MCCOY, THOMAS P., Ph.D. The Effects of Mixture-Induced Local Dependence on Diagnostic Classification. (2015)
Directed by Dr. John T. Willse. 144 pp.

Diagnostic Classification Models (DCMs) have been extensively researched in recent psychometric literature for providing mastery skill profiles for diagnostic feedback (Henson, Templin, & Willse, 2009). DCMs are multidimensional confirmatory latent class models (LCMs) where latent classes represent skill mastery profiles and latent attributes are categorical (mastery or non-mastery). DCMs make a central assumption that once mastery profiles are accounted for that items are independent, referred to as local independence (LI). Construct irrelevant variance (e.g., differential item functioning (DIF), speededness, test wiseness, item-to-skill misspecification) or underrepresentation (extra dimensionality, inappropriate definitional grain-size of defined skills) could introduce systematic within-class variation which would violate LI.

Using connections of LCMs with mixture IRT models, this study explores the effects of introducing systematic within-class variation on diagnostic classification. The log-linear cognitive diagnosis model (LCDM) is extended to include continuous abilities, akin to a multidimensional item response theory (MIRT) model with underling mixtures due to skill mastery/nonmastery. Data were then simulated for different ability variances related to distribution overlap conditions. Multiple DCMs are then fit using the LCDM framework in a simulation study. Impact on classification and local dependence detection are summarized. It was found that as mixture overlap increased due to companion ability variance that diagnostic classification in DCMs greatly suffered, but

can be detected by Yen's *Q*3. The relationship of the degree of inaccuracy and effect sizes based on ability variance and group separation is delineated. Recommendations for practitioners are given along with areas for future study.

THE EFFECTS OF MIXTURE-INDUCED LOCAL DEPENDENCE ON DIAGNOSTIC CLASSIFICATION

by

Thomas P. McCoy

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

Greensboro
2015

Approved by

_____
Committee Chair

To Lisa, Mom, Dad, and Chris.

APPROVAL PAGE

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

Committee Chair _____
John T. Willse

Committee Members _____
Randall D. Penfield

_____
Robert A. Henson

_____
Terry A. Ackerman

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

This section could be very long as there are many people who helped me make this possible. To Dr. Willse, I cannot write enough here. Thank you for being an outstanding mentor, listener, expert, and adviser. I know there is no way any of this could have happened without your direction and insight through our countably infinite interactions. I am so grateful to you.

To Dr. Wallace, this was never possible without your support. Thank you so much. To Dr. Henson, your generosity in meeting and providing sage advice was appreciated more than I can write. Thank you so much. To Dr. Cowling, I never would have started this journey without your encouragement and providing opportunity for me. I will never forget that and thank you so much. To Dr. Couper, I did it. To Dr. Barba, your encouragement, conversation, support, and caring will not be forgotten. Thank you so much. To Dr. Penfield, thank you so much for providing keen insight and also making our Department a wonderful place to learn as a student. To Dr. Ackerman, your suggestions and willingness to visit were so much appreciated. Thank you very much for your encouragement and time. To Dr. Chalhoub-Deville, your teachings had a profound impact on my learning and approach to this project. Thank you so much.

To my family, friends, and colleagues, there is not enough space left to write enough to give you due credit. This was not possible without your support, love, and encouragement, especially from my wonderful wife Lisa and both of our families.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Statement of the Problem

Measurement can be negatively impacted by unaccounted for sources of heterogeneity. One source of such heterogeneity could be unobserved (latent) groups within samples of examinees (i.e., "mixtures" of examinees), for which mixture item response theory (IRT) models have been proposed. In presenting estimation methods for the mixture Rasch IRT model, Willse (2011) calls for future research to investigate connections between mixture IRT and latent class models (LCMs). von Davier and Rost (2007) remark that when the variance of the latent ability distribution within the mixture IRT model goes to zero within each mixture, then mixture IRT resembles a LCM. Willse (2011) notes this can be accomplished by fixing the examinee ability parameters to zero, and that when nonzero ability variance is present that this can be conceptualized as a LCM with a violation of local independence (LI). Violating LI in LCMs could have a negative effect on modeling class membership, estimating model parameters, and evaluating model fit, thereby calling into question validity of results.

One kind of LCMs that have grown in stature in recent educational research has been diagnostic classification models (DCMs). DCMs are multidimensional confirmatory LCMs which are item response models with categorical latent variables. The categories of the latent variables have been previously interpreted in educational

literature as skill mastery and non-mastery. DCM estimates have been used to provide diagnostically rich feedback to examinees as a profile of the likelihood of obtaining various skills. The process of assessment design, skill mastery estimation, and reporting is referred to as diagnostic measurement (DM) (Rupp, Templin, & Henson, 2010). Madison and Bradshaw (2014) note that DCMs are well-suited to accomplish DM because of their increased reliability relative to other multidimensional psychometric models and also their practical efficiency.

All uses and interpretations from assessments with non-trivial stakes should be fair, reliable and valid (AERA, APA, NCME, 2014). K-12 schools are increasingly developing or purchasing interim and formative assessments to identify learning deficiencies well before end-of-year summative assessments (Hansen et al., 2010). DM has increasingly been referenced as a way to provide formative diagnostic feedback of skill mastery for remediation purposes to examinees, as well as for instructors of examinees. One reason for the impetus has been the educational and political landscape in recent decades shaped by the Common Core State Standards (CCSS) Initiative based on funding from the Race to the Top program. Two large consortia, the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment Consortium (SBAC), have underlined the role of diagnostic assessment for enhancing learning to increase student achievement (and therefore presumably college and career readiness).

The educational literature is replete with discussions about formative assessment (FA), and a detailed discussion of defining and use are not given here. An interested

reader can be pointed to the excellent review of FA given by Black and William (1998). Heritage et al. (2009) develop a measure of teacher knowledge and examined what instructional method should be implemented during FA processes. They conclude that if teachers do not know what to do when students need remediation then FA has little value. Perie et al. (2009) discuss interim assessment, which are in between end-of-year summative assessment and day-to-day formative assessment (e.g., benchmark). Interim assessment presents special issues such as material not yet being covered or attempting to be reliable mini-versions of the end-of-year assessment. The authors conclude with some skepticism of interim assessment and suggest that perhaps "resources would be better spent helping teachers learn formative assessment techniques." Shepard (2009) concludes that FA purports to raise student achievement but this still must be validated:

> Heritage et al. (2009) showed that teachers generally had difficulty saying what instructional interventions/appropriate next instructional steps should be used given evidence of what a student did or did not understand.

Huff et al. (2007) present on the demand of diagnostic assessment in K-12 education. Results from a nationally representative random sample of 400 teachers about utility of large-scale state-mandated and commercial assessments for diagnostic remediation and instruction are described. They provide recommendations for future directions of DM for K-12 teachers, and compare cognitive psychology to what K-12 educators wish diagnostic assessment could provide them. A non-trivial proportion of teacher never use results from such assessments, and view their own classroom-based assessments to best provide formative feedback (Huff et al., 2007). They conclude with

discussion about lack of direction on use of diagnostic assessment in K-12 education and future directions which might hold promise:

> When assessing the demand for CDA from educators, it is important to recognize that they are not actually demanding that assessment developers use cognitive models as the basis for assessment design and reporting. What educators are demanding is that they receive instructionally relevant results from any assessments in which their students are required to participate and that these assessments be sufficiently aligned with classroom practice to be of maximum instructional value (p. 24).

Thus, aligned assessments for diagnostic purposes are desired. However, DCMs have also been retrofitted to existing non-diagnostic assessments (e.g., large-scale summative proficiency assessments) in multiple studies. Henson et al. (2014) contend while retrofitting to existing assessments not intended for a DM purpose is still perilous, DCMs (and their extensions) can still be applied in prospectively designed multidimensional assessments where a content "blueprint" has been created by subject matter experts to cover a range of content areas within the span of desired knowledge domains. Use of such a blueprint is common in test design practice as a means towards building content validity into assessments. Thus, test developers and educators attempting to prescribe to CCSS are increasingly driving towards delivering assessments that potentially have some multidimensional aspects to them. Rupp et al. (2010) remark that DCMs may provide a more direct method of proficiency diagnosis (because latent categorical attributes are explicitly modeled), which is a common goal of standard-setting. However, the validity of interpretations and uses of findings from DCMs hinge upon the confidence in if underlying assumptions have been adequately satisfied.

Like all LCMs, DCMs make the central assumption of LI. Violation of LI due to systematic within-skill profile variation could induce inaccuracy of estimating skill profile attainment, thereby introducing invalidity into interpretations and use for DM. Potential sources of systematic within-skill profile variation could arise from the latent trait(s) truly being continuous, construct irrelevant variance or construct underrepresentation. Phenomena related to construct irrelevant variance recognized in previous psychometric research (e.g., Henning, 1989; Ferrier et al., 2011) include: differential item functioning (DIF), test-wiseness, speededness, test related anxiety, fatigue, testing conditions, test exposure, item format, and poor item quality. Phenomena related to construct underrepresentation include: extra multidimensionality (including item bundles/testlets) and issues with definitional grain-size of attributes (Rupp et al., 2010).

If one or more of these sources are introduced through mixtures of underlying continuous abilities with nonzero variance then skill diagnoses could suffer, especially when the mixture distributions overlap non-trivially. The current study will investigate this phenomenon, and delineate at what point practitioners should take caution in interpreting DCM results as mixture overlap increases. The phenomenon is introduced through a multidimensional continuous-categorical item response model (MCCIRM) by way of extending the log-linear cognitive diagnosis model (LCDM) of Henson et al. (2009) to additionally include combinations of continuous abilities (cf. Henson, Templin, Willse, and Irwin, 2014).

Contribution of the Current Study

The current study explored on the connections between mixture IRT and LCM suggested by Willse (2011) within the context of DM. The effect of increasing systematic within-skill profile variation using DCMs induced by mixtures of continuous abilities on skill mastery classification is delineated.

Research Questions and Hypotheses

The four study research questions and six hypotheses are as follows:

*RQ*1: Does increasing variance of continuous abilities in mixtures of mastery/non-mastery groups cause detectable violations of local independence when performing diagnostic classification?

*H*1: Increasing variance of continuous abilities generated from the MCCIRM is detected by increasingly large Yen's *Q*3 statistics based on results from DCMs without continuous ability.

*RQ*2: Does increasing variance of continuous abilities degrade model fit when performing diagnostic classification?

*H*2: Increasing variance of continuous abilities within the MCCIRM substantially degrades item parameter recovery in LCDM estimates without continuous ability.

*H*3: Increasing variance of continuous abilities within the MCCIRM leads to overestimation of attribute-to-attribute correlations under the LCDM without continuous ability.

*RQ*3: Does increasing variance of continuous abilities lower accuracy of diagnostic classification?

*H*4: Increasing variance of continuous abilities within the MCCIRM substantially degrades proportion of correct diagnostic classification based on estimated LCDMs without continuous ability.

*RQ*4: Are effects of increasing variance of continuous abilities on accuracy of diagnostic classification affected by complexity or compensation?

*H*5: Complex structure substantially degrades proportion of correct diagnostic classification based on the LCDM without continuous ability when variance of continuous abilities within the MCCIRM increases.

*H*6: Compensatory versus noncompensatory processes leads to substantially lower proportion of correct diagnostic classification based on the LCDM without continuous ability when variance of continuous abilities within the MCCIRM increases.

Assumptions

Throughout this study *skills* are synonymous with *attributes*, although attributes could be more broadly interpreted outside educational research. Attribute *profiles* are synonymous with attribute *patterns*. For notation, it is also assumed throughout that items are represented by the index *i*, examinees by *e*, latent continuous abilities by *t*, and attributes by *a*.

This study investigates mixture-induced local dependence effects on diagnostic classification through one particular choice of psychometric measurement model: the MCCIRM. However, other modeling mechanisms could have been used to investigate the study research questions, such as use of the general diagnostic model (abbreviated GDM; von Davier, 2005). Validity of the current study findings would be in doubt if

conclusions were model dependent. Future research could attend to investigating what, if any, differences arise when comparing MCCIRM versus GDM methods under equivalent conditions.

