee's posterior probability of residing in each group. In Kaplan et al.'s (2015) research, the GDI performed very similarly to the MPWKL, but the GDI's computational time was significantly lower than that of the MPWKL.

Finkelman et al. (2014) present a method of incorporating response time into a CD-CAT selection index. They use a CDM for the item response and a lognormal model for the response times. Their algorithm is based on Fan et al.'s (2012) criterion of evaluating information per time unit. Finkelman et al.'s (2014) method may be preferred when the time to completion for each examinee should be limited.

These CD-CAT item selection algorithms discussed are all designed to work with binary measures for response accuracy. Generalizations of these algorithms to continuous response models may not be straightforward, if possible at all. For

example, the GDI is based on the weighted variance of success probabilities. In such models, the probability of a particular response cannot be represented by single number p, as in the binary models. Rather, the probabilities are given by a distribution, which is a function of t. In binary response models, the mean and variance are given by $E(X) = p$ and $Var(X) = p(1-p)$, respectively. Thus, p, the probability of a correct response, simply represents the mean of the Bernoulli distribution, and the variance can be derived directly from the same parameter. Thus, it is sufficient to work only with the mean, on which the GDI is based.

In the continuous response applications, however, representing each distribution by only their means may obscure some information about the distribution, because the variance cannot be written as a function of the mean. For example, for a lognormal random variable, T, both its mean and variance, which are given by $E(T) = (\mathrm{e}^{\mu+\sigma^2})$ and $Var(T) = (\mathrm{e}^{\sigma^2} - 1)(\mathrm{e}^{2\mu+\sigma^2})$, respectively, are functions of *both* parameters μ and σ. For example, Figure 3.1 shows a variety of differently-shaped lognormal distributions, all of which have the same mean of approximately 12.18, but variances ranging from approximately 255 to 2833. Therefore, an index based on the concept of the GDI, but that takes into account information beyond just the mean, is needed for continuous response models.

The SHE method sums entropies over possible responses, but in a continuous model, the sum would need to be replaced with an integral, as would the KL. Furthermore, the KL is only defined if the support of each distribution is nonzero across the union of the support of both distributions, due to the division in the formula, which may render the index undefined in a continuous model. This may happen when the responses from one group are much larger than those for the other group, as in the case of highly discriminating items. In binary models, such as the DINA and G-DINA, this problem will generally not occur because there are only two possible responses. Even for very high quality items, all groups will likely have a nonzero probability of a correct response. Thus, a new algorithm

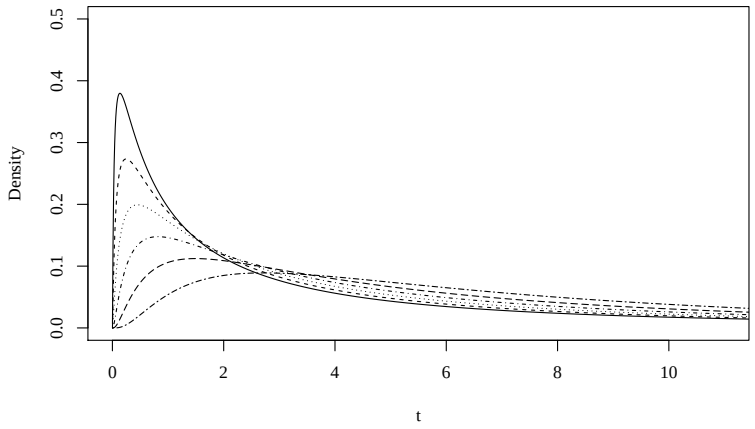

Figure 3.1: Various Lognormal Distributions with Identical Means

that avoids these problems is necessary for continuous response models.

An additional consideration regarding KL-based indices is that the KL only measures the divergence between two distributions. Therefore, to develop a CD-CAT algorithm for use with continuous response models, it is best to work with the continuous version of the generalized DINA (de la Torre, 2011), known as the continuous-generalized DINA (C-G-DINA; Minchen & de la Torre, 2016), which will be reviewed in the next section. The rationale for developing the algorithm in the context of the C-G-DINA models is that the C-DINA model only produces two latent groups for each item, potentially allowing for an existing method, such as the MPWKL (Kaplan et al., 2015) or an adaptation of the MPWKL, to be used, notwithstanding the aforementioned issue of division by zero. However, the C-G-DINA model may partition examinees into more than just two groups per item, necessitating a new method altogether.

3.3 Cognitive Diagnosis Models

Dichotomous-Response CDMs

One of the most basic CDMs is the deterministic inputs, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model. Under this model, which may be a simplification of reality, examinees must possess all skills that are required by the item to respond correctly; possessing skills beyond those that are required is of no additional benefit. If a student lacks any of the skills being measured by the question, he or she is expected to respond incorrectly. This response process is modeled mathematically with the latent response variable, which is defined as $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$, where α_{ik} is binary, and denotes the presence or absence of the k^{th}, $k = 1, \ldots, K$, skill for examinee i, $i = 1, \ldots, N$, and q_{jk} denotes whether or not attribute k is required for item j, $j = 1, \ldots, J$.. Its item response function (IRF) can be defined as

$$P(X_{ij} = 1|\boldsymbol{\alpha}_i, s_j, g_j) = (1 - s_j)^{\eta_{ij}}(g_j)^{1-\eta_{ij}}, \tag{3.1}$$

where g_j and s_j are the two item parameters - guessing and slip - that define the DINA model. The guessing parameter models the probability that an examinee without all the required skills responds correctly; the slip parameter models the probability that, in spite of having all the required attributes, an examinee responds incorrectly.

The generalized-DINA (G-DINA; de la Torre, 2011) model is a generalization of the DINA model in which a probability of providing a correct response is estimated for each of the combinations of necessary attributes. Specifically, the G-DINA model defines $K_j^* = \sum_{k=1}^{K} q_{jk}$, which represents the number of required attributes for item j. The G-DINA model estimates unique probabilities of success for each of the $2^{K_j^*}$ possible latent groups that are based on the main effects of

having mastered each attribute, plus all possible two- to K_j^*-way interactions.

Continuous-Response CDMs

As the continuous-response analog to the DINA model, the continuous-DINA (C-DINA) was recently proposed by Minchen, de la Torre, and Liu, (in press). The latent response variable, η_{ij}, is identical to that used in the DINA model. Thus, the IRF of the C-DINA model is given as

$$P(T_{ij} \leq t|\boldsymbol{\alpha}_i) = \int_0^t [f_{j0}(t_{ij})]^{1-\eta_{ij}}[f_{j1}(t_{ij})]^{\eta_{ij}} dt_{ij}, \tag{3.2}$$

where

$$f_{j\eta}(t_{ij}) = \frac{1}{t_{ij}\sqrt{2\pi\sigma_{j\eta}^2}} \exp\Big[-\frac{(\ln t_{ij} - \mu_{j\eta})^2}{2\sigma_{j\eta}^2}\Big], \tag{3.3}$$

and T_{ij} is the continuous response, $f_{j0}(t_{ij})$ and $f_{j1}(t_{ij})$ are the response distributions for the $\eta_j = 0$ and 1 groups. The item parameters are given by $\mu_{j\eta}$ and $\sigma_{j\eta}$, which are the mean and variance, respectively, of the logarithm of the responses. Note that, whereas the DINA model defines two *probabilities* for each item, the C-DINA defines two *distributions* for each item. In the case of the C-DINA model, the lognormal distribution is used, in part because it is amenable to modeling response time (van der Linden, 2006), but other distributions, such as the Beta, could be used as well (Minchen, de la Torre, & Liu, in press). Binary CDMs estimate correct response probabilities for each latent group; these probabilities can actually be viewed as the mean of the Bernoulli distribution governing their responses. The C-DINA, however, requires the estimation of two parameters for *each* of the lognormal distributions, resulting in twice as many parameters as are required in the DINA model.

In the same way the G-DINA model generalizes the DINA, the C-G-DINA

model generalizes the C-DINA. The C-G-DINA model (Minchen & de la Torre, 2016) includes features of both the C- and G-DINA models. Rather than partitioning examinees into just two groups for each item as the DINA and C-DINA models do, the C-G-DINA model partitions examinees into $2^{K_j^*}$ groups for each item, just as the G-DINA model does. Similarly, the C-G-DINA model estimates a lognormal distribution for each of these groups.

The C-G-DINA model is given as follows. Let the cumulative distribution function of the response be given by

$$P(T_{ij} \leq t|\boldsymbol{\alpha}_i) = \int_0^t f_{j\eta}(t_{ij})dt_{ij}, \tag{3.4}$$

where

$$f_{j\eta}(t_{ij}) = \frac{1}{t_{ij}\sqrt{2\pi\sigma_{j\eta}^2}} \exp\Big[-\frac{(\ln t_{ij} - \mu_{j\eta})^2}{2\sigma_{j\eta}^2}\Big], \tag{3.5}$$

and where $\eta = 1 \dots 2^{K_j^*}$, representing the collection of reduced latent groups for a given item. Whereas the C-DINA model estimates two lognormal distributions for each item, the C-G-DINA model estimates $2^{K_j^*}$ lognormal distributions for each item. Note also that, as with the G-DINA model, mastering additional attributes beyond those required to solve the problem has no effect under this model. From this parameterization, it can be seen that mastering a particular combination of the required attributes may result in a unique μ and σ, which in turn may result in unique lognormal distributions of responses.

Figure 3.2 shows the differences between the dichotomous and continuous versions of the DINA and G-DINA models in the case where $K_j^* = 2$, resulting in $2^{K_j^*} = 4$ groups, and where both attributes are required to solve the problem ($K_j^* = 2$). The top panel shows that the DINA and C-DINA models both partition examinees into two groups, whereas the bottom panel shows that the G-DINA

and C-G-DINA models partition examinees into $2^{K_j^*} = 2^2 = 4$ groups.

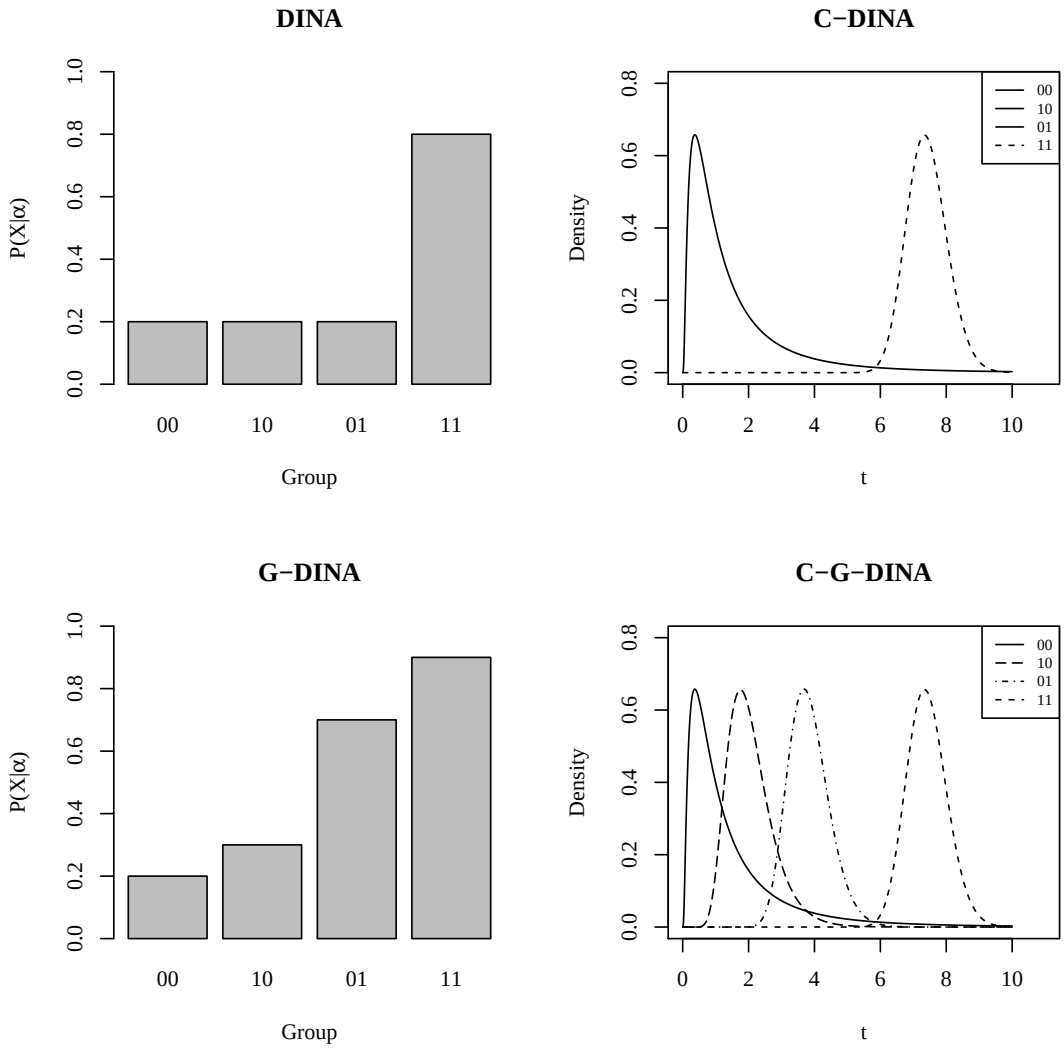

Figure 3.2: Binary and Continuous CDMs

3.4 The Jensen-Shannon Divergence

The Jensen-Shannon divergence (JSD; Lin, 1991) is an index that measures the degree of divergence in multiple probability distributions. Its function is similar to that of the KL information, but it can be extended to more than two probability

distributions, and is given by

$$JSD_{\{w_1,w_2,\ldots,w_n\}}(P_1, P_2, \ldots, P_n) = H\Big(\sum_{i=1}^{n} w_i P_i\Big) - \sum_{i=1}^{n} w_i H(P_i). \tag{3.6}$$

In this formula, $H(X)$ is the Shannon Entropy (Shannon, 2001), computed as

$$H(X) = E[I(X)] = E[-ln(P(X))] = -\int P(x) ln[P(x)] dx, \tag{3.7}$$

and $w_1, w_2, \ldots, w_n$ is a series of weights, and $P_1, P_2, \ldots, P_n$ are the probability density functions under consideration.

The JSD computes the Shannon Entropy of the mixture of all probability distributions, each respectively weighted by w_n, from which the sum of the weighted Shannon Entropies of each of the individual probability distributions is subtracted. Shannon Entropy is maximized for flat distributions and minimized for degenerate distributions. Thus, the collection of distributions that will maximize the JSD relative to other collections of distributions will be one in which the mixture distribution is flatter, but where the individual distributions have taller peaks. It also can clearly be seen from Equation 3.6 that the measure is symmetric, whereas other divergence measures, such as the KL, are not.

3.5 The JSD as an Item Selection Index

Our use of the JSD as an item selection index in the context of CD-CAT is based on the concept of choosing the most discriminating next item for each examinee. Minchen, de la Torre, and Liu (in press) defined discrimination in the continuous response setting to be the degree of separation between the response distributions of the latent groups. For the C-G-DINA model, the concept of discrimination closely follows the G-DINA discrimination index (GDI; de la Torre & Chiu, 2016). The GDI is based on the variance of the response probabilities

for the latent groups. For the C-G-DINA, the discrimination is defined as the total amount of dispersion among all probability distributions (Minchen & de la Torre, 2016). Thus, the JSD naturally lends itself for use both as a measure of item discrimination and as an item selection measure in CD-CAT for continuous response models.

The selection algorithm involves two steps. The first is to compute the JSD for all remaining items in the bank using the current estimate of the posterior class membership probabilities as the weights. The second is to administer the item that maximizes the JSD for each individual, and recompute the posterior distribution based on the examinee's response. For the sake of notation, assume that examinees have completed $j-1$ items, and that the j^{th} item is being selected.

Let $P(\boldsymbol{\alpha}_{\eta_j}|\boldsymbol{t}_i^{(j-1)})$ be the *combined* posterior distribution according to the q-vector of the candidate item, $\boldsymbol{q}_j$, whose elements are given by

$$p(\eta_j = \eta|\boldsymbol{t}_i^{(j-1)}) = \sum_{\boldsymbol{\alpha}_l:\eta_j=\eta} P(\boldsymbol{\alpha}_l|\boldsymbol{t}_i^{(j-1)}), \tag{3.8}$$

where $\eta_j = 1, \ldots, 2^{K_j^*}$, $l = 1, \ldots, 2^K$, and where $\boldsymbol{t}_i^{(j-1)}$ is the vector of responses for the i^{th} examinee for items $1, \ldots, j-1$. There are now only $2^{K_j^*}$ different values. Now, the JSD can be used as an item selection index, and is defined for the j^{th} item as

$$JSD_{\{P(\boldsymbol{\alpha}_1|\boldsymbol{t}_i^{(j-1)}),\ldots,P(\boldsymbol{\alpha}_{2^{K_j^*}}|\boldsymbol{t}_i^{(j-1)})\}}\big[f_{j1}(t_{ij}), \ldots, f_{j\eta_j}(t_{ij})\big]_{ij} =$$

$$H\Big[\sum_{\eta_j=1}^{2^{K_j^*}} P(\boldsymbol{\alpha}_{\eta_j}|\boldsymbol{t}_i^{(j-1)}) f_{j\eta_j}(t_{ij})\Big] - \sum_{\eta_j=1}^{2^{K_j^*}} P(\boldsymbol{\alpha}_{\eta_j}|\boldsymbol{t}_i^{(j-1)}) H\big(f_{j\eta_j}(t_{ij})\big), \tag{3.9}$$

where

$$f_{j\eta_j}(t_{ij}) = \frac{1}{t_{ij}\sqrt{2\pi\sigma^2_{j\eta_j}}} \exp\Big[- \frac{(\ln t_{ij} - \mu_{j\eta_j})^2}{2\sigma^2_{jl_j}}\Big]. \tag{3.10}$$

An Example

We now turn to an example in which the JSD for several candidate C-DINA items is computed for an examinee. To simplify this example, assume that $K = 3$, that there is one of each possible item types, and that all seven items have identical item parameters, yielding coincidental item response curves, which are shown in Figure 3.3, for both groups.

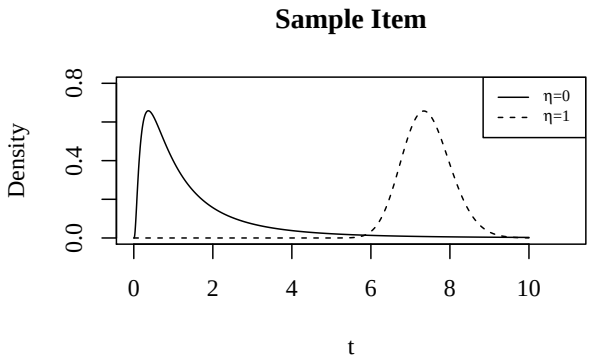

Figure 3.3: JSD Example: C-DINA Item

Table 3.1 shows the posterior probability of membership in each of the latent classes. The probabilities in bold and standard fonts represent those that constitute the $\eta = 0$ and 1 latent groups, respectively, under the candidate item. Each of these sets of probabilities are summed for each of the latent groups, producing the weights shown in the left panel of Table 3.2. The weights and the probability densities of the curves in Figure 3.3 are used to compute the JSD for examinee i on item j according to Equation 3.9. In the right panel of Table 3.2 are the first and second terms and JSD computed using Equation 3.9. In this case, the 110

item has the highest JSD value (.44), and the 111 item has the lowest JSD value (.07). Because the 110 item has the largest JSD for this examinee, it would be the next one administered.

Table 3.1: JSD Example (1)

	Latent Class Posterior Probabilities							
Item	000	100	010	001	110	101	011	111
100	**0.05**	0.01	**0.05**	**0.01**	0.80	0.05	**0.01**	0.02
010	**0.05**	**0.01**	0.05	**0.01**	0.80	**0.05**	0.01	0.02
001	**0.05**	**0.01**	**0.05**	0.01	**0.80**	0.05	0.01	0.02
110	**0.05**	**0.01**	**0.05**	**0.01**	0.80	**0.05**	**0.01**	0.02
101	**0.05**	**0.01**	**0.05**	**0.01**	**0.80**	0.05	**0.01**	0.02
011	**0.05**	**0.01**	**0.05**	**0.01**	**0.80**	**0.05**	0.01	0.02
111	**0.05**	**0.01**	**0.05**	**0.01**	**0.80**	**0.05**	**0.01**	0.02

Table 3.2: JSD Example (2)

	Weights		JSD		
Item	w_0	w_1	T1	T2	JSD
100	**0.12**	0.88	1.33	.98	.35
010	**0.12**	0.88	1.33	.98	.35
001	**0.91**	0.09	1.62	1.36	.25
110	**0.18**	0.82	1.45	1.01	.44
101	**0.93**	0.07	1.58	1.37	.21
011	**0.97**	0.03	1.50	1.39	.10
111	**0.98**	0.02	1.47	1.40	.07

Note. T1 and T2: First and second terms in testing Equation 3.9, respectively.

3.6 Design and Analysis

The goal of the simulation study was to demonstrate that the JSD is a viable item selection index in the context of continuous response items in CD-CAT. This was done by analyzing the improvement the JSD algorithm provided over a random selection of items, using either the classification rates of the examinees or the number of items required for a given level of classification certainty, depending

on the stopping rule. Additionally, the number of times each item was administer was examined.

Classification accuracy at both the attribute- and vector-levels were analyzed. The correct attribute classification (CAC) accuracy was calculated for each attribute pattern l as

$$CAC_l = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} I[\boldsymbol{\alpha}_{ik} = \hat{\boldsymbol{\alpha}}_{ik}]}{N \times K}, \tag{3.11}$$

and the correct vector classification (CVC) was calculated for each attribute pattern l as

$$CVC_l = \frac{\sum_{i=1}^{N}\prod_{k=1}^{K} I[\boldsymbol{\alpha}_{ik} = \hat{\boldsymbol{\alpha}}_{ik}]}{N}, \tag{3.12}$$

where $N = 1000$ was the number of examinees per attribute pattern.

The simulation study included the following factors: item quality (low and high), test length (fixed and variable), and model (C-DINA and C-G-DINA). Item quality was defined as the discrimination of the item as quantified by the JSD, the computational details of which will be given later. For the fixed test length conditions, short ($J = 5$) and long ($J = 10$) tests were used. Fixed length test conditions allow for the classification accuracy to by analyzed after the administration of an arbitrary number of items. For the variable test length conditions, the criteria used was the minimax of the posterior distribution, as in Kaplan et al. (2015), which required that the highest posterior node meet or exceed a specified value. The values used were .50 and .75, to represent varying levels of uncertainty in classification. For example, the .5 minimax condition required that each examinee's posterior distributions had a maximum value (height) of at least .5, which means that 50% of the mass is concentrated on a single latent class, whereas the .75 condition requires for 75% of the mass to be concentrated on a single latent class. Higher minimax values are associated with lower degrees of

classification uncertainty. The algorithm continued administering items to each examinee until this condition is satisfied. The variable test length conditions allowed for the comparison of the number of items required to achieve a particular level of certainty under various conditions.

The construction of the item bank and the examinee distribution was similar to that outlined in Kaplan et al. (2015). First, the item bank contained 310 items, which were comprised of each of the 31 possible q-vectors for the $K = 5$ case, replicated 10 times. Such an item bank allowed for the algorithm to choose the same item type for all items in all fixed-length conditions. In each of the item quality conditions, all items in the bank were of roughly the same quality. A small amount of noise was added to the item parameters to reduce the possibility of multiple items have an identical JSDs. A total of 6,000 examinees were generated, comprised of 1,000 from each of the following attribute patterns: [00000], [10000], [11000], [11100], [11110], and [11111]. Using this subset of attribute vectors allowed for easier comparison of item usage. Maximum a posteriori (MAP) estimation was used for examinee classification. No attribute patterns were precluded from estimation, meaning that examinees could have been classified in a pattern other than the six used to generate the data. All computations were performed in R (R Core Team, 2015). The first two items administered were chosen randomly, and the starting posterior distributions for all examinees were flat.

The expected item discriminations for the C-DINA and C-G-DINA were designed to be approximately the same for each of the discrimination conditions, which was accomplished via the following procedure. First, μ parameters for the C-DINA items were chosen. For the low discrimination condition, $\mu_0 = 1$ and $\mu_1 = 2$ were used; for the high discrimination condition, $\mu_0 = 1$ and $\mu_1 = 3$ were used. Next, σ parameters were chosen that resulted in curves of approximately a fixed height; these parameters were fixed across all 31 items. Then, the JSD was computed for each item using the attribute distribution given above. Note that

the raw, or equally-weighted, discrimination was the same for each item, but the JSD varies depending on the attribute distribution. For example, a 11010 item would have weights of .33 and .67 for the $\eta = 0$ and 1 conditions. The JSD was then averaged across all items, providing an average test discrimination.

The next challenge was to select C-G-DINA parameters for the 31 items that resulted in an average test discrimination similar to that found for the C-DINA. To do so, the necessary number of μ parameters (i.e., $2^{K_j^*}$) were equally spaced between a lower and a higher bound, and corresponding σ parameters were chosen for each distribution so that the resulting curves were the same height as the C-DINA curves. Next, the JSD was computed for each item, and then averaged across all items. Again, using a 11010 item as an example, the weights for each of the latent group would be .17, .17, 0, 0, .33, 0, 0, and .33. Finally, ranges of the μ parameters (separate ranges for low and high discriminations) were adjusted such that the average test discrimination was similar to that of the same discrimination condition for the C-DINA model.

Another way to examine the comparability of the discrimination conditions is to compute the examinee classification rates that result from the various sets of item parameters. To that end, a small simulation study using a 15-item test and the JSDs prescribed by the aforementioned procedure was conducted. Under the low discrimination condition, the attribute classification rates were .973 and .975 for the C-DINA and C-G-DINA models, respectively; for the high discrimination condition, the rates were higher, as in 1.000 and .996, respectively. The vector classification rates for the low discrimination condition were .896 and .897, and for the high discrimination condition, 1.000 and .983 for the C-DINA and C-G-DINA models, respectively. The JSDs for the low and high discrimination conditions were approximately .39 and .54, respectively. Therefore, the low and high discrimination conditions produce similar classification rates for each model.

3.7 Results

Fixed Test Length Results

Tables 3.3 and 3.4 show the mean attribute- and vector-level classifications for examinees by attribute pattern for all conditions involving the five- and 10-item tests, respectively. The weighted averages were computed by assigning the following weights to each of the attribute patterns: 1/32 (00000), 5/32 (10000), 10/32 (11000), 10/32 (11100), 5/32 (11110), 1/32 (11111). These weights reflect what the proportion of examinees with a given number of attributes would be if the attribute patterns were uniformly distributed.

These tables reveal several important results. Most importantly, the JSD always performed at least as good as, but usually better than, random item selection. Improvements were more substantial for the low discrimination conditions because items that are more discriminating are more informative and result in higher classification rates, regardless of the way they are administered. Improvements were also more substantial for the vector-level criterion. Additionally, the JSD resulted in nearly perfect attribute- and vector-level classification rates for all conditions involving a 10-item test. The lowest classification rate was .97 for the low discrimination/vector-level/C-G-DINA condition for the 00000 pattern.

Another general finding was that the classification rates generally rose as the number of attributes in the pattern increased. The increase was much more substantial for the random item selection method because the rates for the patterns with fewer attributes were much lower than those for the JSD. For example, for the low discrimination condition, the vector-level classification rates for the five-item test with data generated from the C-G-DINA model ranged from .59 for examinees with no attributes to .91 for examinees with all five attributes. The same quantities for random item administration ranged from .11 to .90. The corresponding ranges for the C-DINA data were .69 to 1.00, and .01 to 1.00,

Table 3.3: Mean Classification Rates by Attribute Pattern: 5-Item Tests

Generating Model	Pattern	Classification Type: Attribute, Low Disc. JSD	Attribute, Low Disc. Ran.	Attribute, High Disc. JSD	Attribute, High Disc. Ran.	Vector, Low Disc. JSD	Vector, Low Disc. Ran.	Vector, High Disc. JSD	Vector, High Disc. Ran.
C-G-DINA	00000	0.89	0.70	0.93	0.75	0.59	0.11	0.70	0.16
	10000	0.91	0.80	0.95	0.87	0.66	0.31	0.82	0.49
	11000	0.92	0.88	0.98	0.93	0.68	0.56	0.91	0.73
	11100	0.94	0.92	0.98	0.95	0.74	0.68	0.90	0.78
	11110	0.98	0.96	1.00	0.98	0.88	0.82	0.98	0.90
	11111	0.98	0.98	1.00	0.99	0.91	0.90	1.00	0.96
	Wt. Avg.	0.94	0.89	0.97	0.93	0.73	0.60	0.90	0.72
C-DINA	00000	0.92	0.55	0.98	0.62	0.69	0.01	0.90	0.02
	10000	0.92	0.69	0.96	0.78	0.74	0.18	0.80	0.26
	11000	0.90	0.81	0.99	0.88	0.58	0.35	0.94	0.48
	11100	0.93	0.91	1.00	0.95	0.69	0.62	1.00	0.76
	11110	0.98	0.98	1.00	1.00	0.91	0.90	1.00	0.98
	11111	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Wt. Avg.	0.93	0.85	0.99	0.90	0.71	0.50	0.95	0.62

Note. Disc.: Discrimination. Ran.: Random. Wt. Avg: Weighted Average

respectively.