Diagnostic classification models are assumed to be a useful methodology for skill mastery classification successfully when LI is not violated. However, there are other factors affecting classification performance of DCMs reported by previous research that must be considered. To minimize threats to validity of findings, these other factors examined in current study investigation are held constant or are attended to throughout study planning, execution, and dissemination.

Further, violating LI when applying DCMs is assumed to be possible and can be detected by the empirical methodology employed within the study. Although violating LI by design is performed, the detection of assumption violation will be attempted by existing methods recognized as appropriate in non-DCM scenarios. These methods are selected based on review of current literature.

Additionally, violating the key assumption of LI in DCM is hypothesized to lessen usefulness for its intended purpose: to successfully classify examinees into profiles of skills. Therefore, appropriately gauging performance of DCM through appropriate metrics (e.g., proportion with correct classification) based on accepted modeling and estimation practices is important to investigate. To this end, accepted methods of modeling, estimation, and reporting of classification are reviewed for completeness and appropriateness.

Items with only simple structure or average complexity of two were studied. It is possible in practice that items can be of higher complexity, which could impact results of the current study. Future research could examine scenarios where greater complexity is present.

This study was limited to scenarios of large examinee sample size and with reasonably adequate ratios of items to attributes/abilities. Yet, low-stakes formative assessment has been commonly suggested as one of the promising uses of DCMs. In this case, many items or many examinees may not be available because the formative assessment may be occurring at the grade-level or classroom-level and therefore large-scale testing scenarios are not occurring with such a prospective diagnostic purpose in mind. Additional future studies could investigate how classification is degraded when fewer items are used or when sample size shrinks under equivalent conditions. However, the findings from the current can provide a benchmark upon which to compare such studies to when less data are available.

Another key assumption is that the intended DCM specification is correct otherwise and that all items follow the same DCM specification. If a conjunctive, noncompensatory process is driving item responses then only those DCMs appropriate for these kinds of processes are utilized. That is, the current study is also not concurrently investigating other effects of model misspecification in DCMs.

Effects of findings for linking and equating test forms were not explored in this study. Future research could attend to what consequences LI violations have on these common psychometric testing practices (Xu & von Davier, 2008; Rupp et al., 2010).

Further, it is assumed in this study that no other sources of heterogeneity such as DIF are present. However, DIF can occur in practice whereby many previous researches have studied its effects under the IRT framework. Bozard (2008) studied invariance testing in DCMs. Future investigations can examine what impact DIF and other sources of heterogeneity could have on classification performance when also present concurrently.

This study is limited to a perspective that the implemented assessment has been prospectively designed for diagnostic purposes and that a DCM has not been just retrofitted to a large-scale summated assessment for other purposes. Henson et al. (2014) give a discussion of the latter with many relevant considerations. Rupp and Templin (2008) have discussed limitations related to retrofitting diagnostic models to assessments built for other purposes.

The current study also assumes that the test design item-to-attribute $\underset{\sim}{Q}$ and item-to-ability $\underset{\sim}{C}$ matrices have correct specifications, and that they avoid the problematic properties delineated in Madison and Bradshaw (2014). Many previous authors (Rupp & Templin, 2008b; de la Torre, 2008; Kunina-Habenicht et al., 2012; Chiu, 2013) have discussed impact of *Q*-matrix misspecification on skill mastery classification under DCMs and offered various strategies. Rigorous validation efforts, including qualitative studies of high quality (e.g., think-alouds), are necessary to verify correct *Q*-matrix specification according to item writing, cognitive processes, and skill definition (Madison & Bradshaw, 2014). Future research could adapt these strategies to study their utility under the current study's scenarios when within-mastery class variation is present.

A main assumption of the study methodology is that knowledge acquired from simulation findings will be applicable to real-word situations. This assumption may not have teeth through a variety of threats to validity. One threat is if simulation conditions are not similar to real-world scenarios. This threat is attempted to be mitigated by proper identification and inclusion of real-world situations (i.e., simulation based on published model parameter estimates from applied studies using DCM on real assessment data).

Finally, this study only examined the effects of increasing continuous ability variance for the diagnostic classification enterprise. A similar companion study could be performed in the mixture IRT setting, whereby effects of shrinking continuous ability variance on rank-ordering examinees along ability continuum(s) could be investigated. It could be hypothesized that as continuous ability mixtures become more "discretized" then rank-ordering performance could suffer non-trivially (e.g., Markon & Kruger, 2006).

CHAPTER II

REVIEW OF THE LITERATURE

Literature Review Methods

Current methods of literature review were used according to Martella et al. (2013).

The Connection of Mixture Item Response Theory and the Latent Class Model

von Davier (2005) presents the general diagnostic model (GDM) which he conceptualizes as a generalized partial credit item response theory (IRT) model that could be used as a diagnostic classification model (DCM) when ability is constrained to -1 for nonmasters and +1 for masters. Later, von Davier and Rost (2007) remark that mixture IRT resembles a latent class model (LCM) when variance of the latent ability distribution within the mixture IRT model goes to zero within each mixture. Thus, when person parameters (ability) in mixture IRT are fixed to any constant, the model collapses to a LCM with the number of classes equal to the original number of mixtures. Willse (2011) notes that an alternative interpretation of a mixture Rasch IRT model is an LCM with nonrandom within-class variation, which is a violation of an assumption denoted as local independence (LI). First, IRT and DCM are introduced next and then their connections are further discussed.

Item Response Theory (IRT)

IRT aspires to estimate an examinee's standing on a continuous latent ability and parameters that characterize items (Lord & Novick, 1968; Lord, 1980; Finch, French, & Immekus, 2014). IRT models can be divided into two broad categories: those that model scored items that are dichotomous and those that model polytomous items (Finch et al., 2014). Some researchers view IRT as an improvement over classical test theory (CTT) because item difficulty using p-values, discrimination using corrected point-biserial correlations, and examine sum scores are sample dependent and will change depending on examinee characteristics (Finch et al., 2014). Additionally, error of measurement is assumed constant across the score range (Finch et al., 2014).

Three central assumptions are made in IRT: (a) monotonically increasing relationship of probability of correct response and latent ability, (b) unidimensionality, and (c) local independence (cf. section on *Local Independence in IRT* below) (Finch et al., 2014). The three-parameter logistic IRT model (3PL) is credited to Birnbaum (1957; 1958). It is an IRT model for dichotomous item responses and has the form (Hambleton, van der Linden, & Wells, 2010)

$$P\left(X_{ei}=1\,|\,\theta_e\right)=g_i+\left(1-g_i\right)\frac{\exp\left[a_i\left(\theta_e-b_i\right)\right]}{1+\exp\left[a_i\left(\theta_e-b_i\right)\right]}, \qquad (1)$$

Where:

X_{ei} is the dichotomous item response to the *i*-th item for the *e*-th examinee where $X_{ei} = 1$ for a correct response and $X_{ei} = 0$ for an incorrect response

$e = 1, \ldots, E$ for the *e*-th examinee

$i = 1, \ldots, I$ for *i*-th item

θ_e is the continuous latent trait ("ability") for the *e*-th examinee

g_i is the pseudo-guessing/lower asymptote parameter for the *i*-th item

a_i is the discrimination parameter for the *i*-th item

b_i is the difficulty parameter for the *i*-th item.

If $g_i = 0$ for all items, then the 3PL reduces to the 2PL IRT model. If for the 2PL model a common *a* parameter for all items is specified such that $a_i = a$, then the model becomes a 1PL IRT model. Finally, if $a = 1$, then the IRT model is referred to as the Rasch model (Rasch, 1960). The continuous latent ability can be thought of as a random effect for examinees and often is assumed to have a standard Normal distribution (i.e., $\theta_e \sim Normal(0,1)$).

Multidimensional IRT (MIRT)

When there are multiple latent continuous abilities posited, a MIRT model may be considered. Two types of MIRT models are discussed next: (a) compensatory (CMIRT) and (b) noncompensatory (NCMIRT).

Compensatory MIRT Model (CMIRT)

A form of the CMIRT model is given by (Chalmers & Flora, 2014)

$$P\left(X_{ei}=1\,|\,\underset{\sim}{\theta}_e,\underset{\sim}{a}_i,d_i,g_i\right)=g_i+\left(1-g_i\right)\left(\frac{\exp\left(\sum_{t=1}^{T}a_{it}\theta_{et}+d_i\right)}{1+\exp\left(\sum_{t=1}^{T}a_{it}\theta_{et}+d_i\right)}\right) \quad (2)$$

Here, there is only one overall item difficulty parameter $\left(d_i\right)$ while there one item discrimination parameter $\left(a_{it}\right)$ per latent ability $\left(\theta_{et}\right)$. Because contributions from individual abilities are strictly additive, the model is denoted as *compensatory*.

Noncompensatory MIRT Model (NCMIRT)

The form of the NCMIRT model is given by (Reckase, 2009)

$$P\left(X_{ei}=1\,|\,\underset{\sim}{\theta}_e,\underset{\sim}{a}_i,\underset{\sim}{d}_i,g_i\right)=g_i+\left(1-g_i\right)\left[\prod_{t=1}^{T}\left(\frac{\exp\left(a_{it}\theta_{et}+d_{it}\right)}{1+\exp\left(a_{it}\theta_{et}+d_{it}\right)}\right)\right]. \quad (3)$$

For the NCMIRT model, there is one item difficulty parameter and item discrimination parameter per latent ability (i.e., both subscripted by *t*). Because contributions from individual abilities are multiplied in the product term, the model is denoted as *noncompensatory*.

Product MIRT Model

Chalmers and Flora (2014) describe estimation for NCMIRT models and compare them to "product" MIRT models (pMIRT) with direct products of continuous latent abilities (cf. Eq.13 in Chalmers & Flora (2014)) to capture latent interaction effects for noncompensatory processes. An example they describe with two abilities is (written using the logit of the probability)

$$\text{logit}\left(P\left(X_{ei}=1 \mid \underset{\sim}{\theta}_e, \underset{\sim}{a}_i, d_i\right)\right) = a_{i1}\theta_{e1} + a_{i2}\theta_{e2} + a_{i12}\left(\theta_{e1}\cdot\theta_{e2}\right) + d_i . \quad (4)$$

They remark that this model can have a probability response surface similar to the NCMIRT model described above. The form above resembles a LCDM with two attributes except here the latent variables are continuous abilities instead of categorical attributes. Rupp et al. (2010) call the use of latent interactions to model processes that cannot be assumed compensatory as "effectively noncompensatory." Denoting discrimination parameters using γ instead of a and fixing all pseudo-guessing parameters g_i to zero, the pMIRT model can be generalized to have T continuous abilities by

$$\text{logit}\left(P\left(X_{ei}=1 \mid \underset{\sim}{\theta}_e, \underset{\sim}{\gamma}_i, d_i\right)\right) = \sum_{t=1}^{T}\gamma_{it}\theta_{et} + \sum_{t=1}^{T-1}\sum_{u>t}^{T}\gamma_{itu}\left(\theta_{et}\cdot\theta_{eu}\right) + \ldots + d_i \quad (5)$$

Where the "+…" leaves over to higher-order product terms above just pairwise. However, Chalmers and Flora (2014) just investigated this model with only pairwise terms.

Additionally, we can assume an $I \times T$ item-to-theta design matrix $\left(\underset{\sim}{C}\right)$ that is known *a priori* such that each of its elements $c_{it} = 1$ if the *i*-th item is measured by the *t*-th continuous latent ability and $c_{it} = 0$ otherwise. With these assumptions, Eq. (5) could then be expressed as

$$\text{logit}\left(P\left(X_{ei} = 1 \mid \underset{\sim}{\theta}_e, \underset{\sim}{\gamma}_i, d_i\right)\right) = \sum_{t=1}^{T} \gamma_{it}\theta_{et}c_{it} + \sum_{t=1}^{T-1}\sum_{u>t}^{T} \gamma_{itu}\theta_{et}c_{it}\theta_{eu}c_{iu} + \ldots + d_i \quad (6)$$

Now, when all γ_{itu} and other discrimination parameters for higher-order interactions are fixed to zero, the pMIRT model reduces to a 2PL C-MIRT model. In this way, one might possibly test if any γ_{itu} and above are significantly different from zero, which could provide empirical evidence as to if a compensatory assumption is reasonable (given model assumptions are satisfied). Rupp et al. (2010) critique such statistical based approaches for this decision in the DCM setting, such as using AIC or BIC for improved fit with one assumption over the other because " in practice, the data and theory are seldom at a level of quality needed."

Advantages of the pMIRT model (with up to only pairwise interactions of abilities) discussed by Chalmers and Flora (2014) included greater stability (i.e., higher convergence rates) and faster computing times relative to NCMIRT. The disadvantages include poorer relative fit to a NCMIRT, and challenging parameter estimation especially for standard errors when sample size was 1,000 (less so for N = 2,500+). One suggested reason for the poorer relative fit despite the pMIRT model possibly producing similar

probability response surfaces is that if γ_{itu} becomes too large in magnitude, then even when $\underset{\sim}{\theta}_e$ decreases the response surface may increase (Chalmers & Flora, 2014).

Latent Class Models (LCMs)

The diagnostic classification models discussed in this study are a form of LCM. In LCMs, examinees are assumed to belong to one of *C* discrete classes where $c = 1,\ldots, C$ and where class membership is unknown. That is, unknown membership implies classes are latent and hence models for latent classes specify categorical latent variables. The probability that the *e*-th examinee has a correct response to the *i*-th item, given that they are in the *c*-th latent class can be written as $P\left(X_{ei}=1 \mid c\right)=\pi_{ic}$ (Rost, 1990; Rupp et al., 2010). Dichotomous or polytomous item responses for examinees can be observed for the latent class model. Item responses are assumed to be conditionally independent given membership in the *c*-th latent class (Skrondal & Rabe-Hesketh, 2004). Focusing on dichotomous responses only, Rupp et al. (2010) remark that estimates of π_{ic} can also be interpreted as the item difficulty, which again is class-specific (Rupp et al., 2010). Here, the LCM can be written as (Rupp et al., 2010)

$$P\left(\vec{X}_e=\vec{x}_e\right)=\sum_{c=1}^{C}\upsilon_c\prod_{i=1}^{I}\pi_{ic}^{x_{ei}}\left(1-\pi_{ic}\right)^{1-x_{ei}} \tag{7}$$

Where:

$\vec{x}_e$ is the observed response pattern across all *I* items for the *e*-th examinee

$e = 1,\ldots,E$ examinees

$i = 1,\ldots,I$ items

x_{ei} is the observed scored response for the *e*-th examinee on the *i*-th item

υ_c is the latent class membership probability for the *c*-th class ("mixing proportion")

$c = 1,\ldots,C$ latent classes where class membership is latent and not known *a priori*

$\sum_{c=1}^{C} \upsilon_c = 1$ so there are $c-1$ latent class membership probabilities to be estimated

π_{ic} is the probability of a correct response to the *i*-th item for an examinee given that they belong to latent class *c* (i.e., $\pi_{ic} = P\left(X_{ei} = 1 \mid c\right)$).

For each latent class, the product term above provides the joint probability of a particular response pattern (Rupp et al., 2010). Hancock and Samuelsen (2008) give more details on LCM and extensions thereof.

Diagnostic Classification Models (DCMs)

Rupp and Templin (2008) define DCMs as the following:

> Diagnostic classification models (DCM) are probabilistic, confirmatory multidimensional latent variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables. The predictor variables are combined in compensatory and noncompensatory ways to generate latent classes. DCM enable multiple criterion-referenced interpretations and associated feedback for diagnostic purposes, which is typically provided at a relatively fine-grain size. This feedback can be, but does not have to be, based on a theory of response processing grounded in applied cognitive

psychology. Some DCMs are further able to handle complex sampling designs for items and respondents, as well as heterogeneity due to strategy use.

Henson (2009) argued that DCMs do not enable criterion-referenced interpretations but rather are norm-referenced, and that classification into mastery groups are sample dependent. The same three authors later defined DCMs to be: "Statistical models with discrete latent variables that are used to classify respondents into one of several distinct latent classes associated with distinct attribute profiles" (Rupp et al., 2010). This definition of DCMs is assumed throughout the remainder of this study.

The first task in DCMs can be to estimate the various $P\left(X_{ei}=1\,|\,c\right)=\pi_{ic}$ of the LCM through proposed item response models with categorical latent variables for attributes after imposition of constraints (and therefore are confirmatory in nature). Then, through a chosen structural model, the mixing proportions υ_c are estimated and attribute profile and marginal attribute mastery are computed.

Use of DCMs for Diagnostic Measurement

Rupp et al. (2010) define diagnostic measurement (DM) to be, "The determination of whether respondents have mastered/possess several attributes with the aid of a diagnostic assessment and a suitable latent-variable model." The same authors define a diagnostic assessment to be an assessment that is designed to provide classifications of respondents (Rupp et al., 2010). Diagnosis is defined to be the act of precisely analyzing a problem and identifying its causes for the purpose of classification-based decision making (Rupp et al., 2010). With these three definitions explicated,

DCMs are posited to provide an appropriate mechanism by which to perform DM for skills' mastery diagnosis.

Condensation Rules for DCMs

A condensation rule describes the relationship between observed and latent random variables (Maris, 1995; Maris, 1999). They afford a prescription of how skills are "condensed" (i.e., combined) to produce a latent response for a given DCM (Rupp et al., 2010). Maris (1995) defined three condensation rules still currently considered in DCM: (a) conjunctive, (b), disjunctive, and (c) drop-off. Let skill mastery of the *a*-th skill be denoted as $\alpha_a = 1$ and non-mastery as $\alpha_a = 0$ for $a = 1, 2, \ldots, A$ skills. The current study adopts these definitions for describing how latent skills combine, and according to the following adapted from Rupp et al. (2010) and Zhao (2013):

Table 1. Definition of Model Types and Their Possible Condensation Rules

Model Type	*Condensation Rule*	*Mathematical Form*	*Interpretation*
Compensatory	Disjunctive	$1 - \left[\prod_{a=1}^{A} (1 - \alpha_a) \right]$	At least one of the skills has to be applied (and can compensate for lack of other skills)
Noncompensatory	Conjunctive	$\prod_{a=1}^{A} \alpha_a$	All of the required skills have to be applied

Model Type	*Condensation Rule*	*Mathematical Form*	*Interpretation*
Noncompensatory	Drop-off	$= 0$ if $\alpha_1 = 0$ $= 1$ if $\alpha_1 = 1$ and $\alpha_2 = 0$ ⋮ $= A$ if $\alpha_1 = \alpha_2 = \ldots = \alpha_A = 1$	For polytomous condensation. Process involves A components executed sequentially. Passing component $a+1$ only possible if the a-th component was also passed.

Only conjunctive and disjunctive condensation rules as defined above are considered in the current study since all studied DCMs here have dichotomous condensed evaluands, as discussed in Rupp et al. (2010). Almond and Shute (2009) remark that performance is dominated by the weakest skill in the conjunctive rule and dominated by the strongest skill under the disjunctive rule. They also consider other rules not discussed further in the current study.

An Introduction to Core DCMs

There are multiple studies that give excellent introductions to common DCMs such as Rupp and Templin (2008), Henson et al. (2009), and Rupp et al. (2010). "Core" DCMs (Rupp et al., 2010) studied further are now discussed, and follow similar description as in these works.

Deterministic Input Noisy "And" Gate (DINA)

The DINA model (Macready & Dayton, 1977; Haertel, 1989; Junker & Sijtsma, 2001) is a common model investigated in DCM literature, likely due to its accessible interpretation. The DINA model maps examinee skill sets onto expected item responses

by use of an item-by-skill *Q*-matrix (Tatsuoka, 1985), which is specified *a priori*. The *Q*-matrix is an $I \times A$ matrix of indicators, q_{ia}, of whether the *a*-th skill ($a = 1,\ldots, A$) must be mastered for *i*-th item ($i = 1,\ldots, I$) to be correctly answered by an examinee (Tatsuoka, 1985). Given the *Q*-matrix, the DINA models the probability of a correct response, $X_{ei} = 1$, for the *e*-th examinee ($e = 1,\ldots, E$) to item *i* as

$$P\left(X_{ei} = 1 \,|\, \underset{\sim}{\alpha}_e\right) = \left(1 - s_i\right)^{\eta_{ei}} \left(g_i^{\left(1-\eta_{ei}\right)}\right) \tag{8}$$

The expected item response pattern of the *e*-th examinee is $\underset{\sim}{\eta}_{e\cdot} = \left(\eta_{e1}, \eta_{e2}, ..., \eta_{eI}\right)$, where η_{ei} is a latent variable for the *e*-th examinee possessing all required attributes for answering the *i*-th item. It is the deterministic input of the DINA model, and has been referred to as the *condensation kernel* (Rupp et al., 2010). Here, the *i*-th entry of the expected response pattern for the *e*-th examinee is defined as

$$\eta_{ei} = \prod_{a=1}^{A} \alpha_{ea}^{\,q_{ia}}, \tag{9}$$

Where a skill is represented by α_{ea} and the respective *Q*-matrix entry by q_{ia}. The skill α_{ea} is an entry in an attribute profile vector, $\underset{\sim}{\alpha}_e$, for the *e*-th examinee and is 1 if the *a*-th attribute has been mastered and is 0 if not mastered. Thus, the DINA model is a multidimensional binary latent trait model that requires all attribute skills to be mastered

in order for the latent ideal response pattern entry $\eta_{ei} = 1$. Because all skills must be mastered for $\eta_{ei} = 1$, this model uses a conjunctive condensation rule and is considered noncompensatory (i.e., cannot compensate for lack of one skill by way of another skill).

The DINA model has two item-specific parameters to be estimated which are assumed to be the *same across attributes*: slipping (s_i) and guessing (g_i). In the above, slipping is defined as the probability of an observed incorrect item response when in fact the examinee possesses all the required skills or attributes for the item (i.e., $s_i = P\left(X_{ei} = 0 \mid \eta_{ei} = 1\right)$ where η_{ei} is defined as above). Guessing is defined as the probability of an observed correct item response when in fact the examinee does not possess at least one desired attribute: $g_i = P\left(X_{ei} = 1 \mid \eta_{ei} = 0\right)$.

The unique $\underset{\sim}{\alpha}_e$'s from the DINA model are the latent classes where classification into mastery or non-mastery is desired. There exists 2^A possible $\underset{\sim}{\alpha}_e$. Like other DCMs, it is assumed that for the *e*-th examinee, the item responses $X_{e1}, X_{e2}, ..., X_{eI}$ are conditionally independent given $\underset{\sim}{\alpha}_e$ (de la Torre, 2009; Wang & Douglas, 2013). Finally, there is one s_i and one g_i parameter per item with equality constraints imposed across attributes, and slipping and guessing is subject to the order constraint that $\left(1 - s_i\right) > g_i$ (de la Torre, 2009). In sum, for each item there are only two uniquely determined parameters affecting response probabilities in the DINA model: s_i and g_i.

Compensatory Reparameterized Unified Model (CRUM)

The CRUM (Hartz, 2002; Henson et al., 2009) is a special case of the general diagnostic model (von Davier, 2005), where the logit of probability of correct response is given by

$$\text{logit}\left(P\left(X_{ei}=1\,|\,\underset{\sim}{\alpha}_e\right)\right)=\sum_{a=1}^{A} r_{ia}^{*}\alpha_{ea}q_{ia}-\pi_i^{*}\,, \tag{10}$$

Where all $r_{ia}^{*}>0$. This model is partially compensatory as not possessing other skills ($\alpha_{ea'}=0$ for $a'\neq a$) does not impact the contribution on remaining necessary skills (Henson et al., 2009).