The attribute patterns with fewer attributes were usually more difficult to estimate, and the JSD algorithm offers a substantial improvement over random selection for these attribute patterns. One exception to this pattern was perfect CAC and CVC rates for the 11000 attribute pattern for C-DINA data with the 10-item test. Upon closer inspection, it was found that these anomalous rates were likely due to the prior distribution, which was updated after each item following the second item. Without a prior distribution, the posterior nodes typically took on one of two values, with one of the values being very small. For example, if there were two large nodes with the remainder being close to zero, the approximate value of the two large nodes was around .5. By using the updated prior distribution, one of the large nodes was made slightly larger than the other. For the 11000 pattern, the correct prior node was often slightly higher than the node corresponding to the other nonzero posterior node, causing the pattern to

Table 3.4: Mean Classification Rates by Attribute Pattern: 10-Item Tests

		Classification Type							
		Attribute				Vector			
Generating		Low Disc.		High Disc.		Low Disc.		High Disc.	
Model	Pattern	JSD	Ran.	JSD	Ran.	JSD	Ran.	JSD	Ran.
C-G-DINA	00000	0.99	0.79	1.00	0.84	0.97	0.19	1.00	0.34
	10000	1.00	0.90	1.00	0.95	0.98	0.56	1.00	0.79
	11000	1.00	0.96	1.00	0.97	0.98	0.81	1.00	0.86
	11100	1.00	0.95	1.00	0.99	0.99	0.79	1.00	0.95
	11110	1.00	0.97	1.00	1.00	0.99	0.86	1.00	0.98
	11111	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Wt. Avg.	1.00	0.95	1.00	0.97	0.98	0.76	1.00	0.88
C-DINA	00000	1.00	0.70	1.00	0.78	0.99	0.17	1.00	0.27
	10000	1.00	0.83	1.00	0.89	0.99	0.36	1.00	0.47
	11000	1.00	0.93	1.00	1.00	0.99	0.68	1.00	1.00
	11100	1.00	0.98	1.00	0.94	1.00	0.91	1.00	0.70
	11110	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
	11111	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Wt. Avg.	1.00	0.94	1.00	0.96	1.00	0.75	1.00	0.80

Note. Disc.: Discrimination. Ran.: Random. Wt. Avg: Weighted Average

be correctly classified at a high rate. It is important to note, however, that even though the pattern was classified correctly at a high rate, the highest node in the posterior was frequently only around .5.

Variable Test Length Results

Table 3.5 shows the average number of items required to attain posterior minimax values of .50 and .75 for each of the attribute patterns, and Table 3.6 shows the average efficiency of the tests using the JSD algorithm relative to the tests with random selection. The JSD outperformed random item selection for all attribute patterns except for the 11111 attribute pattern for the low discrimination condition with data generated from the C-DINA model, for which the differences were no larger than .08 items. To rule out noise as the cause, this condition was replicated with 10,000 examinees, but similar results were obtained. The reason for this is unclear and warrants further study.

Table 3.5: Average Number of Items Administered by Attribute Pattern

Generating Model	Pattern	Minimax: 0.50				Minimax: 0.75			
		Low Disc.		High Disc.		Low Disc.		High Disc.	
		JSD	Rand	JSD	Rand	JSD	Rand	JSD	Rand
C-G-DINA	00000	6.39	13.53	5.63	10.07	7.40	33.20	6.45	25.28
	10000	6.29	12.39	5.55	9.19	7.76	25.67	6.40	18.05
	11000	6.31	10.79	5.51	8.23	7.85	19.00	6.31	12.84
	11100	6.24	9.65	5.26	7.16	7.75	15.17	6.45	10.07
	11110	6.08	8.24	4.97	6.30	7.20	12.77	5.94	8.38
	11111	5.96	7.16	4.75	5.56	6.94	9.22	5.56	7.09
C-DINA	00000	6.64	19.93	5.16	14.16	7.38	44.90	5.92	33.46
	10000	6.62	20.36	5.18	14.35	7.65	45.09	5.84	36.50
	11000	6.17	17.18	5.10	13.83	7.21	25.64	5.47	19.17
	11100	5.82	10.60	4.58	8.56	6.78	14.50	4.76	9.96
	11110	4.99	6.18	3.58	4.46	5.86	7.85	4.14	5.50
	11111	3.82	3.70	3.02	3.09	4.39	4.38	3.15	3.36

Note. Disc.: Discrimination. Rand: Random.

Table 3.6: Average Efficiency of the JSD CAT Algorithm by Attribute Pattern

Generating Model	Pattern	Minimax: 0.50		Minimax: 0.75	
		Low Disc.	High Disc.	Low Disc.	High Disc.
C-G-DINA	00000	0.47	0.56	0.22	0.26
	10000	0.51	0.60	0.30	0.35
	11000	0.58	0.67	0.41	0.49
	11100	0.65	0.74	0.51	0.64
	11110	0.74	0.79	0.56	0.71
	11111	0.83	0.85	0.75	0.78
C-DINA	00000	0.33	0.36	0.16	0.18
	10000	0.33	0.36	0.17	0.16
	11000	0.36	0.37	0.28	0.29
	11100	0.55	0.53	0.47	0.48
	11110	0.81	0.80	0.75	0.75
	11111	1.03	0.98	1.00	0.94

Note. Disc.: Discrimination.

Similar to the findings in the previous section, attribute patterns with more attributes generally required fewer items to reach the required level of certainty in the posterior distribution within a given set of conditions. This reduction in the number of items required was more pronounced for the .75 minimax requirement,

mainly because of the large increases in the number of items required for examinees with fewer attributes. Compared with the random selection, there was much less variation in the average number of items administered across the attribute patterns for the JSD algorithm. For example, for the low discrimination, .75 minimax condition for the C-DINA model, the JSD algorithm required an average of 7.38 (efficiency = .16) and 4.39 (efficiency = 1.00) items for the 00000 and 11111 attribute patterns. The corresponding rates for random selection were 44.90 and 4.38. Although examinees with fewer attributes required more items than those with more attributes for both algorithms, the increases were much more modest for the JSD algorithm.

Although not presented here in full, the number of items required to reach the desired level of certainty under random item administration can be exorbitant. For example, for at least one 00000 examinee in the low discrimination/C-DINA/.75 condition, 300 items were required when using random item selection, but only 12 for the same condition when using the JSD. Despite the fact that the efficiencies shown in Table 3.6 for two conditions were greater than or equal to one, all maxima for the JSD algorithm were strictly less than their random counterparts.

Finally, comparing the average and maximum numbers of items required for the C-DINA and C-G-DINA models yielded mixed results. For example, the average number of items required when using the JSD was typically less for the C-DINA; however, for random selection, the C-DINA typically required more items for attribute patterns with fewer specifications, and fewer items otherwise, than the C-G-DINA. This finding was often true for the maxima as well.

Item Usage

To gain additional insight into the nature of the JSD algorithm, the final step of analysis was to examine item usage. Figures 3.4 and 3.5 show the overall

item usage proportions for all 31 item types for each of the attribute patterns for the C-DINA and C-G-DINA models, respectively. When applying the JSD to C-DINA data, the item preference followed a somewhat consistent pattern. As the number of attributes in the pattern increased, the JSD moved from selecting mostly single-attribute items to items that measure more attributes. In the case of the 00000 examinees, items administered were almost exclusively single-attribute items.

In the case of the 11111 examinees, however, the preference appeared to be mainly for three- and four-attribute items. The items that were used the most all measured the fifth attribute. Interestingly, there was virtually no preference for single-attribute items. The item usage for the remaining attribute patterns was more moderate than either the 00000 or the 11111 patterns. For example, the 11110 examinees tended to be administered primarily two-, and three-, and four-attribute items, with all of them measuring the fourth attribute. The 11100 examinees were administered many 11100 items.

In contrast to the item usage of the C-DINA model, the JSD preferred single-attribute items, and sometimes two-attribute items, for all attribute patterns when using the C-G-DINA model. For the 00000 and 10000 patterns, the strongest preference was toward 10000 items. For the 11000 pattern, the strongest preference was instead for the 01000 item. There was also a slightly-elevated preference for 01100 items, but it was still less than all single-attribute items. For 11100 examinees, the most used items were 00100, 00010, and 01100, all of which were used with approximately the same frequency. The 11110 examinees were administered the 00010 and 00001 items at the highest rates, and also the 00110 items. Finally, the most frequently administered items for the 11111 examinees were 00001 and 00011. Only when examinees had three or more attributes were there substantial rates of use for two-attribute items. Items with three or more attributes were rarely used for any attribute pattern.

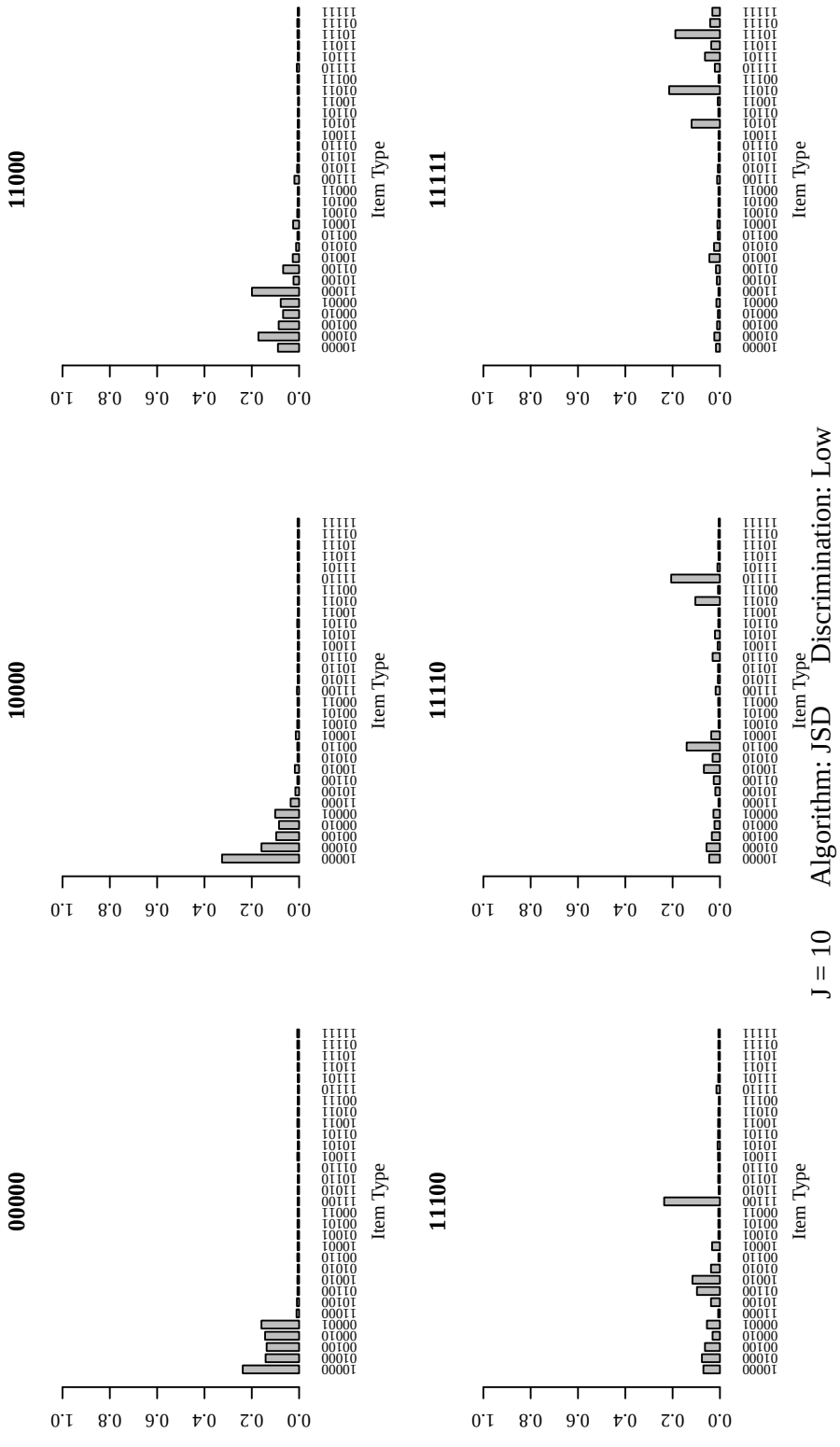

Figure 3.4: Overall Item Usage: C-DINA

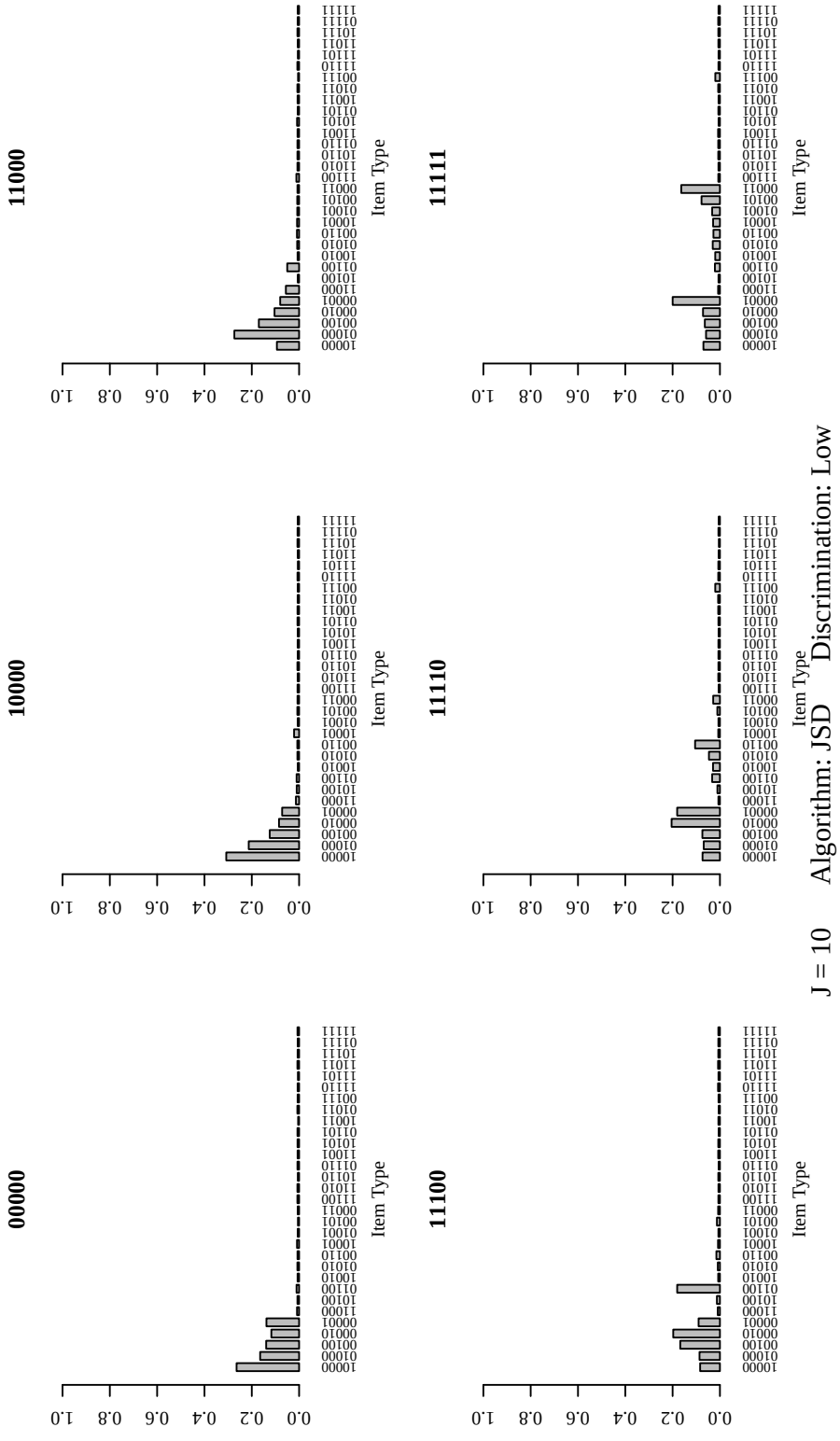

Figure 3.5: Overall Item Usage: C-G-DINA

In summary, the usage patterns of the JSD for both models are similar for 00000, 10000, and 11000 attribute patterns, in which mainly single-attribute items were chosen. However, for the remaining attribute patterns, the C-DINA results show that the JSD tended to choose items with more attributes, particularly for the 11111 pattern, whereas the C-G-DINA results show that the JSD chose primarily one- and two-attribute items.

3.8 Real Data Example

Minchen, de la Torre, and Liu (in press) analyzed a set of balance-scale data, originally collected and analyzed by van der Maas and Jansen (2003), with the C-DINA model. The JSD CAT algorithm proposed in this paper was applied to that data set. Parameter estimates obtained in the aforementioned article were used in this example.

Although the true classifications are not known, the pseudo-true classifications are assumed to be those obtained when administering all 40 items. To emphasize the distinction between the unknown true classifications and the pseudo-true classifications, CAC^* and CVC^* are used denote the latter. If all 40 items were administered, both of these indices would be equal to one.

Table 3.7 shows the classification rates for balance scale data, according to each of three levels of the two types of stopping rules. When a fixed number of items were administered, high correct classifications were obtained with very short tests. In this example, administering only five items resulted in CAC^* and CVC^* rates of .90 and .85, respectively. By increasing the test length to 20, those rates rose to .99, indicating near perfect classification rates. Correct classifications when using the minimax stopping conditions also show that reasonably high rates were attainable without imposing a stringent criterion. For example, CAC^* and CVC^* rates for the .50 condition were .83, and .77, respectively.

Table 3.7: Balance Scale Data CAT Results

Stopping Rule		CAC^*	CVC^*
Test Length:	5	.90	.85
	10	.96	.92
	20	.99	.99
Posterior Minimax:	.5	.83	.77
	.75	.90	.85
	.95	.96	.95

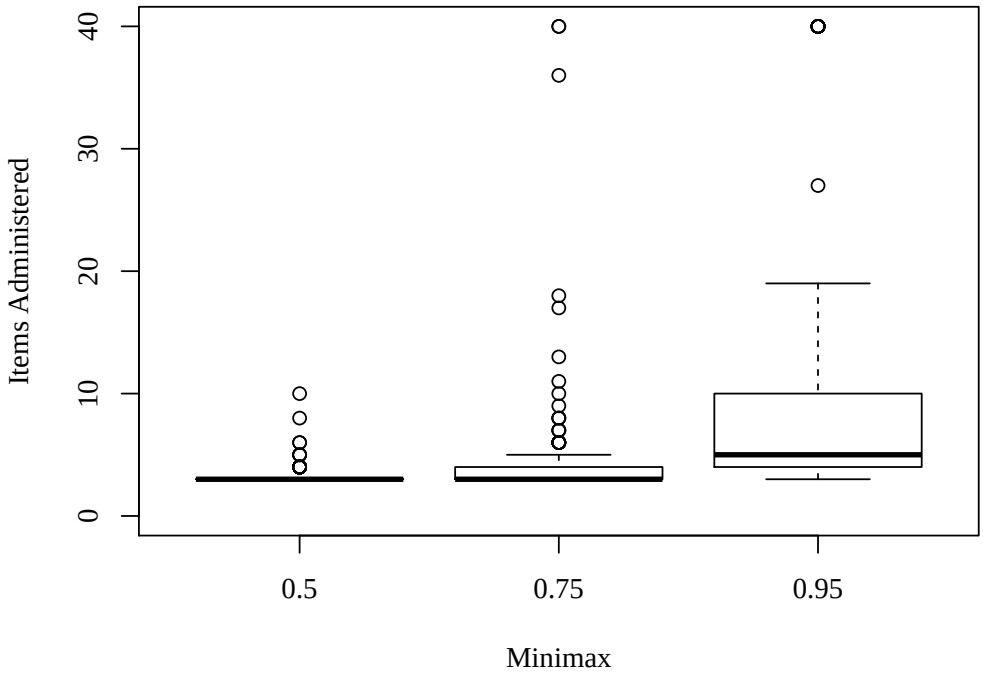

Figure 3.6: Number of Items Administered by Minimax Condition

Figure 3.6 shows the distribution of the number of items administered under each of the posterior minimax conditions. The lowest minimax condition resulted in very few items being administered, with most examinees being administered about four items. The .75 minimax condition still resulted in many examinees being administered around four of five items, but there also were a number of examinees that were administered more items. The .95 criterion still resulted in

most examinees only being administered a small number of items. The median for this condition was less than 10, and all but a few examinees were administered less than half of the items in the bank. When taken together, the test lengths shown Figure 3.6 and the classification rates shown in in Table 3.7 both suggest that applying the JSD CAT algorithm proposed in this paper can result in a dramatic reduction in the number of items required to obtain good classification rates.

3.9 Discussion and Conclusion

The purpose of this article was to adapt the JSD for use as an item selection method in CD-CAT for continuous response CDMs. Doing so also affirms the JSD's use as a measure of item discrimination in continuous response CDMs. The JSD was shown to perform better than random item selection in nearly all conditions. For fixed-length tests, the improvements were most substantial for low discrimination conditions and at the vector classification level, as well as for attribute patterns with fewer attributes, which were inherently more difficult to classify, and thus required more items. For tests in which the stopping rule was determined by the posterior distribution, the improvements were most substantial for attribute patterns with fewer attributes, and for the higher minimax requirement.

As the minimax requirement was increased, the JSD generally required just a few more items, whereas random selection required many more items, particularly for attribute patterns with fewer attributes. Under some conditions, namely those that corresponded to lower correct classification rates on the fixed-length tests, the maximum number of items administered to meet the posterior minimax requirement was excessive (i.e., well over 100), but this was never the case for the JSD.

The overall item usage for each of the six attribute patterns for the C-DINA and C-G-DINA models discussed above appeared to be somewhat different. The C-DINA model administered more complex items to examinees with more attributes, whereas single-attribute items, and sometimes two-attribute items, were dominant for the C-G-DINA model. The JSD rarely chose items with three or more attributes.

Although this simulation study explored a variety of important factors, there are still a number of issues that remain to be investigated. One example is that the item bank was perfectly balanced and used items of all possible types. This is likely a substantial simplification of reality, but it was necessary in the early stages of development in an effort to better understand the behavior of the algorithm. The items were also designed so that the average discrimination of a test was similar across the models, but this depends on both the attribute pattern distribution and the composition of the test. Adjusting either of these factors could yield different results. Additionally, future studies could compare the JSD to the GDI when applied to continuous data. Although a loss of information can result from using the GDI when the variances of the response distributions are different, the degree to which this would degrade results is unknown.

We offer a final note on the support of the JSD. Specifically, the JSD can range from 0 to $\log(w)$ (Castner, 2014), where w is the number of probability distributions being compared, meaning that the JSD can be larger for items with more attributes under the C-G-DINA model. One implication of this is that items with more attributes could have higher JSDs than items with a smaller number of attributes. To place all computations on a similar scale, JSD values could be normalized by dividing by their maxima, namely $\log(w)$.

Such an adjustment would fix the range from 0 to 1, but those values would only be comparable if the JSD demonstrated linear behavior across the entirety of its range, which is unknown at this time. Nonetheless, making all JSD values

comparable may make the values for items with a smaller number of attributes larger relative to items with more attributes, possibly skewing the preference of an adjusted version of this algorithm more towards single-attribute items. The benefit of adjusting the JSD for the C-G-DINA model, however, is unclear, because the JSD always performed equivalently to or better than random item selection and also usually administered one- and two-attribute items. Any adjustments made of this type may further skew the item preference towards single-attribute items, at which point item exposure may become a concern.

References

Barrada, J., Olea, J., Ponsada, V., & Abad, F. (2009). *Test overlap rate and item exposure rate as indicators of test security in CATs.* In Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved from http://iacat.org/sites/default/files/biblio/cat09barrada.pdf.

Castner, J. A. (2014). *Measures of cognitive distance and diversity.* Available at http://www.columbia.edu/ jac2130/Research/MeasuresOfCognitiveDistance AndDiversity.pdf.

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

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

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

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

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

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

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

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics, 37*, 655-670.

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

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

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*(3), 167-188.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *Information Theory, IEEE Transactions on, 37*(1), 145-151.

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

Minchen, N. D., & de la Torre, J. (2016, July). *The continuous G-DINA model and the Jensen-Shannon divergence.* Paper presented at the International Meeting of the Psychometric Society, Asheville, NC.

Minchen, N. D., de la Torre, J., & Liu, Y. (in press). A cognitive diagnosis model for continuous response. *Journal of Educational and Behavioral Statistics.*

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Revuelta, J. (1995). *El control de la exposici6n de los items en tests adaptativos informati- zados [Item exposure control in computerized adaptive tests].* Unpublished master's dissertation, Universidad Autonoma de Madrid, Spain.

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

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

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review, 5*(1), 3-55.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8,* 481-520.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child development, 46*(2), 1-84.

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

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*(1), 1-9.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*, 181-204.

van der Maas, H. L. J., & Jansen, B. R. J. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology, 85*, 141-177.

Xu, X., Chang, H., & Douglas, J. (April 2003). *Computerized adaptive testing strategies for cognitive diagnosis.* Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.

Xu, G., Wang, C., & Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 69*, 291-315.

Chapter 4

Study III: A Q-Matrix Validation Method for Continuous Response CDMs

Abstract

An integral component of many cognitive diagnosis models is the Q-matrix, which specifies the attributes measured in each item. Because the Q-matrix is constructed by experts and is subject to error, an important area of research within cognitive diagnosis modeling is to validate its specifications. An array of statistical methods have been developed for this purpose, but most have been developed in the context of a binary response. However, due to the continuous nature of the response variable under recently proposed continuous response models, a new method is required. In this paper, such a method is developed for use in the context of a generalized continuous response model that can also be applied to constrained versions of the model. A simulation study was carried out to analyze its performance, and a real data example was included. Results from the simulation study demonstrated the method's viability, and showed that its performance improved as item quality increased.

Keywords: cognitive diagnosis models, continuous response, Q-matrix validation,

DINA model, C-DINA model, G-DINA model, C-G-DINA model

A Q-MATRIX VALIDATION METHOD FOR CONTINUOUS RESPONSE CDMS

4.1 Introduction

The purpose of many traditional assessments is to rank-order students. In doing so, students can be compared with one another, as is done in norm-referenced tests, or they can be compared against some fixed criterion, as is done in criterion-referenced tests. Such tests serve an important function in educational assessment, but they are generally not designed to provide diagnostic information; rather, they usually provide a single score on a continuous scale. Although attempts have been made to extract diagnostic information they have had limited success (e.g., de la Torre, 2012; de la Torre & Karelitz, 2009; de la Torre & Minchen, 2014).

Conversely, cognitively diagnostic assessments (CDAs; de la Torre & Minchen, 2014) are designed from their very inception to be diagnostic, and yield score profiles that report students' mastery and non-mastery on a set of discretely-defined attributes. Based on such information, teachers can modify their classroom instruction to best serve students' needs. Whereas traditional assessments are typically analyzed using item response theory (IRT) or classical test theory (CTT), CDAs require an alternative class of statistical models referred to as cognitive diagnosis models (CDMs).

Like IRT models, CDMs are item response models, but rather than estimating levels of a continuous latent trait, they group examinees into latent classes. In educational tests, these classes represent distinct combinations of skills. CDMs can also be applied in medical settings to determine patients' clinical disorders (Templin & Henson, 2006; de la Torre, van der Ark, & Rossi, 2015), as well as in situational judgement testing (Sorrel et al., 2016).

A critical component of many CDMs is the Q-matrix (Tatsuoka, 1983), which is usually a binary loading matrix that identifies which skills are measured in each item. The Q-matrix is constructed in the test development phase (e.g., Tjoe & de la Torre, 2013a; Tjoe & de la Torre, 2013b; Tjoe & de la Torre, 2014) and is assumed to be correct in most analyses; however, if its entries are not all correct, model misfit may result. Therefore, procedures have been developed to validate its entries.