Deterministic Input, Noisy "Or" Gate (DINO)

The DINO model (Templin & Henson, 2006) is the analogous compensatory version of the DINA model (Rupp et al., 2010). The following disjunctive condensation kernel is used:

$$\omega_{ei}=1-\prod_{a=1}^{A}\left[\left(1-\alpha_{ea}\right)^{q_{ia}}\right] \tag{11}$$

This disjunctive condensation kernel implies that possessing at least one attribute can completely compensate for the lack of all others (Rupp et al., 2010). Given this, the DINO probability of correct response is defined as

$$P\left(X_{ei}=1\,|\,\omega_{ei}\right)=\left(1-s_i\right)^{\omega_{ei}}\,g_i^{\left(1-\omega_{ei}\right)}\,, \tag{12}$$

Where slipping and guessing are now defined as:

$$s_i = P\left(X_{ei} = 0 \mid \omega_{ei} = 1\right) \tag{13}$$

And

$$g_i = P\left(X_{ei} = 1 \mid \omega_{ei} = 0\right). \tag{14}$$

Like the DINA, the DINO model constrains slipping and guessing to be the same across attributes. Again for each item there are only two uniquely determined parameters affecting response probabilities: s_i and g_i.

The Log-Linear Cognitive Diagnosis Model (LCDM)

The LCDM of Henson et al. (2009) defines a family of DCMs using a loglinear with latent variable modeling specification (Haberman, 1974; Haberman, 1979; Hagenaars, 1993). The functional form of the LCDM is given as

$$P\left(X_{ei} = 1 \mid \underset{\sim}{\alpha}_e, \underset{\sim}{\lambda}_i, \eta_i\right) = \frac{\exp\left[\underset{\sim}{\lambda}_i^{\mathrm{T}} \underset{\sim}{\mathrm{h}}\left(\underset{\sim}{q}_i, \underset{\sim}{\alpha}_e\right) + \lambda_{i0}\right]}{1 + \exp\left[\underset{\sim}{\lambda}_i^{\mathrm{T}} \underset{\sim}{\mathrm{h}}\left(\underset{\sim}{q}_i, \underset{\sim}{\alpha}_e\right) + \lambda_{i0}\right]}, \tag{15}$$

Where

$$\underset{\sim}{\lambda}_i^{\mathrm{T}} \underset{\sim}{\mathrm{h}}\left(\underset{\sim}{q}_i, \underset{\sim}{\alpha}_e\right) = \sum_{a=1}^{A} \lambda_{ia} \alpha_{ea} q_{ia} + \sum_{a=1}^{A-1} \sum_{b>a}^{A} \lambda_{iab} \alpha_{ea} q_{ia} \alpha_{eb} q_{ib} + \ldots \tag{16}$$

And where model terms are defined by:

X_{ei} is the scored item response for the *e*-th examinee on the *i*-th item, where $X_{ei} = 1$ for a correct response and $X_{ei} = 0$ for an incorrect response

$e = 1, 2, ..., E$ for the *E* examinees

$i = 1, 2, ..., I$ for the *I* items

$\underset{\sim}{\alpha}_e$ is a vector of length *A* of categorical latent attributes for the *e*-th examinee

$\underset{\sim}{\lambda}_i^{\mathrm{T}} \underset{\sim}{h}\left(\underset{\sim}{q}_i, \underset{\sim}{\alpha}_e\right)$ is the *kernel* for the LCDM and is defined by the above so that:

λ_{i0} is the logit of correct response for the *i*-th item when no attributes have been mastered (intercept term). It should be noted that $\lambda_{i0} = -\eta_i$ as some prior research uses $-\eta_i$.

α_{ea} is the categorical latent attribute for the *e*-th examinee for mastery of the *a*-th attribute, so that $\alpha_{ea} = 1$ if the *e*-th examinee has mastered attribute *a* and $\alpha_{ea} = 0$ otherwise. Here, *a* = 1, 2, …, *A* for the *A* attributes

λ_{ia} is the weight (logit increment) for item *i* on the *a*-th categorical latent attribute α_{ea}

q_{ia} is an indicator variable for whether the *i*-th item requires the *a*-th categorical latent attribute, so that $q_{ia} = 1$ if the *i*-th item requires the *a*-th attribute and $q_{ia} = 0$ otherwise.

λ_{iab} is the weight (logit increment) for the pairwise interaction of the *a*-th categorical latent attribute α_{ea} and the *b*-th categorical latent attribute α_{eb} ($b > a$) for the *i*-th item

"+…" represents higher-order interaction terms above pairwise (up to the *A*-th way interaction).

Using Eq. (24) with up to third-order interactions, Eq. (23) can equivalently be represented by way of the logit of correct response, $\text{logit}\left(P\left(X_{ei}=1\,|\,\underset{\sim}{\alpha}_e,\underset{\sim}{\lambda}_i,\lambda_{i0}\right)\right)$, given by

$$\sum_{a=1}^{A}\lambda_{ia}\alpha_{ea}q_{ia}+\sum_{a=1}^{A-1}\sum_{b>a}^{A}\lambda_{iab}\alpha_{ea}q_{ia}\alpha_{eb}q_{ib}+\lambda_{i123}\alpha_{e1}q_{i1}\alpha_{e2}q_{i2}\alpha_{e3}q_{i3}+\lambda_{i0} \qquad (17)$$

This representation of the LCDM is referred to as the saturated LCDM in the current study.

LCDM Representations of Core DCMs

Rupp et al. (2010) remark that "Constraints can be placed on the parameters in $\underset{\sim}{\lambda}_i$ so that the probability of a correct response increases in accordance with the DCM that is represented by the LCDM." Thus, the core DCMs previously discussed can be represented by the LCDM framework. LCDM representations are described next for these core DCMs: the DINA, CRUM, and DINO.

DINA Representation in the LCDM

The DINA model can be represented as an LCDM with all main effects set to zero and positive interaction terms. The functional form of the DINA LCDM for three skills is given by

$$\text{logit}\left(P\left(X_{ei}=1\,|\,\underset{\sim}{\alpha}_e,\underset{\sim}{\lambda}_i,\lambda_{i0}\right)\right)=\lambda_{i123}\alpha_{e1}q_{i1}\alpha_{e2}q_{i2}\alpha_{e3}q_{i3}+\lambda_{i0} \qquad (18)$$

The relationships to the original DINA model item parameters (for two attributes case) are described in Henson et al. (2009) and are given by

$$\lambda_{i0} = \ln\left(\frac{g_i}{1-g_i}\right) \tag{19}$$

And

$$\lambda_{iab} = -\ln\left(\frac{g_i}{1-g_i}\right) + \ln\left(\frac{1-s_i}{s_i}\right) \tag{20}$$

CRUM Representation in the LCDM

The CRUM can be represented as an LCDM with all interactions set to zero and positive main effects. Therefore, the functional form of this restricted LCDM is given by

$$\text{logit}\left(P\left(X_{ei}=1 \mid \underset{\sim}{\alpha}_e, \underset{\sim}{\lambda}_i, \lambda_{i0}\right)\right) = \sum_{a=1}^{A} \lambda_{ia}\alpha_{ea}q_{ia} + \lambda_{i0}\,. \tag{21}$$

DINO Representation in the LCDM

With up to three-way interaction effects, the DINO model can be represented as an LCDM with all main effects and interactions set with equality constraints, and further where main effects are positive, two-way interactions are negative, and the three-way interaction is positive. Given these constraints, the functional form of the DINO LCDM

parameterization for the $\text{logit}\left(P\left(X_{ei}=1\,|\,\underset{\sim}{\alpha}_e,\lambda_i,\lambda_{i0}\right)\right)$ is given by (putting the negative sign outside of double-sum)

$$\sum_{a=1}^{A}\lambda_i\alpha_{ea}q_{ia}-\sum_{a=1}^{A-1}\sum_{b>a}^{A}\lambda_i\alpha_{ea}q_{ia}\alpha_{eb}q_{ib}+\lambda_i\alpha_{e1}q_{i1}\alpha_{e2}q_{i2}\alpha_{e3}q_{i3}+\lambda_{i0}\,. \quad (22)$$

When higher than three-way interactions are considered (but not studied here), the sign of the λ_i for these higher-order interaction effects alternates as described in Henson et al. (2009).

Relationships of Core DCM Parameterizations to LCDM Representations

The LCDM representations of the three core DCMs considered in this study have been explicated, but the relationships to each models original parameterization are import to further consider. Knowing these, previously reported estimates from the original models could be further incorporated into the current study. The following table describes the relationship of item parameters in the original core DCMs relative to their LCDM representations for the case where two attributes are required by items (complexity = 2):

Table 2. Conversion of Core DCM Parameters to Equivalent LCDM Representation Parameter*

Model	*Original parameter*	*Original Parameter to LCDM representation*	*LCDM representation to Original Parameter*
DINA	s_i	$s_i = \frac{\exp(-\lambda_{i0} - \lambda_{iab})}{1+\exp(-\lambda_{i0} - \lambda_{iab})}$	$\lambda_{iab} = -\ln\left(\frac{g_i}{1-g_i}\right) + \ln\left(\frac{1-s_i}{s_i}\right)$
	g_i	$g_i = \frac{\exp(\lambda_{i0})}{1+\exp(\lambda_{i0})}$	$\lambda_{i0} = \ln\left(\frac{g_i}{1-g_i}\right)$
CRUM	r_{ia}^*	$r_{ia}^* = \lambda_{ia}$	$\lambda_{ia} = r_{ia}^*$
	π_i^*	$\pi_i^* = -\lambda_{i0}$	$\lambda_{i0} = -\pi_i^*$
DINO	s_i	$s_i = \frac{\exp(-\lambda_{i0} - \lambda_i)}{1+\exp(-\lambda_{i0} - \lambda_i)}$	$\lambda_i = -\ln\left(\frac{g_i}{1-g_i}\right) + \ln\left(\frac{1-s_i}{s_i}\right)$
	g_i	$g_i = \frac{\exp(\lambda_{i0})}{1+\exp(\lambda_{i0})}$	$\lambda_{i0} = \ln\left(\frac{g_i}{1-g_i}\right)$

**Note.* Derivations made under assumption items require two attributes (complexity = 2).

Impact of Simple versus Complex Structure on LCDM Representations

Under *complex structure*, defined to be any item requiring more than one attribute, the above formulae describe the LCDM representations of the three core DCMs studied. Under *simple structure*, defined to be all items each requiring only one attribute, all of the LCDM representations reduce to be the same model given by

$$\text{logit}\left(P\left(X_{ei}=1 \mid \underset{\sim}{\alpha}_e, \underset{\sim}{\lambda}_i, \lambda_{i0}\right)\right) = \sum_{a=1}^{A} \lambda_{ia}\alpha_{ea}q_{ia} + \lambda_{i0}\,. \qquad (23)$$

Thus, only this one model representation has to be considered in this case.

Applied DCM Studies in the Literature

There have been many technical studies on DCMs reported that have focused on statistical and psychometric aspects usually performed with simulation studies. It is promising that recent literature has seen a substantial growth in DCM application studies, which were reviewed to inform the current study. The following table provides details of recent studies, their scope and DCM(s) utilized, sample, and range of item parameter estimates reported.

Table 3. Applied DCM Studies from Previous Literature and Range of Item Parameter Estimates

Author(s)	**Sample**	**No. Items / No. Attributes**	**DCM(s) used**	**Range of estimates**
Bradshaw et al. (2014)	990 5th-7th grade teachers (for fraction arithmetic)	27 items (MC; few CR) / 4 attributes (19 simple items, 8 complexity=2)	LCDM (saturated)	η_i : mean= -1.38 λ_{ia} : mean= 1.40 to 3.23 λ_{iab} : mean= 1.41
de la Torre & Douglas (2004)*	Fraction subtraction for 2,144 students (Tatsuoka, 2002)	20 items / 8 attributes (up to complexity=5)	Higher-order DINA	s_i = 0.04 to 0.33; g_i = 0.00 to 0.44
Henson & Templin (2007), Feng et al. (2014)	ESL using ECPE assessment on 2,922 examinees	28 items / 3 attributes (average complexity = 1.32 items per attribute)	Reduced RUM in Feng et al. (2014)	π^*: 0.70 to 0.97 r^*: 0.36 to 0.92 across skills
Henson, Templin, Willse, & Irwin (2015)	2318 students on Reading EOGs	73 items / 2 attributes	Categorical Bi-factor (constrained MCCIRM)	Not presented

Author(s)	Sample	No. Items / No. Attributes	DCM(s) used	Range of estimates
Jang (2009a)	2,703 test takers of L2 reading comprehension TOEFL internet-based test	37 items / 9 final skills (range from 4 items per skill to 12)	Fusion model (maximum complexity = 3)	π^*: 0.56 to 0.99 r^*: 0.34 to 0.84 across skills
Jurich & Bradshaw (2014)	Higher Ed: Proficiency in evaluating psychosocial research; 1,710 students at 2 time points	17 items / 4 attributes (simple structure)	LCDM	η_i : -1.248 to 0.938 (mean = -0.310) λ_{ia} : 0.540 to 3.341 (average = 0.859 to 1.602)
Kim (2011)	480 essays for English for Academic purpose (TOEFL iBT) with 10 teachers	35 descriptors (with 2 prompts) / 5 skills	Reduced RUM (complexity = 3 or less except for 1 item loading on all 5 skills)	π^*: 0.54 to 0.99 r^*: 0.05 to 0.88 across skills
Kunina-Habenicht et al. (2009)	464 German 3rd and 4th graders in 10 classes in 6 schools on math	8 to 27 per skill / 5 skills	2-parameter GDM	GDM loadings range from 0.51 to 0.99
Lee & Sawaki (2009)	Listening & Reading from TOEFL iBT for 3,139 ESL students	Listening: 34 / 4 Reading: 39 / 4	PC-GDM, Fusion model, cLCA	Not presented
Li & Suen (2013)	2,019 students for Reading comprehension of ELA in Michigan	20 reading items / 5 skills (2 items complexity=3, rest split between simple/complex)	Fusion model	π: 0.619 to 0.995 r^*: 0.237 to 0.958 across skills

Author(s)	Sample	No. Items / No. Attributes	DCM(s) used	Range of estimates
Templin & Henson (2006)	593 potential college underclassmen pathological gamblers	41 items / 10 attributes (criteria) (average of 5.5 items per skill)	DINO (average complexity = 1.34)	s_i : 0.08 to 0.66 g_i : 0.01 to 0.56 (cf. Table 4)
Zhao (2013)	3 sets: (1) Fraction subtraction, (2) Math in TIMSS 2007 1,131 4th graders, (3) FCSA 2011 in 1,629 middle & high school US students	(1) 20 / 8 (2) 23 / 15 (3) 20 / 5	DINA, Fusion model	(1) s_i : 0.03 to 0.33, g_i : 0.01 to 0.48, π*: 0.72 to 0.99, (mean = 0.93) *r**: 0.00 to 0.95 across skills (mean: 0.01 to 0.40) (2) s_i : 0.03 to 0.90, g_i : 0.00 to 0.76, π*: 0.11 to 0.99, (mean = 0.80) *r**: 0.00 to 0.97 (mean: 0.02 to 0.44) across skills (3) s_i : 0.01 to 0.63, g_i : 0.24 to 0.81, π*: 0.41 to 0.99, *r**: 0.00 to 0.95 across skills

**Note*. Primarily simulation study.

Many of the technical studies that were primarily simulation based also typically reported an application to real assessment data, even if retrofitted after the fact to an assessment engineered to be unidimensional and for the purpose of rank-ordering examinees along an ability continuum. These studies are not enumerated in the above table, except for de la Torre and Douglas (2004) which reported one of the first DCM applications on a heavily cited data source: the fraction subtraction data of Tatsuoka (2002). The remainder of the studies reported on DCM applications in prospective studies for skill mastery diagnosis. Their range of estimates provides real-word linkages based on relationships to their corresponding LCDM representations for describing logit separation between mastery groups.

Assessing Fit of DCMs

One of the most salient issues facing current research efforts using DCMs has been fit assessment, including item fit, person fit and globally for model fit. Many studies have focused on validation of the specified *Q*-matrix, and how certain misspecifications impact estimation and ultimately classification. A novel study by Madison and Bradshaw (2014) looked at impact of different *designs of correctly specified Q*-matrices on classification and made recommendations. The following table describes studies in the current literature that have examined DCM fit assessment from various perspectives:

Table 4. DCM Fit Studies

Author(s)	**Item, Person, or Model fit examined?**	**DCM(s) used**	**Measures of Fit**	**Relevant Conclusions**
Chen, de la Torre, & Zhang (2013)	Model	DINA, DINO, R-RUM, G-DINA (equivalent to earlier presented LCDM)	LL, AIC, BIC, residuals based on p-values, log-OR item pairs or correlations	AIC and BIC picked true or saturated models; correlation and log-OR had comparable performance
de la Torre & Douglas (2004)	Item and Model	DINA, higher-order DINA	pairwise log-ORs and Bayes factors	Supported higher-order DINA where appropriate
de la Torre & Lee (2013)	Item	DINA, DINO	Wald test	"Excellent" power when sample size is sufficiently high for these DCMs.
Henson et al. (2015)	Item and Model	LCDM, Categorical Bifactor (with simple structure)	AIC, BIC, residuals, Yen's $Q3$, eigenanalysis permuted Q-matrix	The chosen fit measures adequately convey improved model fit where it conceptually should.
Kunina-Habenicht et al. (2012)	Item and Model	LCDM	Newly proposed MAD (cf. Eq. (3)) RMSEA (cf. Eq. (4))	More sensitive to overspecification of Q-matrix than underspecification. AIC, BIC sensitive to both. Excluding true 3-way interactions did not affect classification.

Author(s)	Item, Person, or Model fit examined?	DCM(s) used	Measures of Fit	Relevant Conclusions
Jurich (2014)	Model	LCDM	Limited information (LI) M_2, RMSEA	Full information fit is problematic for realistic test lengths. Both M_2 and RMSEA performed well for fit in LCDM (unlike for MIRT)
Madison & Bradshaw (2014)	Model	LCDM	Only examine convergence rates, classification performance, and reliability	At least 1 factorially simple item per skill is required for identification. Not all *Q*-matrix designs are equal in quality.
Rupp et al. (2010)	Item, Person, and Model	All core DCMs and the LCDM	AIC, BIC, Resampling, Posterior predictive model checking (PPMC), LI statistics, Person fit statistics.	As of 2010, assessing DCM fit is an evolving field. If the DCM does not fit or the *Q*-matrix is misspecified, then LI can be violated.
Sinharay & Almond (2007)	Item and Model	2-Parameter Latent Class (2LC) Model	Bayesian residuals, PPMC	PPMC can be successively used to evaluate fit (but may be slow)
Templin & Henson (2006)	Item and Model	Higher-order DINO	Monte Carlo, RMSEA of Pearson *r*, Cohen's κ (cf. their Eq. (5))	Practical methods for model fit in DCM remains an open research question.

Author(s)	Item, Person, or Model fit examined?	DCM(s) used	Measures of Fit	Relevant Conclusions
Zhao (2013)	Model	DINA, fusion model	Posterior predictive model checking using NC as measure of discrepancy	Choice of discrepancy measure for PPMC is important; Person-fit measures are not widely used.

It is evident that there is wide heterogeneity in approaches to assessing fit for internal structure validity in DCMs, so that the more recently proposed approaches could be recommended. Commonalities among approaches are that: (a) global model fit is often assessed (although in alternative ways), (b) calls remain for increased research into model fit methods, and (c) methods for assessing violation of local independence are lacking (exceptions are Hansen, 2013 and Henson et al., 2015).

Implementation of DCMs

There are multiple software platforms that facilitate estimation of DCMs. Templin and Hoffman (2013) describe how to estimate LCDMs using a SAS® macro (which is freely available) to call M*plus* (Muthén & Muthén, 1998-2012) software and store results in created SAS® datasets. Shu, Henson, and Willse (2013) used the 'LCDM.exe' software package of Henson (2008) to estimate LCDMs. Zhao (2013) used OpenBUGS to estimate the DINA model using MCMC and provided code. The flexMIRT® software (Houts & Cai, 2013) uses a general modeling framework described as

$$\text{logit}\left(P\left(y_i = 1 \mid \underset{\sim}{x}, \underset{\sim}{\xi}\right)\right) = \alpha + \sum_{a_1=1}^{A} \beta_{a_1} x_{a_1} + \sum_{a_1=1}^{A} \sum_{a_2=1}^{(a_1-1)} \beta_{a_1,a_2} x_{a_1} x_{a_2} + \ldots + \sum_{a=1}^{p} \lambda_a \xi_a \quad (24)$$

Where x's are latent 0/1 attributes, ξ's are latent continuous variables, and +…+ indicates higher-order interaction terms. Examples of implemented DCMs that come with flexMIRT include: C-RUM, DINA, DINO, and a testlet DINA model. Fitting DCMs via the GDM is available using the *mdltm* software (von Davier, 2005; von Davier & Xu, 2009). Finally, the R package 'CDM' Robitzsch et al. (2013) is freely available, which can estimate the DINA, DINO, GDINA, polytomous GDINA, and GDM models. The 'NPCD' package (Zheng, Chiu, & Douglas, 2013) in R can be used for nonparametric distance-based classification as described in Chiu and Douglas (2013).

This study uses Robitzsch et al. (2013) CDM package in R for estimation. Robitzsch et al. (2013) remark that estimation using the expectation-maximization (E-M) algorithm is performed based on de la Torre (2011).

The Assumption of Local Independence

Equation (7) gave the joint probability of the observed response pattern in the LCM. This equation only holds when the assumption of local independence (LI) is tenable. The direct implication of this assumption allows one to estimate this joint probability using the product term of all π_{ic}, which is the probability of correct response to the *i*-th item for an examinee member of the *c*-th latent class. If this assumption is violated, then this *measurement* portion of the latent class model breaks down. Without this, we could not then sum across the mixing proportion estimates to arrive at the joint

probabilities of item response strings because the product term would not be accurate. Because all DCMs discussed in this study have been previously defined as a form of LCM, violating LI could invalidate the resulting diagnostic classification. Forms of independence including LI are discussed in more detail next.

Independence of Probabilistic Events

A countable number of *n* events (e.g., an item response on a test), $E_1, E_2, \ldots, E_n$, are defined to be mutually independent (MI) if and only if the probability of their joint occurrence (i.e., joint probability) is given by the product of the probabilities of each of the events given by

$$P\left(\bigcap_{i=1}^{n} E_i\right) = P\left(E_1, E_2, \ldots, E_n\right) = \prod_{i=1}^{n} P\left(E_i\right) \tag{25}$$

This is known as the *multiplication rule* for mutually independent events. An event E_i could be a test response. A less restrictive form of independence is pairwise independence (PI). Here, a countable number of events are pairwise independent if and only if every pair of events, *j* and *k*, are independent as described by

$$P\left(E_j \bigcap E_k\right) = P\left(E_j, E_k\right) = P\left(E_j\right) \cdot P\left(E_k\right) \tag{26}$$

It should be explicitly stated that MI implies PI is satisfied.

Conditional Independence

Dawid (1979; 1980) gives treatment of conditional independence. The definition of pairwise conditional independence is given as

$$P\left(E_j \bigcap E_k \mid E_i\right) = P\left(E_j \mid E_i, E_k \mid E_i\right) = P\left(E_j \mid E_i\right) \cdot P\left(E_k \mid E_i\right), \qquad (27)$$

For all $i \neq j \neq k$. In the context of DCM this can be represented by

$$P\left(Y_{ei} = 1 \bigcap Y_{ei'} = 1 \mid \underset{\sim}{\alpha}\right) = P\left(Y_{ei} = 1 \mid \underset{\sim}{\alpha}, Y_{ei'} = 1 \mid \underset{\sim}{\alpha}\right) = P\left(Y_{ei} = 1 \mid \underset{\sim}{\alpha}\right) \cdot P\left(Y_{ei'} = 1 \mid \underset{\sim}{\alpha}\right) \qquad (28)$$

For all items $i \neq i'$ and where $Y_{ei} = 1$ is a correct response to the *i*-th item for the *e*-th examinee and $\underset{\sim}{\alpha}$ is the vector of dichotomous latent skills.