Many Q-matrix validation procedures, some of which will be discussed later, have been developed for use with binary responses, but polytomous (de la Torre, 2009a; Ma & de la Torre, 2016) and continuous (e.g., ; Minchen & de la Torre, 2016; Minchen, de la Torre, & Liu, in press) responses exist and have been the topic of a growing body of research. One readily-available continuous response in computer-based testing programs is response time, which is nearly free to capture (van der Linden, 2006). Although care must be exercised in the way response time is used (van der Linden, 2006), it is additional information from a psychometric standpoint. Van der Linden (2007) introduced a framework to use response time to improve ability estimation in the context of IRT.

Other continuous response types include "probability testing," in which examinees estimate the probabilities that various alternatives are correct. Typical multiple-choice questions without partial credit can be viewed as a special case of probability testing in which the chosen answer is assigned a 100% probability of being correct. Probability testing can reveal more information per question (de Finetti, 1965). Another type of continuous response measure is simply to place a mark on a continuum indicating one's level of endorsement. Such a response format could be approximated by a Likert scale, for which a graded-response model may be appropriate, but in the case of a sufficiently large number of response categories, the variable essentially becomes continuous (Samejima, 1969). Noel and Dauvier (2007) and Noel (2014) developed IRT models for such a response

format.

Although continuous response measures have been considered in IRT, they have received only limited attention in the CDM literature. Minchen, de la Torre, and Liu (in press) recently proposed a CDM that takes continuous measures as input. They demonstrated the model's viability through a simulation and showed its applicability to a real data set. Their study, as in many studies, assumed that the Q-matrix was correctly specified, which may not in fact be true. To address this concern, this paper presents a Q-matrix validation procedure based on standard regression and model selection procedures for continuous response CDMs. In addition to the results from a simulation study, a real data example is presented. Before discussing the method in detail, a brief review of the CDMs relevant to this work is offered.

4.2 Cognitive Diagnosis Models

CDMs are restricted latent class models through which examinees' class memberships are estimated. These classes are defined as the set of permutations of skill patterns permissible under the attribute structure. In some cases, the set may include all combinations, whereas in other situations, such as an attribute structure in which mastering one skill presupposes the mastery of another skill, only a subset of skill patterns may be permissible.

As mentioned earlier, the Q-matrix represents the skills required for each item, and is used in most CDMs. The Q-matrix is of dimension $J \times K$, where J represents the number of items on the test, and K represents the number of skills measured on the test. Entries in the Q-matrix are denoted by q_{jk}, $j = 1, \ldots, J$ and $k = 1, \ldots, K$, and are 1 if the j^{th} item requires the k^{th} attribute and 0 otherwise. As mentioned before, the Q-matrix is generally assumed to be observable, known, and correct because it is determined during test development

in a lengthy and iterative process that involves researchers, educators, examinees, and psychometricians (Tjoe & de la Torre, 2013a; Tjoe & de la Torre, 2013b; Tjoe & de la Torre, 2014). Because judgement is involved, statistical methods have been developed to verify the entries of the Q-matrix. Some of these methods will be reviewed later.

The i^{th} examinee's attribute pattern is represented by a binary K-length vector, denoted by $\boldsymbol{\alpha}_i$, where $i = 1, \ldots, N$. Unlike the entries in the Q-matrix, attribute patterns are unobservable (i.e., latent) and need to be estimated. Examinees are partitioned into a maximum of $L = 2^K$ latent classes, including a class for examinees who possess none of the K attributes. The probability of success for examinee i on item j is a function of both $\boldsymbol{\alpha}_i$ and $\boldsymbol{q}_j$, the nature of which is defined by the specific CDM.

The DINA Model

The deterministic inputs, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model is a parsimonious CDM. With $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ as the DINA-specific latent response variable, the probability of a correct response is given as

$$P(X_{ij} = 1 | \boldsymbol{\alpha}_i, s_j, g_j) = (1 - s_j)^{\eta_{ij}} (g_j)^{1-\eta_{ij}}, \tag{4.1}$$

where g_j and s_j are the guessing and slip parameters for item j, respectively. From the definition of η_{ij}, it can be shown that each DINA item partitions examinees into two latent groups: Those examinees who possess all attributes that the item requires, and those who are missing at least one required attribute, for whom $\eta_{ij} = 1$ and 0, respectively.

Although the DINA model may be too simple to accurately represent reality (de la Torre, 2011; Henson & Douglas, 2005), it is both readily interpretable and

parsimonious, making it simple with which to work. De la Torre (2009), and Culpepper (2015) provide more details on marginal maximum likelihood estimation and Bayesian estimation of the model, respectively.

The G-DINA Model

In an effort to relax the strict and perhaps unrealistic constraints of the DINA model, de la Torre (2011) proposed the *generalized*-DINA (G-DINA) model, in which the DINA model's assumption that all attribute patterns contained in the $\eta_j = 0$ group have the same probability is relaxed. Specifically, the G-DINA model allows for some of these attribute patterns (i.e., those that differ on the measured attributes) to have unique probabilities of success.

Let $K_j^* = \sum_{k=1}^{K} q_{jk}$ denote the number of attributes required to solve item j. Then, define $\boldsymbol{\alpha}_{gj}^*$ to be a K_j^*-length vector that retains only the entries from $\boldsymbol{\alpha}_l$ for which $q_{jk} = 1$, where $g = 1, \ldots, 2^{K_j^*}$. The $2^{K_j^*}$ $\boldsymbol{\alpha}_{gj}^*$ vectors are referred to as the *reduced attribute patterns* and represent the *reduced latent groups* under item j. Then, the probability of a correct response for each reduced latent group is expressed as a function of the effect of mastering a given combination of the attributes required for that item, as in

$$P(X = 1 \mid \boldsymbol{\alpha}_{gj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{gk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{gk}\alpha_{gk'} + \ldots$$

$$\cdots + \delta_{12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{gk}, \tag{4.2}$$

where δ_{j0} is the intercept, δ_{jk} are the main effects due to mastering α_k, and all other δ parameters are the $2-$ to K_j^*-way interaction effects. Because the individual contribution of each attribute and all $2-, 3-, \ldots, K_j^*-$way interactions are

modeled, the G-DINA is a saturated model. De la Torre (2011) also defined the additive-CDM (A-CDM), which can be obtained by constraining all interaction terms in the G-DINA model to 0, leaving only the main effects and the intercept. Other additive models can be defined from other saturated models with different link functions.

The C-DINA Model

The continuous-DINA (C-DINA; Minchen, de la Torre, & Liu, in press) is a CDM that takes continuous responses, rather than discrete (e.g., dichotomous or polytomous) responses, as the input. The model's latent response variable is defined as it is in the DINA model, namely, $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$. Due to the continuous nature of the response, an integral is used to write the the item response function (IRF), which is given as

$$P(T_{ij} \leq t|\boldsymbol{\alpha}_i) = \int_0^t [f_{j0}(t_{ij})]^{1-\eta_{ij}} [f_{j1}(t_{ij})]^{\eta_{ij}} dt_{ij}, \tag{4.3}$$

where

$$f_{j\eta}(t_{ij}) = \frac{1}{t_{ij}\sqrt{2\pi\sigma_{j\eta}^2}} \exp\Big[- \frac{(\ln t_{ij} - \mu_{j\eta})^2}{2\sigma_{j\eta}^2}\Big], \tag{4.4}$$

where T_{ij} is the response of examinee i on item j, $f_{j\eta}(t_{ij})$ is the distribution of responses for group η on item j, and $\mu_{j\eta}$ and $\sigma_{j\eta}^2$ are the item parameters for $f_{j\eta}(t_{ij})$. Whereas in typical CDMs a correct response probability is estimated, the C-DINA model estimates a response *distribution* for each latent group, namely, the lognormal probability density function in Equation 4.4. Each function is defined by $\mu_{j\eta}$ and $\sigma_{j\eta}$, resulting in a total of $2 \times 2J = 4J$ structural parameters, whereas the DINA model only requires $2J$ parameters. The item parameters, $\mu_{j\eta}$ and $\sigma_{j\eta}^2$, are estimated as the posterior-weighted mean and variance of the

logarithm of the response times for item j, respectively, for groups $\eta = 0$ and 1.

The C-G-DINA Model

The continuous generalized DINA (C-G-DINA; Minchen & de la Torre, 2016) model generalizes the C-DINA model in exactly the same way that the G-DINA generalizes the DINA model. Rather than partitioning examinees into just two groups, the C-G-DINA partitions them into $2^{K_j^*}$ latent groups on each item. The resulting model is a straightforward extension of the C-DINA. Its IRF is defined as

$$P(T_{ij} \leq t|\boldsymbol{\alpha}_i) = \int_0^x f_{j\eta}(t_{ij})dt_{ij}, \tag{4.5}$$

where

$$f_{j\eta}(t_{ij}) = \frac{1}{t_{ij}\sqrt{2\pi\sigma_{j\eta}^2}} \exp\Big[-\frac{(\ln t_{ij} - \mu_{j\eta})^2}{2\sigma_{j\eta}^2}\Big], \tag{4.6}$$

and where $\eta = 1 \dots 2^{K_j^*}$. Thus, the C-G-DINA estimates lognormal response distributions for each of the $2^{K_j^*}$ groups. Parameter estimation is similar to that of the C-DINA model, except that there are $2 \times 2^{K_j^*}$ parameters for each item.

4.3 Q-Matrix Validation

A variety of methods exist to validate the Q-matrix. De la Torre (2008) developed the sequential EM-based δ-method, designed to work with the DINA model. In this method, the correct Q-vector for an item minimizes the sum of its slip and guessing parameters. As a result, the difference in the correct response probabilities of the two groups is maximized. This method was subsequently generalized by de la Torre and Chiu (2016) for use in the G-DINA model, who developed the G-DINA model discrimination index (GDI). The GDI is defined as the weighted

variance of the probability of a correct response for each of the reduced latent groups under the candidate q-vector, where the weights are equal to the proportion of examinees contained in each of the groups. Using their framework, the fully specified q-vector maximizes the GDI; their procedure is to choose the q-vector that accounts for a sufficient proportion of the variance relative to the fully-specified q-vector.

Liu, Xu, and Ying (2012) developed a method in which the Q-matrix is learned from the data. Their method does not attempt to correct misspecified Q-matrix entries; instead it attempts to build the entire Q-matrix from scratch. The centerpiece of their method is the latent response variable, which summarizes the interaction between the person attributes and the item attributes as discussed earlier. Although they have only tested their method on the DINA model, they indicate that extensions to other models, such as the deterministic inputs, "or" gate (DINO; Templin & Henson, 2006), can be made. However, it remains to be seen how the method will perform with models of a different class (i.e., continuous response models) or more complex models.

Chiu (2013) developed the Q-matrix refinement method, which is based on Chiu and Douglas' (2013) nonparametric classification method, and has been shown to be effective, even with relatively high levels of misspecification in the Q-matrix. However, the method is based on the residual between the response X_{ij} ideal response $\boldsymbol{\eta}_{ij}$. In a continuous models such as the C-DINA and C-G-DINA, the ideal response only denotes the group, not the expected response.

DeCarlo (2012) also presented a Q-matrix validation procedure, but in his method, certain entries in the Q-matrix are thought of as random variables and are estimated with model parameters. This method has been shown to work well with a reparameterized version of the DINA model. The limitation of this method, though, is that the elements of the Q-matrix to be validated must be determined a priori; the extent to which this is a drawback depends on the application.

Misspecifications in the portion of the Q-matrix not under consideration appear to negatively affect the recovery of the uncertain elements.

In the language of factor analysis, these Q-matrix validation methods can be categorized as either exploratory or confirmatory, where the latter contains only DeCarlo's (2012) method of those discussed, because it only validates predetermined entries. (His method, however, can also be used in an exploratory manner.) Exploratory methods can be either parametric or nonparametric, and can be based on clustering, statistical learning, or some type of index, each of which has its own advantages and disadvantages. An important note, however, is that each of these methods have been developed in the context of dichotomous response data.

4.4 Proposed Method: Weighted Least Squares Q-Matrix Validation Procedure

The proposed weighted least squares (WLS) method is based on the logic of model selection in regression. At the heart of the method is the idea that the correct latent groupings of examinees should produce the most homogeneous groups. In other words, the within-group variation should be at a minimum. This is an exhaustive search method that evaluates every possible q-vector for all candidate items. Using this method, each candidate q-vector partitions examinees into $2^{K_j^*}$ (as defined earlier) latent groups for each item. Note that the C-G-DINA model is implicit in this algorithm.

The method chooses the q-vector (e.g., grouping) that results in response distributions that are maximally homogeneous with the smallest number of specifications, which is defined as the *correct* q-vector by de la Torre and Chiu (2016). Other *appropriate* (de la Torre & Chiu, 2016) q-vectors with more specifications can also be found that result in homogeneous groups, but they would require

additional parameters. The method is outlined as follows:

1. Data are obtained either from an actual test or from a simulation. In a simulation study, the data are generated according to some CDM and a Q-matrix that is presumed to be correct. Next, the model is estimated with a provisional Q-matrix, which may contain misspecifications. The resulting posterior distribution for each examinee is captured, which is denoted as $P(\boldsymbol{\alpha}|\mathbf{T}_i)$ and is of dimension 1×2^K.

2. For each item j, there are a total of $2^K - 1$ possible q-vectors of length K that can be formed for a given value of K, ($\mathbf{0_{1\times K}}$ is not included), only one of which is correct. Of these q-vectors, one of them will be $\boldsymbol{q}_f = \mathbf{1_{1\times K}}$, which will be referred to as the *full* q-vector. The remaining $C = 2^K - 2$ q-vectors contain at least one zero. Denote q-vectors in this set as $\boldsymbol{q}_c$, where $c = 1, \ldots, C$. A single $\boldsymbol{q}_c$ is selected.

3. The posterior distribution obtained in step 1 is then collapsed based on the latent groupings formed by the candidate q-vector, resulting in $2^{K_c^*} < 2^K$ groups, where $K_c^* = \sum_{k=1}^{K} q_{ck}$. The resulting distribution will be referred to as the *candidate posterior distribution*, denoted as $P_c(\boldsymbol{\alpha}|\mathbf{T}_i)$, is of dimension $1 \times 2^{K_c^*}$, and whose elements are computed as

$$P_c(\boldsymbol{\alpha}_{\eta_c}|\mathbf{T}_i) = \sum_{\boldsymbol{\alpha}_l:\eta_c=\eta} P(\boldsymbol{\alpha}_l|\boldsymbol{T_i}). \tag{4.7}$$

   where $\eta_c = 1, \ldots, 2^{K_c^*}$ are the reduced latent groups formed under $\boldsymbol{q}_c$.