Local Independence in IRT

In IRT literature there have been different forms of LI discussed. The "Strong" form of local independence is given by (Embretson & Reise, 2000)

$$P\left(\underset{\sim}{X} = \underset{\sim}{x} \mid \theta\right) = \prod_{i=1}^{I} P_i\left(X_i = x \mid \theta\right), \qquad (29)$$

Where $\underset{\sim}{x}$ is the examinee's vector of responses and x is their response to the *i*-th item. Here, knowing that a person gets an item correct does not change the conditional probability of getting any other item correct. There is also a "Weak" form of local independence, given by (Embretson & Reise, 2000)

$$\mathrm{Cov}\left(X_i, X_{i'} \mid \theta\right) = 0\,, \tag{30}$$

For all items $i \neq i'$. This indicates that pairs of items share no covariance once the latent trait(s) have been accounted for. This is a weaker form of LI because higher order dependencies among items are allowed here but not in the Strong form of local independence above.

The LI assumption is the foundation for model-fitting algorithms that can provide goodness-of-fit indices and terms for residual covariance. Specifically, IRT assumes the Strong form of local independence since it is "full-information" as it models the entire response string. Factor analysis of categorical variables assumes the Weak form of local independence since it is "limited-information" as it models the item covariances (and possibly means) which are summary statistics (rather than item response strings directly from raw data).

Stout (1987) and Stout (1990) suggested that a state of "essentially unidimensional" is enough to satisfy IRT assumptions. Here, a test is considered essentially unidimensional when the average between-item residual covariances after fitting a one-factor limited-information model approaches zero as the length of the test increases. Therefore, this notion of essentially unidimensional is under a Weak form of LI. Consequently, a test can be considered essentially unidimensional if LI approximately holds in a sample of test takers who are approximately equal on the latent trait. Stated another way, just because a test is essentially unidimensional does not mean the Strong form of LI has been satisfied.

To understand what impact violations of local independence can have on IRT, one can inspect the likelihood equation used to estimate IRT model parameters. In general, the likelihood is given by the joint probability of observing each scored item response

$$L\left(\text{data} \mid \text{parameter value}\right) = L\left(X_1 = x_1, X_2 = x_2, \ldots, X_I = x_I \mid \theta_e\right) \quad (31)$$

This joint probability can only be broken down into the multiplication of the conditional probabilities of each item response given the value of the latent trait if the Strong form of LI using Eq. (29) can be assumed, and is described by

$$L\left(X \mid \theta_e\right) = \prod_{i=1}^{I} \left(P_{ei} \mid \theta_e\right)^{x_i} \left(1 - P_{ei} \mid \theta_e\right)^{(1-x_i)} \quad (32)$$

Local Independence in DCMs

Lazarsfeld and Henry (1968) define local independence for LCMs to mean that within a class items are all independent of one another. This implies mutual independence of the strong form and not just item pairwise independence as in the Weak form (Lazarsfeld & Henry, 1968). However, the following quote underlines that no latent variable model would likely perfectly account for the associations among manifest variables in practice, "In nearly every social science application, we are quite sure, *a priori*, that no manifest observation will be completely determined by the assumed position of a respondent on an underlying latent scale" (Lazarsfeld & Henry, 1968).

While it is hard to disagree with this, the practical question is: What level of systematic variation causes a violation of the latent model assumption that items are independent after accounting for class? Henning (1989) provides a nontechnical treatment of LI regarding. Lazarsfeld and Henry (1968) went into depth about this assumption in their treatment of latent structure models:

> A series of accounting equations links the latent parameters and the empirically observed response frequencies. These equations derive from the central substantive idea of the whole system, called the *principle of local independence*, following a suggestion of Mosteller.

They describe local independence starting with the following set of equations

$$P\left(I = i, J = j\right) = P\left(I = i\right) \cdot P\left(J = j\right) \tag{33}$$

Which is subsequently represented symbolically as $p_{ij} = p_i p_j$. However, their description *also* included other non-pairwise relationships between the joint and marginal probabilities:

$$\begin{aligned} p_{ij} &= p_i p_j \\ p_{ijk} &= p_i p_j p_k \\ &\vdots \\ p_{1,2,\ldots,J} &= p_1 p_2 \cdot \ldots \cdot p_J \ . \end{aligned} \tag{34}$$

Thus, Lazarsfeld and Henry (1968) define LI as the Strong form for latent structure models (i.e., conditional mutual independence). For DCMs in particular, the joint

distribution of item responses given the parameters in the latent space (the skill profile vectors $\underset{\sim}{\alpha}$ and model parameters $\underset{\sim}{\Lambda}$) under the conditional LI assumption can be written as:

$$P\left(Y_{e1} = y_{e1},\ Y_{e2} = y_{e2},\ldots,Y_{eI} = y_{eI} \mid \underset{\sim}{\alpha},\ \underset{\sim}{\Lambda}\right) = \prod_{i=1}^{I} P\left(Y_{ei} = y_{ei} \mid \underset{\sim}{\alpha},\ \underset{\sim}{\Lambda}\right) \quad (35)$$

There are also other assumptions in DCM, such as the probability of correct response is non-decreasing for attribute mastery (or as more attributes are mastered). That is: $P\left(X_{\square i} = 1 \mid \underset{\sim}{\alpha} = \hat{\alpha}\right)$ is non-decreasing in $\hat{\alpha}$ for items i = 1, 2, …, I. This assumption is known as *monotonicity*. Latent trait models such as IRT also assume unidimensionality of latent ability (obviously models assuming such exclude MIRT and other extensions). DCMs are multidimensional in their latent parameters by the previous definition.

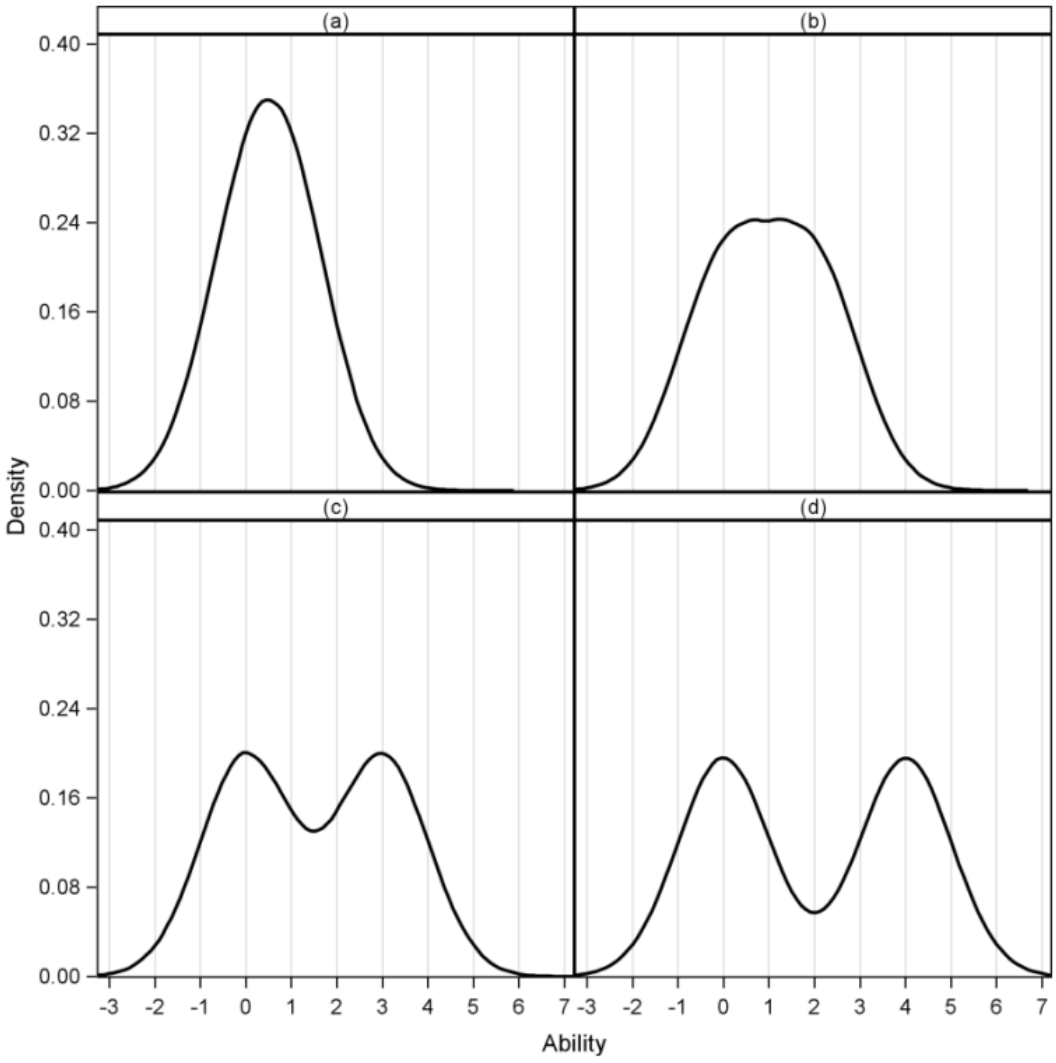
When LI is violated this can result in biased item parameter estimates and therefore incorrect estimates of the probability of mastery and ultimately classification. Rosenbaum (1998) contends that the degree of violation may be more intense depending upon its source. Thus, it is useful to consider various conceptual sources of within-class variation that could cause a violation in LI (cf. *Conceptual sources of systematic within-skill profile variation* below).

Mixture Models

Mixture models are a general approach for modeling data that is assumed to stem from different groups (or clusters) but where group membership is unknown (Frick et al.,

2012). To illustrate, the following Figure 1.2 from McLachlan and Peel (2000) present density plots of two *mixtures* of Normal distributions for a few choice effect sizes:

Figure 1. Plot of Mixture Density of Two Univariate Normal Components in Equal Proportions with Common Variance



Note: $\sigma^2 = 1$ and means $\mu_A = 0$ and $\mu_B = \Delta$ in the cases: (a) $\Delta = 1$; (b) $\Delta = 2$; (c) $\Delta = 3$; (d) $\Delta = 4$ (Figure 1.2 from McLachlan & Peel (2002); reproduced with permission; Copyright 2000 with Wiley).

Similar phenomena are investigated for the DCM setting in this study.