   As an example, Table 4.1 shows the the process of forming the candidate posterior distribution when $K = 3$ and $\boldsymbol{q}_c = [110]$. Because the third attribute is not required under this q-vector, it becomes irrelevant, which is shown in the middle set of columns. In the next set of columns to the right, this attribute is removed entirely, which results in a reduction of

attributes. Here, latent classes 1 and 4, 2 and 6, 3 and 7, and 4 and 8 become indistinguishable, resulting in four reduced latent groups. The elements of the posterior obtained in step 1 associated with indistinguishable latent groups are summed, resulting in a candidate posterior distribution.

Table 4.1: Forming the Candidate Posterior Distribution with q-vector [110]

Full									Reduced		
Class	$\boldsymbol{\alpha}$								$\boldsymbol{\alpha}^*$		Group
1	0	0	0	$\rightarrow$	0	0	0	$\rightarrow$	0	0	1
2	1	0	0	$\rightarrow$	1	0	0	$\rightarrow$	1	0	2
3	0	1	0	$\rightarrow$	0	1	0	$\rightarrow$	0	1	3
4	0	0	1	$\rightarrow$	0	0	1	$\rightarrow$	0	0	1
5	1	1	0	$\rightarrow$	1	1	0	$\rightarrow$	1	1	4
6	1	0	1	$\rightarrow$	1	0	1	$\rightarrow$	1	0	2
7	0	1	1	$\rightarrow$	0	1	1	$\rightarrow$	0	1	3
8	1	1	1	$\rightarrow$	1	1	1	$\rightarrow$	1	1	4

4. The candidate posterior-weighted mean log responses are obtained as

$$\mu_{\eta_c} = \frac{\sum_{i=1}^{N} ln(t_{ij}) P_c(\boldsymbol{\alpha}_{\eta_c}|\mathbf{T}_i)}{\sum_{i=1}^{N} P_c(\boldsymbol{\alpha}_{\eta_c}|\mathbf{T}_i)}. \tag{4.8}$$

In this instance, weighting effectively partitions each examinee's response into each reduced latent group according to the posterior probability of residing in that group.

5. The differences between the log responses and the mean posterior-weighted log responses computed in Equation 4.8 in step 4 are obtained, squared, and summed across examinees and latent groups. This quantity is the sum of squared errors for $\boldsymbol{q}_c$, and is given by

$$SSE_c = \sum_{\eta_c=1}^{2^{K_c^*}} \sum_{i=1}^{N} P_c(\boldsymbol{\alpha}_{\eta_c}|\mathbf{T}_i) \Big[ln(t_{ij}) - \mu_{\eta_c} \Big]^2. \tag{4.9}$$

Note that the log responses for each latent group will be normal under the

correct q-vector.

6. Similarly, the sum of squared errors for the full model is computed as

$$SSE_f = \sum_{l=1}^{2^K} \sum_{i=1}^{N} P(\boldsymbol{\alpha}_l|\mathbf{T}_i)\Big[ln(t_{ij}) - \mu_l\Big]^2, \tag{4.10}$$

where

$$\mu_l = \frac{\sum_{i=1}^{N} ln(t_{ij})P(\boldsymbol{\alpha}_l|\mathbf{T}_i)}{\sum_{i=1}^{N} P(\boldsymbol{\alpha}_l|\mathbf{T}_i)}. \tag{4.11}$$

Note that the computations for SSE_f and μ_l are identical to those for SSE_c and μ_{η_c}, respectively, except that the original posterior distribution obtained in step 1 is used to compute the former. This is because there is no collapsing of latent classes under $\boldsymbol{q}_f$.

7. The total sum of squares is computed by subtracting the mean of the log responses from each log response, squaring, and summing over individuals. This quantity is given by

$$SST = \sum_{i=1}^{N} \Big[ln(t_{ij}) - \mu_j\Big]^2, \tag{4.12}$$

where

$$\mu_j = \frac{\sum_{i=1}^{N} ln(t_{ij})}{N}. \tag{4.13}$$

8. Using SSE_c, SSE_f, and SST, each candidate q-vector's proportion of variance explained can be computed as

$$R_c^2 = 1 - \frac{SSE_c}{SST}, \tag{4.14}$$

and

$$R_f^2 = 1 - \frac{SSE_f}{SST}. \tag{4.15}$$

9. The ΔR^2 F-statistic is computed for each of the C candidate q-vectors as

$$F_c = \frac{R_f^2 - R_c^2}{df_M(F) - df_M(R)} \div \frac{1 - R_f^2}{df_E(F)}, \tag{4.16}$$

and follows an $F[df_M(F) - df_M(R), df_E(F)]$ distribution, where the degrees of freedom of the full model is $df_M(F) = 2^K$, the model degrees of freedom of the reduced model is $df_M(R) = 2^{K_c^*}$, and the error degrees of freedom of the full model is $df_E(F) = N - 2^K$.

10. P-values are obtained for each of the C ΔR^2 F-statistics, denoted as p_c. A Bonferroni correction is applied by setting $\alpha^* = \alpha/g$, where $g = 2^C$. Define

$$W = \{\boldsymbol{q}_c | p_c > \alpha^*\}, \tag{4.17}$$

and

$$Z = \{\boldsymbol{q}_c \in W : K_c^* = K_{c'}^* \quad \forall \quad c \text{ and } c'\}. \tag{4.18}$$

11. The decision rule is to choose

$$\hat{\boldsymbol{q}}_j = \begin{cases} \boldsymbol{q}_c : K_c^* = arg\ min_c \ \forall\ \boldsymbol{q}_c \in W, & \text{if } |W| \geq 1 \text{ and } |Z| = 0 \\ \boldsymbol{q}_c : p_c = arg\ max_c \ \forall\ \boldsymbol{q}_c \in Z, & \text{if } |Z| \geq 1 \\ \mathbf{1}_{1\times K}, & \text{if } |W| = 0 \end{cases}, \tag{4.19}$$

where $|\cdot|$ denotes the number of elements in a set.

12. Repeat steps 2-11 for all items under consideration.

The full model, $\boldsymbol{q}_f$, will necessarily produce the highest R^2 value because it contains more predictors than any other $\boldsymbol{q}_c$ (i.e., it is appropriate). Thus, the most parsimonious model (i.e., the q-vector with the smallest number of specifications) that is not significantly different from the full model is chosen.

This procedure is essentially a regression-style model selection method in which the more complex models are penalized through a loss in the error degrees of freedom. If the generating model is known to be C-DINA, then the problem is simplified. All q-vectors will partition examinees into exactly two groups when applied to the C-DINA model; thus, all q-vectors represent models with equal degrees of freedom: $N - 2$. In such a case, the analogous model selection method would be to simply choose the q-vector that results in the largest R^2 value or, equivalently, the smallest SSE value. This method was explored in the simulation study and real data examples and will be referred to as the *Max R^2 method*.

Note that this procedure appeals to the robustness of the F-test to the violation of the homogeneity of variance assumption. A slightly different version of the F-statistic may be used that avoids this assumption by approximating the distribution of the sum of variances and using the Satterthwaite correction (1946) to compute the degrees of freedom. This statistic, however, may violate the assumption that the chi-squared statistics in the numerator and denominator of the F-statistic are independent. One potential way to minimize the degree of violation of this assumption would be to classify examinees rather than partitioning their responses into the various latent classes based on their posterior distributions. Using only the mode of the posterior rather than the entire distribution, however, could result in a loss of information.

We make a final note about the similarity of the proposed method to the GDI (de la Torre& Chiu, 2016). Each method attempts selects a q-vector on the basis of maximizing the variability in the responses of the latent groups. However, the key difference between the two methods is that the WLS method

works directly with the posterior-weighted responses, whereas the GDI uses the posterior-weighted correct response probabilities; the former are observed, the latter are estimated. A secondary difference is that the WLS method is based on a statistical test, whereas the GDI relies on a user-specified cutoff.

Design and Analysis

The goal of the simulation study was to evaluate the performance of the WLS method while manipulating item quality (low, medium, and high discrimination), sample size ($N = 500, 1000, 2000$), number of Q-matrix misspecifications $(8, 15)$, and generating model (C-DINA and C-G-DINA). The Q-matrix shown in Table 4.2 was used for all conditions. The WLS method was tested by misspecifying the Q-matrix, applying the method to each item, and observing the results. The results for the C-DINA and C-G-DINA models were compared and, in the case of the C-DINA model, the WLS method was compared to the Max R^2 method. The method was also applied to a real data set.

Misspecifications were chosen at random and were limited to one per item. Misspecifying 8 and 15 of the 150 entries in the Q-matrix amounts to just over 5% and exactly 10% of entries, resulting in just over a quarter and exactly half of the items being misspecified, respectively. In a procedure that will be discussed next, the discriminations were chosen such that the results for the C-DINA and C-G-DINA models would be comparable.

To manipulate item quality in the DINA model, the guessing and slip parameters would be adjusted. Smaller values for those parameters would correspond to higher quality items because each of the groups responds both more consistently and differently from each other. The same reasoning holds in continuous response models. C-DINA items can be made more discriminating by choosing distributional parameters that result in the two lognormal distributions being further apart (Minchen, de la Torre, & Liu, in press). For the generalized models,

Table 4.2: Simulation Study Q-matrix

	Attribute						Attribute				
Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0	16	0	1	0	1	0
2	0	1	0	0	0	17	0	1	0	0	1
3	0	0	1	0	0	18	0	0	1	1	0
4	0	0	0	1	0	19	0	0	1	0	1
5	0	0	0	0	1	20	0	0	0	1	1
6	1	0	0	0	0	21	1	1	1	0	0
7	0	1	0	0	0	22	1	1	0	1	0
8	0	0	1	0	0	23	1	1	0	0	1
9	0	0	0	1	0	24	1	0	1	1	0
10	0	0	0	0	1	25	1	0	1	0	1
11	1	1	0	0	0	26	1	0	0	1	1
12	1	0	1	0	0	27	0	1	1	1	0
13	1	0	0	1	0	28	0	1	1	0	1
14	1	0	0	0	1	29	0	1	0	1	1
15	0	1	1	0	0	30	0	0	1	1	1

items are made more discriminating for the G-DINA by adjusting the variance of the success probabilities of the latent groups (de la Torre & Chiu, 2016), and for the C-G-DINA, by spacing out the response distributions. Figure 4.1 shows an example of an item with low and high discriminations under each of the aforementioned models. Note that in this example the variances of the distributions were constrained to be similar for the purposes of illustration.

To quantify item discrimination, we used the Jensen-Shannon divergence (JSD; Lin, 1991), which quantifies the total divergence among a system of probability distributions. The JSD was first used in CDM as a computerized adaptive testing item selection algorithm for the C-DINA and C-G-DINA models (Minchen & de la Torre, 2016). In their algorithm, the next item chosen is the one that maximizes the JSD of the item response curves, weighted according to the current estimate of the posterior distribution.

Our use of the JSD here was simply to standardize the levels of discrimination

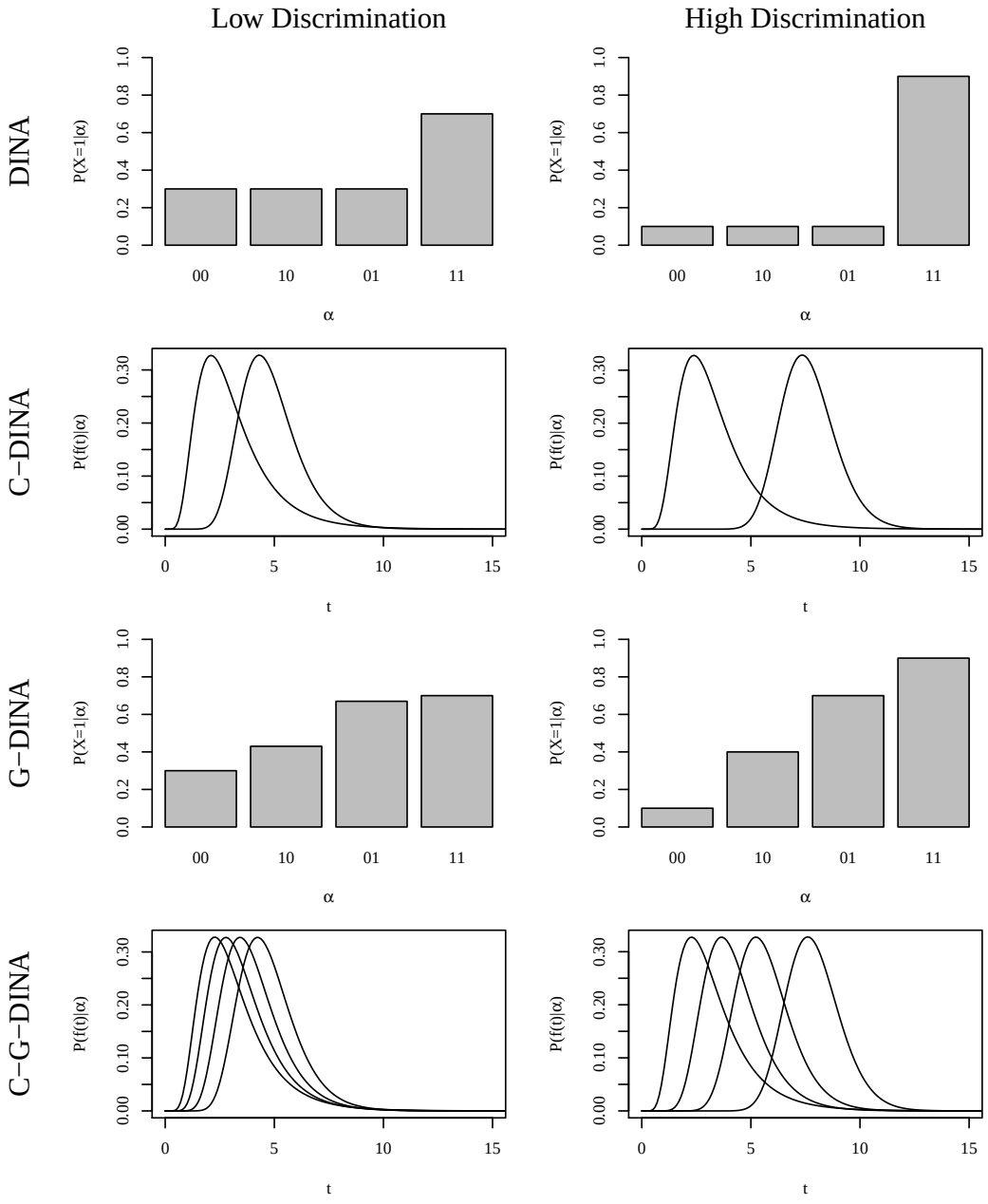

Figure 4.1: Example Items of Low and High Discrimination Under Various CDMs

within a given level for both the C-DINA and C-G-DINA data, which made the results comparable. Item parameters were chosen such that the expected JSD values were approximately equal between the two generating models for all three discrimination conditions at the test level (i.e., the average of all JSDs on the test).

To construct tests for each of the models that had comparable average expected JSDs, two sets of μ_0 and μ_1, one for each of the discrimination conditions, were chosen arbitrarily for the C-DINA model. Corresponding σ_0 and σ_1 parameters were also chosen such that the response curves had a similar height. Parameters for all items were similar, but a small amount of noise was added so that item parameters were not identical, except by chance. Finally, the JSD was computed for all items. The weights used to compute the JSD were the proportion of examinees, that resided in the $\eta = 0$ and 1 groups, assuming a uniform attribute distribution. To find the average test discrimination, JSD values were averaged. These values were approximately .14, .21, and .28 for the low, medium, and high discrimination conditions, respectively.

The next task was to find parameters for the C-G-DINA items that yielded similar average test discriminations to those found under the C-DINA model. Although there were many solutions to this problem, the method we used was to adjust the range of the μ_η parameters, and to equally space them across the range. The lower endpoint used was the same as the lower endpoint used of the C-DINA μ_0. The σ parameters were found such that the curves were approximately the same height as the C-DINA curves. Weights were also assigned to each of the curves for the computation of the JSD, but all weights were equal because of the way the C-G-DINA forms latent groups.

A small simulation study was conducted to compare the classification rates for each of the models for each of the JSD values given above. The correct attribute classification rates for the low, medium, and high discrimination condition were

.915, .966, and .987, respectively, for the C-DINA model, whereas the same rates for the C-G-DINA model were slightly lower at .896, .949, .978, respectively. The correct vector classification rates for the same order of discrimination conditions were .716, .875, and .949, respectively, for the C-DINA model, and .596, .785, and .900, respectively, for the C-G-DINA model. These classification rates suggest that the levels of discrimination are somewhat similar, but not exactly the same.

Results were measured by tabulating the proportion of times that an incorrect q-vector was corrected and that a correct q-vector "corrected" (i.e., misspecified). These quantities are the true positive and false positive rates, respectively. The vector-level true positive rate was computed as

$$T_v^+ = \frac{\sum_{r=1}^{R} \sum_{j \in M} I(\boldsymbol{q}_j = \hat{\boldsymbol{q}}_j)}{\mid M \mid \times R}, \tag{4.20}$$

and the vector-level false positive rate was computed as

$$F_v^+ = \frac{\sum_{r=1}^{R} \sum_{j \notin M} I(\boldsymbol{q}_j = \hat{\boldsymbol{q}}_j)}{[J- \mid M \mid] \times R}. \tag{4.21}$$

The attribute-level true positive rate was computed as

$$T_a^+ = \frac{\sum_{r=1}^{R} \sum_{j \in M} \sum_{k=1}^{K} I(q_{jk} = \hat{q}_{jk})}{\mid M \mid \times K \times R}, \tag{4.22}$$

and the attribute-level false positive rate was computed as

$$F_a^+ = \frac{\sum_{r=1}^{R} \sum_{j \notin M} \sum_{k=1}^{K} I(\boldsymbol{q}_j = \hat{\boldsymbol{q}}_j)}{[J- \mid M \mid] \times R \times K}, \tag{4.23}$$

where M is the set of misspecified q-vectors, $\mid M \mid$ is the size of set M, and $R = 100$ replications. The true and false negative rates can be obtained from these quantities, if desired. Results were averaged over 100 replications; both Q-matrix misspecifications and response data were replicated. All computations

were performed in R (R Core Team, 2015).

4.5 Results

Tables 4.3 and 4.4 display the true and false positive rates for data generated with the C-DINA model using the WLS method, Table 4.5 displays the mean true positive rates for the C-DINA model using the Max R^2 method, and Tables 4.6 and 4.7 display the rates for the same quantities for the C-G-DINA model. Table 4.3 shows that discrimination affected the true positive rate to a greater degree than either the number of misspecifications or the sample size. For example, the vector-level true positive rate for the small sample size condition decreased from .87 to .81 when moving from the 5% to 10% misspecification level, but for either level of misspecification, the rate increased to 100% when the discrimination condition was medium. For the medium and large sample size, increasing the level of misspecification reduced rates from .85 and .85 to .83 and .81, respectively.

As the discrimination increased, the algorithm's ability to determine the correct q-vector increased due to the greater distinctness in the responses of each of the latent groups. Also, the attribute-level true positive rates were always at least as high as their vector-level counterparts, and never less than .96. The reduction in the true positive rates when increasing the level of misspecification was only .01, and was only seen for the low discrimination condition; all other rates were 100%.

This method showed some robustness to the level of misspecification in the Q-matrix, as only small reductions in the true positive rates were seen, and only for the low discrimination condition. Increasing the sample size had a small and inconsistent effect. Near perfect true positive rates were found at medium and high discriminations, for which the method was robust to the number of misspecifications. Finally, greater improvements were seen at the vector level.

Table 4.3: C-DINA WLS Mean True Positive Rates

		T_v^+			T_a^+		
% Misspec.	Disc.	S	M	L	S	M	L
5%	Low	0.87	0.85	0.85	0.97	0.97	0.97
	Med	1.00	1.00	1.00	1.00	1.00	1.00
	High	1.00	1.00	1.00	1.00	1.00	1.00
10%	Low	0.81	0.83	0.81	0.96	0.96	0.96
	Med	1.00	1.00	1.00	1.00	1.00	1.00
	High	1.00	1.00	1.00	1.00	1.00	1.00

Note. Misspec.: Misspecification. Disc.: Discrimination. S: Small sample size. M: Medium sample size. L: Large sample size.

In Table 4.4, false positive rates are tabulated for both misspecification levels, and also for the condition in which no entries were misspecified (0% misspecification), which shows the method's Type I Error. False positive rates were zero (to the second decimal place), with the exception of the low discrimination condition. All false positive rates for the low discrimination condition, however, were well below the significance level of .05. The false positive rates for the 0% misspecification condition were always equal to or lower than those for either the 5% or 10% misspecification conditions, which was likely due to a slight deterioration in the accuracy of the candidate posterior distributions for the conditions in which the Q-matrix was misspecified.

Table 4.4: C-DINA WLS Mean False Positive Rates

		F_v^+			F_a^+		
% Misspec.	Disc.	S	M	L	S	M	L
0%	Low	0.02	0.01	0.02	0.00	0.00	0.00
	Med	0.00	0.00	0.00	0.00	0.00	0.00
	High	0.00	0.00	0.00	0.00	0.00	0.00
5%	Low	0.02	0.02	0.02	0.00	0.00	0.00
	Med	0.00	0.00	0.00	0.00	0.00	0.00
	High	0.00	0.00	0.00	0.00	0.00	0.00
10%	Low	0.03	0.02	0.02	0.01	0.00	0.00
	Med	0.00	0.00	0.00	0.00	0.00	0.00
	High	0.00	0.00	0.00	0.00	0.00	0.00

Note. Misspec.: Misspecification. Disc.: Discrimination. S: Small sample size. M: Medium sample size. L: Large sample size.

Table 4.5: C-DINA Max R^2 Mean True Positive Rates

		T_v^+			T_a^+		
% Misspec.	Disc.	S	M	L	S	M	L
5%	Low	0.96	0.94	0.95	0.99	0.99	0.99
	Med	1.00	1.00	1.00	1.00	1.00	1.00
	High	1.00	1.00	1.00	1.00	1.00	1.00
10%	Low	0.79	0.79	0.79	0.96	0.96	0.96
	Med	0.97	0.97	0.98	0.99	0.99	1.00
	High	0.98	0.99	0.97	1.00	1.00	0.99

Note. Misspec.: Misspecification. Disc.: Discrimination. S: Small sample size. M: Medium sample size. L: Large sample size.

Table 4.5 shows the mean true positive rates for the Max R^2 method for the C-DINA model. With respect to the false positive rates, the Max R^2 method always had a rate of 0 for this set of conditions, whereas the WLS method had false positive rates of 0 only for medium and high discrimination conditions. For the low discrimination condition, the highest false positive rate for the WLS method was for the 10% misspecification, small sample size condition at the vector level (i.e., the least ideal condition and the most stringent performance criteria) of .03, which was still very low.

Next, the comparison of the Max R^2 method to the WLS method with respect to the true positive rates revealed that for the 10% level of misspecification, the WLS method always performed equivalently to or better than the Max R^2 method, with the differences being slightly more notable at the vector level. For example, with a small sample size, the mean true positive vector-level rates were .79, .97, and .98 for low, medium, and high discriminations for the Max R^2 method, whereas the WLS counterparts were .81, 1.00, and 1.00. For the 5% misspecification level, the Max R^2 method outperformed the WLS method only for the low discrimination condition. The vector-level differences were approximately .09 or .10, whereas the attribute level differences were about .02.

Tables 4.6 and 4.7 show the mean true and false positive rates, respectively, for the C-G-DINA model. Taken together, these tables show that the item quality

had the largest effect on the true positive rates compared to the other factors. In examining the true positive rates, the results for the small sample size and 5% level of misspecification condition increased from .56 to .83 and .95 as the discrimination increased from small to medium and large. The rates for the 10% level were very similar. Conversely, increasing the sample size at the 5% level only increased rates by .01 to .03, and the level of misspecification resulted in maximum differences of only about .02.

Unlike Table 4.4, Table 4.7 shows highly elevated false positive rates for the low discrimination condition, and moderately elevated false positive rates for the medium discrimination condition. For the low discrimination conditions, rates were .39 for all levels of misspecification except the 10% level, in which they were .40. For the medium discrimination conditions, rates were between .11 and .13, with the higher rates generally being found in conditions with more misspecifications. For the high discrimination condition, however, false positives were below the nominal level. Again, neither sample size nor level of misspecification had a large effect on the results.

Table 4.6: C-G-DINA WLS Mean True Positive Rates

		T_v^+			T_a^+		
% Misspec.	Disc.	S	M	L	S	M	L
5%	Low	0.53	0.53	0.55	0.89	0.89	0.90
	Med	0.73	0.77	0.77	0.95	0.95	0.95
	High	0.93	0.95	0.93	0.99	0.99	0.99
10%	Low	0.53	0.54	0.54	0.89	0.89	0.90
	Med	0.74	0.76	0.75	0.95	0.95	0.95
	High	0.92	0.93	0.92	0.98	0.99	0.98

Note. Misspec.: Misspecification. Disc.: Discrimination. S: Small sample size. M: Medium sample size. L: Large sample size.

Table 4.7: C-G-DINA WLS Mean False Positive Rates

		F_v^+			F_a^+		
% Misspec.	Disc.	S	M	L	S	M	L
0%	Low	0.44	0.42	0.44	0.09	0.09	0.09
	Med	0.18	0.19	0.18	0.04	0.04	0.04
	High	0.05	0.05	0.05	0.01	0.01	0.01
5%	Low	0.44	0.43	0.43	0.09	0.09	0.09
	Med	0.18	0.18	0.17	0.04	0.04	0.04
	High	0.05	0.07	0.05	0.01	0.01	0.01
10%	Low	0.44	0.44	0.44	0.09	0.09	0.09
	Med	0.21	0.21	0.18	0.04	0.04	0.04
	High	0.05	0.06	0.05	0.01	0.01	0.01

Note. Misspec.: Misspecification. Disc.: Discrimination. S: Small sample size. M: Medium sample size. L: Large sample size.

4.6 Real Data Example

Data Description

In their paper introducing the C-DINA model, Minchen, de la Torre, and Liu (in press) demonstrated the viability of their model on a set of response times (which also included an analysis of response accuracies) to balance scale questions, which was originally collected and subsequently analyzed by van der Maas and Jansen (2003). Each question presented a balance scale, centered on a fulcrum, with one or more equal weights on either side. Examinees were expected to use the positioning and magnitude of the weights to determine if the scale would lean to either side or remain balanced. The combination of the number of weights and their locations determine the type (and difficulty) of the problem.

The final dataset included 146 examinees and 40 questions, and represented a subset of the original dataset. The rationale for the subset can be found in Minchen, de la Torre and Liu (in press). The 40 items were comprised of four types, as shown in Table 4.8. Each of the item descriptions indicates the nature of the questions. For example, conflict-distance means that the side of the scale

with more weights are placed closer to the fulcrum than the weights on the other side. The question cannot be solved by examining only the number and locations of the weights. Rather, the two pieces of information must be integrated using the torque rule. Additional descriptions are available in van der Maas and Jansen (2003). Although it would have been desireable to find a dataset with a greater number of attributes, we were unable to find one that was amenable to this analysis.

Table 4.8: Reduced Q-matrix for the Balance Scale Data

			Attribute	
Item Type	# Items	Description	Distance	Torque
I	10	Simple-distance	1	0
II	10	Conflict-balance B	1	1
III	10	Conflict-distance	1	1
IV	10	Conflict-balance A	1	1

Analysis

All 40 items were examined using the WLS method, and only one correction was made. Item 39, which was a Type III item with $q = 11$, was "corrected" to have $q = 10$. Although it had a slightly lower R^2 value of .13, it was not significantly different than the R^2 value of .16 for the correct q-vector. Unfortunately, the item was not available, so further analysis to remedy the discrepancy cannot be done. However, in this case, the question is whether or not item 39 can be performed without the torque rule. To address this question, the item should be returned to the subject matter experts to investigate whether the question can be solved using an alternative strategy, or perhaps there is a shortcut to solving this problem that does not involve the torque rule.

The Max R^2 method was also applied the data, but it always chose a q-vector of $q_j = 10$. That method, however, should only be considered for data that truly

conforms to the C-DINA. When applying the WLS method to the data, which fits a C-G-DINA model, it was confirmed that the choosing the q-vector that maximized the R^2 value always resulted in choosing $q_j = 11$. Therefore, the WLS method presented in this paper provided results closest to the established Q-matrix for this dataset.