Characteristics of Mixture Distributions

An exemplar of two mixtures of Normal distributions can be considered for describing distributional characteristics. Standard moments characterizing distributions such as mean, standard deviation, skewness, and kurtosis could possibly be used, but there are other characteristics that are relevant for mixtures in particular. One method of visualizing overlapping distributions was described by Linacre (1996), who plotted the ratio of two group's standard deviations (with larger standard deviation in numerator) on the *y*-axis and $\left|\mu_A - \mu_B\right| / \sigma_{smaller}$ on the *x*-axis according to contours of the expected percentage of Normal distribution overlap. Closely related is the well-known Cohen's *d* measure, which is the ratio of the difference in locations of two mixtures to a pooled standard deviation, given by

$$d = \frac{\mu_A - \mu_B}{\sigma_{\text{pooled}}}. \tag{36}$$

Given this measure, Cohen (1977) defined U_3 as a measure of non-overlap, where the percentage of the population distribution for A for which the upper half of the cases of the population B distribution exceeds. Cohen's *d* can be converted to Cohen's U_3 by: $U_3 = \Phi(d)$, where Φ is the cumulative distribution function of the standard Normal distribution. Cohen's *d* can also be converted to an "overlapping" coefficient (*OVL*) using the following formula (Reiser & Faraggi, 1999)

$$OVL = 2\left[\Phi\left(-\frac{1}{2}\cdot|d|\right)\right], \tag{37}$$

Again where Φ is the cumulative distribution function of the standard Normal distribution. There is one additional metric of overlap amount considered here: the Bhattacharyya coefficient (*BC*; Djouadi, Snorrason, & Garber, 1990), which is defined as

$$BC = \exp\left[-\left(\frac{\left(\mu_A - \mu_B\right)^T\left(\mu_A - \mu_B\right)}{8\sigma^2}\right)\right]. \tag{38}$$

The following table describes various combinations of Cohen's *d* and measures of overlap when $\mu_A = 0$ and the σ_{pooled} is held consistent at 0.5 across scenarios:

Table 5. Relationship between Cohen's *d*, U_3, *OVL*, and *BC* Holding σ_{pooled} Constant

μ_A	μ_B	σ_{pooled}	Cohen's *d*	Cohen's U_3	*OVL*	*BC*
0	0	0.5	0	50%	100%	1.000
0	0.1	0.5	0.2	57.93%	92.03%	0.995
0	0.25	0.5	0.5	69.15%	80.26%	0.969
0	0.4	0.5	0.8	78.81%	68.92%	0.923
0	0.5	0.5	1	84.13%	61.71%	0.883
0	1	0.5	2	97.72%	31.73%	0.607
0	1.5	0.5	3	99.87%	13.36%	0.324
0	2.5	0.5	5	100%	1.24%	0.044
0	4.1	0.5	8.2	100%	0%	<0.001

Here is a similar table when $\mu_A = 0$ and $\mu_B = 1.5$, and varying σ_{pooled} across scenarios:

Table 6. Relationship between Cohen's *d*, U_3, *OVL*, and *BC* with varying σ_{pooled}

μ_A	μ_B	σ_{pooled}	Cohen's *d*	Cohen's U_3	*OVL*	*BC*
0	1.5	$\rightarrow+\infty$	0	50%	100%	$\rightarrow$1
0	1.5	7.5	0.2	57.93%	92.03%	0.995
0	1.5	3.0	0.5	69.15%	80.26%	0.969
0	1.5	1.875	0.8	78.81%	68.92%	0.923
0	1.5	1.5	1	84.13%	61.71%	0.883
0	1.5	0.75	2	97.72%	31.73%	0.607
0	1.5	0.5	3	99.87%	13.36%	0.324
0	1.5	0.3	5	100%	1.24%	0.044
0	1.5	0.183	8.2	100%	0%	<0.001

Thus, as the Cohen's *d* decreases to zero, the common variance approaches positive infinity. Likewise, when the numerator of Cohen's *d* is held constant but the value of *d* increases *ad infinitum*, then it is because the common variance has decreased. This study will investigate this relationship for DCMs when there is separation on latent traits between skill masters and nonmasters but within each there remains variance grafted on from continuous abilities due to various conceptual causes. Another perspective is that there is a mixture IRT model with nonzero ability variance where mixtures are separated by mastery states of skills.

Mixture IRT Models

Mixture IRT models are IRT models allowing for unobserved latent subpopulations using mixtures (Rost, 1990; Willse, 2011). This is compelling when the IRT model does not fit overall but is well-fitting within each mixture (Willse, 2011). One application of mixture IRT models has been to investigate possible DIF effects for latent groups (e.g., Cohen & Bolt, 2005; DeMars & Lau, 2011). If person ability

parameters are fixed to a constant, then mixture IRT reduces to a LCM (von Davier & Rost, 2006).

Conceptual Sources of Systematic Within-skill Profile Variation

Without variation within latent skill mastery class, an examinee would only be considered a skill master or nonmaster, and among nonmasters everyone is equally a nonmaster (beyond variation in error prone observed items). Assumed equivalence of this nature could imply that a single treatment (educational remediation) for nonmastery at one "dose" level is sufficient. This conclusion of uniform remediation would be challenging to defend, as many bridges of inferences would have to be supported for the validity of this claim. Thus, some non-systematic, stochastic within-class variation is presumed in the diagnostic measurement process (e.g., random variation from slipping or guessing).

In DCM, latent classes are the profiles of skill mastery; not just mastery of each skill. Therefore, an examinee is classified as having a particular profile (or not). As the number of skills grows large, perhaps it is plausible to conjecture that everyone in the particular skill profile pattern of mastery/non-mastery is on reasonably equal ability potentially. Does the reasonableness of this claim grow as the number of skills grows so large that skill profiles are highly specific and homogenous? Potentially, for example, <0,1,0,1> and <0,1,0,1,0,0,1,0,1,1,0,1> are two vectors of skills of length 4 and 12, respectively. Perhaps it could be argued that placing examinees in the more "specialized" profile of 12 skills could suggest these examinees are more homogeneous with respect to these attributes because if there is enough variation in mastery to differentiate them they

would have been placed in another profile. With only four skills, the classification could perhaps be considered so coarse (or alternatively the definitional grain-size of the attributes) that there is more of an opportunity for within-profile variation than in a scenario with more skills. These considerations lead to a question: Why would there be variation within class (i.e., within skill mastery profile)?

Random/Stochastic Variation

There could be non-systematic, random variation from slipping or guessing as previously defined above, from the stochastic element of responding to the item. This source of variation has been traditionally modeled within the core DCMs previously discussed already.

Systematic Variation from Construct Irrelevant Variance

There could also be variation due to some source of construct irrelevancy (Henning, 1989; Ferrier et al., 2011). These potential sources include:

- Poor items
- Differential item functioning
- Item order
- Test-wiseness
- Test format
- Item format
- Speededness
- Test-related fatigue
- Stakes of the testing

- Examinee motivation
- Test anxiety
- Test exposure
- Test preparation
- Testing conditions (e.g., interruptions)

Systematic Variation from Construct Underrepresentation

There could also be variation due to some source of construct underrepresentation from:

- Multidimensionality is higher than specified
- Issues with definitional grain-size (Rupp et al., 2010)
- Testlets/sets of items with common content/item bundles/item clusters

Thus, systematic sources of within skill mastery profile variation could fall within two broad categories: variation irrelevant to the skills attempting to be measured and underrepresentation of skill profiles (or individual skills themselves). Some researchers (e.g., Rosenbaum, 1988) have separately developed methodology *according to which* of these two sources is problematic. Thus, it is important to distinguish the conceptual sources of such systematic variation.

Mis-specifying the number of attributes in DCM could be considered a misspecification of the *Q*-matrix. It is well studied that inaccurate *Q*-matrices degrade classification performance (e.g., Rupp & Templin, 2008b; de la Torre, 2008; Chiu, 2013). Rupp et al. (2010) discuss issues that arise when the definitional grain-size of attributes is less than adequate.

However, when additional multidimensionality due to the test design arises because of dependencies between groups of items share common content, stimulus, or reading passages even when the number of attributes, their definitional grain-size, and chosen DCM are appropriate, then this is just a test design complexity that should be taken into account. Such adjustments could be done analytically to remove them as a source of invalidity towards uses and interpretations of DCM.

There have been at least two extended DCMs proposed to handle such data. Choi (2010) proposed a diagnostic classification mixture Rasch model to detect examinee heterogeneity, such that LI is only assumed after accounting for classes *and* heterogeneity. Templin (2009) and Henson et al. (2015) propose a categorical bifactor model where general content knowledge is specified as continuous latent ability and specific content area knowledge is specified as categorical latent skill mastery/non-mastery with DCM-like terms. This categorical bifactor model could allow for nuisance variation to be accounted for by the general continuous trait. Templin (2009) remarks that he had yet to see an application of the categorical bifactor model to data where fit was not the best relative to other implemented DCMs. Therefore, the generalized version of the categorical bifactor model discussed in Henson et al. (2015) was selected as the current study's generating model to investigate violation of LI induced from underlying mixture MIRT variation. Henson et al.'s (2015) generalized model is consistent with an assumption of a DCM with systematic variation based on conceptual causes of local dependencies.

Study Generating Model: The MCCIRM

To further motivate the generating model used in this study, let the d_i difficulty parameter from the pMIRT model in Eq. (6) have the following composition

$$d_i = \underset{\sim}{\lambda}_i^{\mathrm{T}} \underset{\sim}{\mathrm{h}}\left(\underset{\sim}{q}_i, \underset{\sim}{\alpha}_e\right) + \lambda_{i0}\,, \tag{39}$$

Where again:

$$\underset{\sim}{\lambda}_i^{\mathrm{T}} \underset{\sim}{\mathrm{h}}\left(\underset{\sim}{q}_i, \underset{\sim}{\alpha}_e\right) = \sum_{a=1}^{A} \lambda_{ia} \alpha_{ea} q_{ia} + \sum_{a=1}^{A-1} \sum_{b>a}^{A} \lambda_{iab} \alpha_{ea} q_{ia} \alpha_{eb} q_{ib} + \ldots \tag{40}$$

This gives the generating model considered in this study, referred to as the *multidimensional continuous-categorical item response model* (denoted MCCIRM). The MCCIRM is an extended version of the model presented in Henson et al. (2015), described by the following

$$\operatorname{logit}\left(P\left(X_{ei} = 1 \mid \underset{\sim}{\theta}_e, \underset{\sim}{\alpha}_e\right)\right) = \underset{\sim}{\gamma}_i^{\mathrm{T}} \underset{\sim}{\mathrm{f}}\left(\underset{\sim}{c}_i, \underset{\sim}{\theta}_e\right) + \underset{\sim}{\lambda}_i^{\mathrm{T}} \underset{\sim}{\mathrm{h}}\left(\underset{\sim}{q}_i, \underset{\sim}{\alpha}_e\right) + \lambda_{i0} \tag{41}$$

Where:

$$\underset{\sim}{\gamma}_i^{\mathrm{T}} \underset{\sim}{\mathrm{f}}\left(\underset{\sim}{c}_i, \underset{\sim}{\theta}_e\right) = \sum_{t=1}^{T} \gamma_{it} \theta_{et} c_{it} + \sum_{t=1}^{T-1} \sum_{u>t}^{T} \gamma_{itu} \theta_{et} c_{it} \theta_{eu} c_{iu} + \ldots \tag{42}$$

And

$$\underset{\sim}{\lambda}_i^{\mathrm{T}} \underset{\sim}{\mathrm{h}}\left(\underset{\sim}{q}_i, \underset{\sim}{\alpha}_e\right) = \sum_{a=1}^{A} \lambda_{ia} \alpha_{ea} q_{ia} + \sum_{a=1}^{A-1} \sum_{b>a}^{A} \lambda_{iab} \alpha_{ea} q_{ia} \alpha_{eb} q_{ib} + \ldots \qquad (43)$$

Where model terms are defined by:

X_{ei} is the scored item response for the *e*-th examinee on the *i*-th item, and $X_{ei} = 1$ for a correct response and $X_{ei} = 0$ for an incorrect response.

$e = 1, 2, ..., E$ for the *E* examinees

$i = 1, 2, ..., I$ for the *I* items

$\underset{\sim}{\theta}_e$ is a vector of length *T* of continuous latent abilities for the *e*-th examinee

$\underset{\sim}{\alpha}_e$ is a vector of length *A* of categorical latent attributes for the *e*-th examinee

$\underset{\sim}{\gamma}_i^{\mathrm{T}} \underset{\sim}{\mathrm{f}}\left(\underset{\sim}{c}_i, \underset{\sim}{\theta}_e\right)$ is the *kernel* for the pMIRT-like aspects of the MCCIRM and is defined by the above so that:

θ_{et} is the *t*-th continuous latent ability for the *e*-th examinee

$t = 1, 2, ..., T$ for the *T* continuous latent abilities

γ_{it} is the weight (logit increment) for the *i*-th item on the *t*-th continuous latent ability

c_{it} is an indicator variable for whether the *i*-th item requires the *t*-th continuous latent ability, so that $c_{it} = 1$ if the *i*-th item requires the *t*-th ability and $c_{it} = 0$ otherwise.

γ_{itu} is the weight (logit increment) for the pairwise interaction of the *t*-th continuous latent ability θ_{et} and the *u*-th continuous latent ability θ_{eu} ($u > t$) for the *i*-th item

"+…" represents higher-order interaction terms above pairwise (up to the *T*-th way interaction). These are not studied further in this report (cf. *Assumptions about scope of Generating Model* below).