Table 4.9: SSE From Real Data Example

$\mathbf{q}_c$	SSE	R^2	P-value	DF
10	35.97	0.13	0.051	144
11	34.49	0.16	-	142

Finally, the entries of the Q-matrix were randomly misspecified at the 5- and 10-% levels and replicated 100 times, with a maximum of one misspecification per item was permitted. True- and false-positive and negative rates were recorded for each level of misspecification, and are shown in Table 4.10. True positive rates were slightly lower for the 10% level than they were for the 5% level, but the differences were not large. For example, the true positive vector rate only decreased by about .02 when doubling the misspecification rate. Also, as expected, vector-wise rates were lower than their corresponding attribute-wise rates.

False positive rates increased slightly as the level of misspecification increased, but these differences were also not large. For example, the false positive vector rate increased from .04 to .06 when increasing the misspecification rate from 5- to 10%. These findings echo a previous finding, which suggests that the level of misspecifications does not have a dramatic effect on the performance of the method.

4.7 Discussion and Summary

The WLS Q-matrix validation method was proposed and tested. The method was developed in the context of a generalized continuous response model, for which the

Table 4.10: Real Data True and False Positive Rates

	Misspecifications	
Index	5%	10%
T_a^+	.92	.91
T_v^+	.88	.86
F_a^+	.02	.03
F_v^+	.04	.06

Max R^2 method would always choose the fully-specified q-vector. Generally, the WLS method has been shown to be an effective method in Q-matrix validation for both the C-DINA and C-G-DINA. For the C-DINA, the results were somewhat mixed; for the lower level of misspecification, the Max R^2 method outperformed the WLS method, but this trend reversed for the higher level of misspecification.

The WLS method was also compared to the Max R^2 method when the C-DINA model was used to generate the data - the only case in which the application of the latter method would be appropriate. The Max R^2 method was only superior to the WLS method in one condition: for the 5% level of misspecification at the vector level. For all other conditions, the methods performed similarly. The advantage of the WLS method is that it is more flexible - it does not assume a restrictive model like the Max R^2 method does. To date and to our knowledge, the C-DINA and the C-G-DINA models are the only continuous response CDMs that have been developed, but they are also perhaps the simplest and most complex models, respectively, possible under this framework. If other intermediate continuous response models were to be developed that were less complex than the saturated model but more complex than the C-DINA model, the WLS method could prove useful.

For the C-G-DINA model, the method had strong performance when the item quality was high, but the performance diminished as the item quality was decreased, a result that was expected. This trend was also seen with the C-DINA data, but to a smaller degree. Although the test composition was designed to

be similar for both models, the development of a measure for discrimination for the C-G-DINA (i.e., the JSD) is in its infancy and needs additional study to more fully understand its behavior. Specifically, the upper bound of the JSD increases as the number of distributions increases, and it is not entirely clear how the JSD values of two systems with different numbers of probability distributions should be compared. The JSD, however, has been shown to be highly effective as a computerized adaptive testing item selection index (Minchen & de la Torre, 2016), thus warranting our use of it here. However, our conclusion that the method worked better on the C-DINA than it did on the C-G-DINA is tentative.

A finding that was consistent across all the simulation study results was that sample size did not affect performance greatly. The level of misspecification affected the C-DINA results, but did not have much effect on the C-G-DINA results. It is conceivable, however, that increasing the level of misspecification to some larger value may negatively affect the performance of the method, but doing so may also affect the convergence of the models. All replications for all conditions in this study converged.

The real data example demonstrated that the WLS method is very consistent with the opinions of experts. Unfortunately, the items are not available to the public, so it was not possible to examine the one item that the method flagged as incorrectly specified. However, it is possible that there is an alternate strategy to solving this problem that does not employ the torque rule. Further consultations with subject matter experts, teachers, and researchers would be required to resolve this discrepancy. The second part of the real data analysis suggested that the method performs reasonably well on real data when simulating misspecifications, and, again, that the level of misspecification does not have a dramatic effect on performance of the method.

Key to our goal of demonstrating the viability of the method is that this study shows that this method may be useful with real data. The Max R^2 method did

not appear to be helpful. The advantage of the WLS method is that it does not require a model specification (although it implicitly assumes a generalized model), whereas the Max R^2 method explicitly assumes a C-DINA model. To the extent that the C-DINA model does not fit the data, the Max R^2 method will fail. A model fit evaluation may be helpful in determining whether to use the Max R^2 method, but even if the C-DINA fits the data, this method should only be preferred if the number of misspecifications is thought to be small. Further study and model development is necessary to understand the behavior of the WLS method for other applications in continuous response CDMs. The method also could be made iterative, an adaptation that would likely improve performance.

As a final note, it should be reiterated that the proposed method relies on the robustness of the F-test to both nonnormality and heterogeneity of variance, both of which may be violated when the various groupings are made. Corrections could potentially be made to this method, or a nonparametric alternative that does not rely on these assumptions could be considered.

References

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

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

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics, 40*, 454-476.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*, 343-362.

de La Torre, J. (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*, 115-130.

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

de la Torre, J. (2012). Application of the DINA Model Framework to Enhance Assessment and Learning. In M. Mok (Ed.), *Self-directed learning oriented assessments in the Asia- Pacific* (pp. 92-110). New York: Springer.

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

de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement, 46*, 450-469.

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

de la Torre, J., van der Ark, L. A., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation*

in Counseling and Development, 0748175615569110.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*, 447-468.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of statistics, 26*, 979-1030.

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

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

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

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

Lin, J. (1991). Divergence measures based on the Shannon entropy. *Information Theory, IEEE Transactions on, 37*, 145-151.

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

Minchen, N. D., & de la Torre, J. (2016, July). *The continuous G-DINA model and the Jensen-Shannon divergence.* Paper presented at the International Meeting of the Psychometric Society, Asheville, NC.

Minchen, N., de la Torre, J., & Liu, Y. (in press). A cognitive diagnosis model for continuous response. *Journal of Educational and Behavioral Statistics.*

Noel, Y. (2014). A beta unfolding model for continuous bounded responses. *Psychometrika, 79*, 647-674.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement, 34*, 1-97.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*, 110-114.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8,* 481-520.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child development, 46*(2), 1-84.

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

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

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.

Tatsuoka, C., Clements, D. H., Sarama, J., Izsak, A., Orril, C. H., de la Torre, J., & Khasanova, E. (2016). Developing workable attributes for psychometric models based on the Q-matrix. *Psychometric Methods in Mathematics Education: Opportunities, Challenges, and Interdisciplinary Collaborations*, 73-96.

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

Tjoe, H., & de la Torre, J. (2013a). Designing cognitively-based proportional reasoning problems as an application of modern psychological measurement models. *Journal of Mathematics Education, 6*, 17-22.

Tjoe, H., & de la Torre, J. (2013b). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal, 26*, 237-255.

Tjoe, H., & de la Torre, J. (2014). On recognizing proportionality: Does the ability to solve missing value proportional problems presuppose the conception of proportional reasoning? *The Journal of Mathematical Behavior, 33*, 1-7.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*, 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287-308.

van der Maas, H. L. J., & Jansen, B. R. J. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology, 85*, 141-177.

Chapter 5
Conclusion

Cognitively diagnostic assessments (CDAs; de la Torre & Minchen, 2014) are designed from the outset to provide diagnostic feedback with respect to a set of discrete attributes. To translate the responses from CDAs into attribute mastery profiles, cognitive diagnosis models (CDMs) are employed to analyze the response data. Whereas the purpose of CDAs is diagnostic, the purpose of the traditional assessement framework is typically to rank-order students.

Although interest in CDMs has grown over the last two decades, most of the technical advancements have been developed in the context of a dichotomous, or sometimes a polytomous, response. However, continuous response measures also exist and have been studied in IRT, but not extensively in CDM. As such, this dissertation builds on the work of Minchen, de la Torre, and Liu (in press) and Minchen and de la Torre (2016), in which the continuous deterministic inputs, noisy "and" gate (C-DINA) and continuous generalized DINA (C-G-DINA) models were developed, respectively.

In the first study, van der Linden's (2007) hierarchical framework was used as a way to jointly estimate latent variables governing response time and response accuracy. Two separate item response models were used in the framework - one for response time and one for response accuracy. For response time, a lognormal model (van der Linden, 2006) was used, and a person-level latent variable, τ, was estimated. On the response accuracy side of the model, a higher-order attribute distribution was used to provide a statistical link between a general ability and

the probability of mastering attributes, allowing a CDM to be used to model response accuracy. (In van der Linden's (2007) original formulation, the general ability was directly responsible for correct item response probabilities.)

Based on this adaptation, both a general ability, θ, and an attribute pattern, $\boldsymbol{\alpha}$, were estimated for each examinee. The simulation study demonstrated that the inclusion of response time improved the estimation of θ, which was expected. It was also hypothesized that improvements in θ estimates would in turn lead to improvements in classification accuracy. Such improvements in $\boldsymbol{\alpha}$ were observed, and they were more substantial both for lower quality items and at the vector level. One unexpected finding was that the range of $\hat{\theta}$ increased when using response time, resulting in better estimates with less bias. Because one of the benefits of the higher-order attribute formulation is that both $\hat{\boldsymbol{\alpha}}$ and $\hat{\theta}$ are provided in a single model (de la Torre & Douglas, 2004), improving the estimation of θ was not a trivial benefit. The effect of the estimation error in τ was not examined systematically, but could be in future work.

The second study discussed the challenges of attempting to apply existing computerized adaptive testing (CAT) algorithms to a continuous response CDM. In particular, extending such algorithms to a generalized continuous response CDM may not be possible without a substantial loss of information. Thus, the Jensen-Shannon divergence (JSD; Lin, 1991), which quantifies the degree of divergence in a system of multiple probability density functions, was adapted for use as a selection algorithm, where examinees' current posterior estimates were used as weights.

The logic of the proposed procedure is to find the next item that is maximally discriminating for each examinee by computing the degree of posterior-weighted variation in the item response function; this logic is identical to that used by Kaplan, de la Torre, and Barrada (2015) when they adapted the generalized DINA (de la Torre, 2011) discrimination index (de la Torre & Chiu, 2016) for use

as a selection algorithm in CDMs with a dichotomous response. The algorithm was found to be superior to random item selection by most measures, although the benefit of using the JSD algorithm over random item selection diminished for examinees who had more attributes.

A few issues remain to be resolved. One is that the range of the JSD is not fixed; as there are more distributions being compared, the maximum possible JSD grows. Note that this will not be a concern when applying the algorithm to C-DINA data, because all items only have two response distributions. When applied to the C-G-DINA, however, this property could lead to items with more attributes having higher JSDs than items with fewer attributes, but based on item usage, this did not appear to be a serious concern. A number of factors in the study could be adjusted to more fully explore the behavior of the JSD. For example, item parameters in this study resulted in lognormal distributions that spanned approximately the same range and were symmetric, both of which need to be adjusted for greater generalization. To examine the performance of the JSD relative to established selection algorithms, the JSD could be compared to the GDI, which works only with the mean of the distributions. In such a study, the response types and shapes of the response distributions would be manipulated.

In the last study, a Q-matrix validation method for use with continuous response was presented. The performance of the method was most affected by item discrimination (quality). The method implicitly assumes a C-G-DINA model. It appeared that performance when using C-DINA data was better than when using C-G-DINA data, in spite of an attempt to ensure that the average discriminations for both models were similar. This finding warrants additional study into the behavior of the JSD. Also, the groupings of log times formed when testing various q-vectors may not form normal distributions, and they may also not have similar variances, leading to a potential violation of the assumptions of the F-test. It is unclear to what degree such violations reduced the effectiveness of the method,

but this could be explored in future studies. As a final note, with regard to both the second and third studies, the descriptors used for item quality (i.e., "low," "medium," "high") should only be interpreted relative to each other. More real data examples are needed to determine what a "typical" set of item parameters are in practice.

Each of these studies explored applications of continuous response in CDMs. The first study modeled two responses per person per item, one of which was continuous, whereas the second and third studies only modeled one continuous response per person per item. The real data examples from all three studies used time as the continuous response variable. From a psychometric standpoint, response time may provide additional information beyond what the response accuracy provides, or it may be of interest by itself. In the case of the former, prudent researchers must understand the meaning of response time if it is to affect the ability estimates, particularly in high-stakes settings. They must also understand the political consequences of using response time to estimate a different ability. Failing to do so may negatively affect the validity of the scores.

With respect to the first study, other types of models that use response time to aid in ability estimation and/or classification accuracy should also be considered. One such model assumes that the response times arise from examinees' attribute profiles, rather than a separate latent trait representing speed. Based on the example provided using van der Maas and Jansen's (2003) data in Chapter 2 and in Minchen, de la Torre, and Liu (in press), the response times are hypothesized to be a function of attribute mastery, providing theoretical justification for using a single α. A possible complication would be examinees who are attempting to apply a skill but are doing so incorrectly. In such a case, response time data may suggest that an examinee possesses a skill, but response accuracy data suggests the opposite (Minchen, de la Torre, & Liu, in press).

Such a model may make use of the developments in the second and third

studies in this dissertation. For example, Q-matrix validation in this type of a dual-response model may be challenging. Perhaps the method presented Chapter 4 may be used in tandem with an existing procedure for dichotomous response. Similarly, the JSD CAT algorithm presented in Chapter 3 may be useful, particularly if it were shown to perform well for dichotomous data.

Finally, a theme that arose across Chapters 3 and 4 was that additional work is needed to better understand the characteristics of the JSD, particularly with regard to its behavior across its support. Presently, it is not entirely clear to what degree it is nessessary to adjust the scale of the JSD such that the upper limit of its range is always one. Although this may be attractive mathematically, it may be unnecessary, unless C-G-DINA items that measure many attributes have lognormal distributions that are very spaced out. Although this issue can be studied from a theoretical standpoint, more continuous response CDM data needs to be collected to determine the typical ranges of the parameters.

References

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

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

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

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

Lin, J. (1991). Divergence measures based on the Shannon entropy. *Information Theory, IEEE Transactions on, 37*(1), 145-151.

Minchen, N. D., & de la Torre, J. (2016, July). *The continuous G-DINA model and the Jensen-Shannon divergence.* Paper presented at the International Meeting of the Psychometric Society, Asheville, NC.

Minchen, N. D., de la Torre, J., & Liu, Y. (in press). A cognitive diagnosis model for continuous response. *Journal of Educational and Behavioral Statistics.*

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*, 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287-308.

van der Maas, H. L. J., & Jansen, B. R. J. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology, 85*, 141-177.