$\underset{\sim}{\lambda}_i^{\mathrm{T}} \underset{\sim}{\mathrm{h}}\left(\underset{\sim}{q}_i, \underset{\sim}{\alpha}_e\right)$ is the *kernel* for the LCDM-like aspects of the MCCIRM, where:

λ_{i0} is the logit of the probability of correct response for the *i*-th item when no attributes have been mastered (intercept term)

α_{ea} is the categorical latent attribute for the *e*-th examinee for having mastered the *a*-th attribute, so that $\alpha_{ea} = 1$ if the *e*-th examinee has mastered the attribute *a* and $\alpha_{ea} = 0$ otherwise, where $a = 1, 2, ..., A$ for the *A* attributes

λ_{ia} is the weight (logit increment) for item *i* on the *a*-th categorical latent attribute α_{ea}

q_{ia} is an indicator variable for whether the *i*-th item requires the *a*-th categorical latent attribute, so that $q_{ia} = 1$ if the *i*-th item requires the *a*-th attribute and $q_{ia} = 0$ otherwise.

λ_{iab} is the weight (logit increment) for the pairwise interaction of the *a*-th categorical latent attribute α_{ea} and the *b*-th categorical latent attribute α_{eb} ($b > a$) for the *i*-th item

"+…" represents higher-order interaction terms above pairwise (up to the *A*-th way interaction). Interactions above three-way are not studied further in this report (cf. *Assumptions about scope of Generating Model* below). Kunina-Habenicht, Rupp, and Wilhelm (2012) reported that excluding higher-order interactions in estimation did not have practical impact on classification.

It should be noted that this MCCIRM should not be defined as a DCM according to Rupp and Templin (2008), who exclude latent trait models with any continuous latent variables as DCMs. Further, it is prudent to consider next some matrices of interest to further define notation: $\underset{\sim}{X}$, $\underset{\sim}{C}$, $\underset{\sim}{Q}$, $\underset{\sim}{A}$. First, $\underset{\sim}{X}$ is the $E \times I$ matrix of scored item responses given by

$$\underset{\sim}{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1I} \\ X_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ X_{E1} & \cdots & \cdots & X_{EI} \end{bmatrix}, \tag{44}$$

Where each X_{ei} is defined above. Next, $\underset{\sim}{C}$ is the $I \times T$ design matrix of which items measure which continuous latent abilities given by

$$\underset{\sim}{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1T} \\ c_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ c_{I1} & \cdots & \cdots & c_{IT} \end{bmatrix}, \tag{45}$$

Where each c_{it} is defined above. Next, $\underset{\sim}{Q}$ is the $I \times A$ design matrix of which items measure which categorical latent attributes (i.e., the *Q*-matrix) described as

$$\underset{\sim}{Q} = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1A} \\ q_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ q_{I1} & \cdots & \cdots & q_{IA} \end{bmatrix}, \tag{46}$$

Where each q_{ia} is defined above. Note that when simple structure is assumed for *i*-th item, that $\sum_{a=1}^{A} q_{ia} = 1$ and for a complexity of two that $\sum_{a=1}^{A} q_{ia} = 2$ (i.e., the row sums of the *Q*-matrix are one and two, respectively). Finally, $\underset{\sim}{A}$ is the $E \times A$ attribute matrix detailing which examinees possess which categorical latent attributes based on each α_{ea} as defined above given as

$$\underset{\sim}{A} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1A} \\ \alpha_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \alpha_{E1} & \cdots & \cdots & \alpha_{EA} \end{bmatrix}. \tag{47}$$

Assumptions about Scope of Generating Model

Further assumptions made about the MCCIRM in order to investigate the study research questions are as follows:

All g_i and γ_{itu} are fixed to zero in the generalized pMIRT aspects of the MCCIRM. Thus, the pMIRT-part of the MCCIRM reduces to a C-MIRT-like contribution.

It is assumed $T = A$ and all *t* correspond to the same value of *a* for all *a*.

Assuming $T = A$, it is further assumed that $\underset{\sim}{C} = \underset{\sim}{Q}$. That is, that the design matrix of continuous latent abilities to items is the same as the design matrix of categorical latent attributes to items. This provides one companion ability for each attribute.

An average complexity of loadings from item to attribute of two is assumed. Simple structure with items loading on only one attribute is also studied.

A total of $A = 3$ attributes are considered in this study.

A single approach to modeling the structural component of each DCM will be used: the saturated (unstructured) structural modeling approach, which estimates the $2^3 - 1 = 7$ parameters for the 8 mixing proportions directly.

EAP estimation of attribute profile probabilities and attribute-wise marginal probabilities will be used, following estimation of item parameters in CDM in R.

Given these assumptions, then the $\text{logit}\left(P\left(X_{ei} = 1 \mid \underset{\sim}{\theta}_e, \underset{\sim}{\alpha}_e\right)\right)$ of the particular MCCIRM examined in this study is given by

$$\sum_{a=1}^{A} \gamma_{ia} \theta_{ea} c_{ia} + \sum_{a=1}^{A} \lambda_{ia} \alpha_{ea} q_{ia} + \sum_{a=1}^{A-1} \sum_{b>a}^{A} \lambda_{iab} \alpha_{ea} q_{ia} \alpha_{eb} q_{ib} + \lambda_{i123} \alpha_{e1} q_{i1} \alpha_{e2} q_{i2} \alpha_{e3} q_{i3} + \lambda_{i0} \quad (48)$$

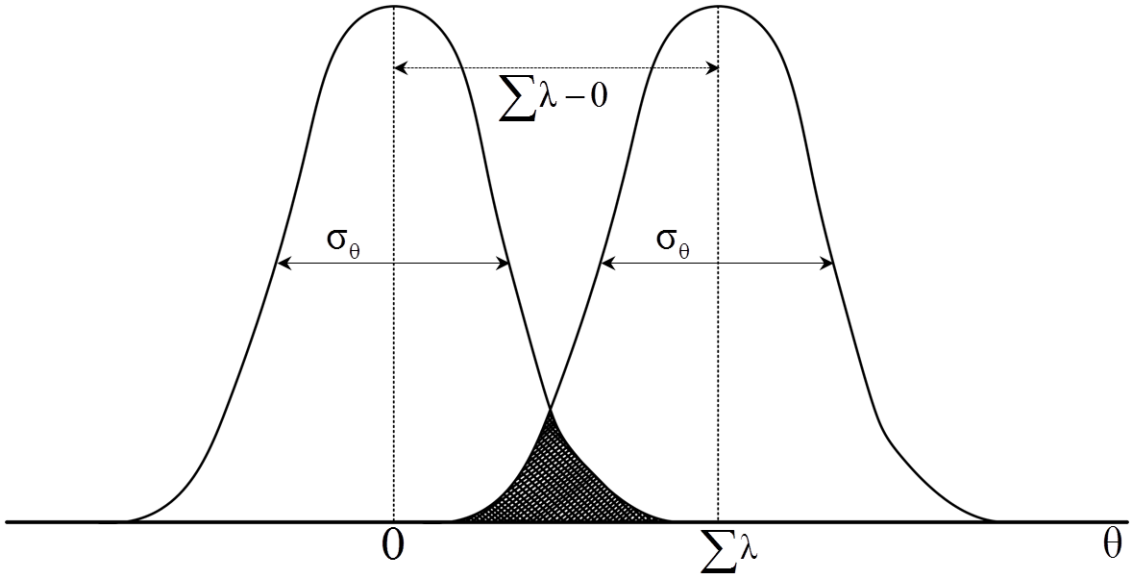
From the above when $A = 3$, we have at most $A + \left[A(A-1)\right]/2 + 1 = 7$ weight terms (λ). These terms in addition to λ_{i0} serve as a single difficulty parameter for a CMIRT model as one perspective. Above and beyond the λ_{i0} term, positive λ_{ia}, λ_{iab}, or λ_{i123} serves to increase the probability of correct item response (holding all else

consistent). Therefore, if all λ_{ia}, λ_{iab}, and λ_{i123} terms are zero (or their sum), then the above model collapses to a 2PL CMIRT model based on Eq. (2) (i.e., where $g_i = 0$).

Continuous Ability Degeneracy from Different Perspectives

Given the particular MCCIRM established for this study in Eq. (48), this could be viewed as a LCDM with nonzero continuous ability variance present based on the C-MIRT aspects that have been added. That is, if all $\theta_{ea} = 0$ in Eq. (48), then the MCCIRM collapses to the LCDM. Likewise, if all $Var(\theta_{ea}) \to 0$, then again the MCCIRM reduces to the LCDM. Thus, we could conceive this MCCIRM as a LCDM with underlying mixture MIRT when $Var(\theta_{ea}) > 0$. So, if the dichotomous assumed nature of any attribute α is interpreted as skill mastery or nonmastery, then Figure 2 conveys the "offset" in location of ability distributions of two Normal mixtures when the sum of the applicable λ_{ia}, λ_{iab}, and λ_{i123} terms in Eq. (48) is > 0 (denoted by $\sum\lambda$), which would be true for masters of 1+ skills:

Figure 2. Ability Offset According to Attribute Mastery



Here, mixtures of the continuous ability are by choice conceptualized to have a mean of zero for complete skill nonmasters to set the ability metric. Then, the mean of the mixture for complete skill masters would by offset by the sum of the LCDM-like parameters, $\sum\lambda$, in the MCCIRM in the logit scale. Thus, this can be considered a mixture IRT model (e.g., for a given continuous ability θ), where the two mixture distributions for skill nonmasters and masters have the same variance, σ_θ^2, but different location parameters of $\mu_\theta = 0$ and $\mu_\theta = \sum\lambda$, respectively.

From the above the MCCIRM can then be considered as a DCM with LI violation as another perspective. In the case that ability variances shrinks then the MCCIRM approaches the standard LCDM. So, the extent to which non-negligible variance of abilities is present could possibly quantify the magnitude of the violation of LI.

Multiple Abilities: The Simple Structure Case

In terms of the MCCIRM equation given in Eq. (48), *simple structure* is defined as the joint condition for the *i*-th item where $q_{ia} = 1$ when the *i*-th item requires the *a*-th attribute and all other $q_{ia'} = 0$ for $a \neq a'$ and likewise for c_{ia}. Given this, when there are multiple abilities with nonzero variance but items have simple structure, again the CMIRT portion of the assumed MCCIRM model given in Eq. (48) can be conceptualized as the specification of mixtures of two ability distributions per ability dimension, given by

$$\theta_{ea} \sim \left(1 - \pi_{\mathrm{M}}\right) \cdot Normal\left(0,\ \sigma^2_{\theta_{ea}}\right) + \pi_{\mathrm{M}} \cdot Normal\left(\sum \lambda_a,\ \sigma^2_{\theta_{ea}}\right), \quad (49)$$

Where θ_{ea} is the *a*-th ability for the *e*-th examinee and is Normally distributed with a mean of zero and variance of $\sigma^2_{\theta_{ea}}$ if the examinee is a non-master of the *a*-th skill, and is Normally distributed with a mean of $\sum \lambda_a$ and variance of $\sigma^2_{\theta_{ea}}$ if the examinee is a skill master of the *a*-th skill. As the $\sum \lambda_a$ increases above zero, the overlap between these ability distributions becomes less for a given level of ability variance $\sigma^2_{\theta_{ea}}$ and therefore diagnostic classification would conceptually be more recoverable (i.e., classification is more accurate) because skill mastery raises the probability of correct response to a greater degree. As the $\sum \lambda_a$ approaches zero, the overlap between these ability distributions becomes greater for a given level of ability variance $\sigma^2_{\theta_{ea}}$ and therefore

diagnostic classification would conceptually be less accurate. Likewise, for a given $\sum \lambda_a$ value as the common mixture distribution ability variance $\sigma^2_{\theta_{ea}}$ increases then so does the overlap between the ability mixtures. Similarly, for a given $\sum \lambda_a$ value as the common mixture distribution ability variance $\sigma^2_{\theta_{ea}}$ decreases to zero then the overlap between the ability mixtures decreases as well. This suggests the following proportionality,

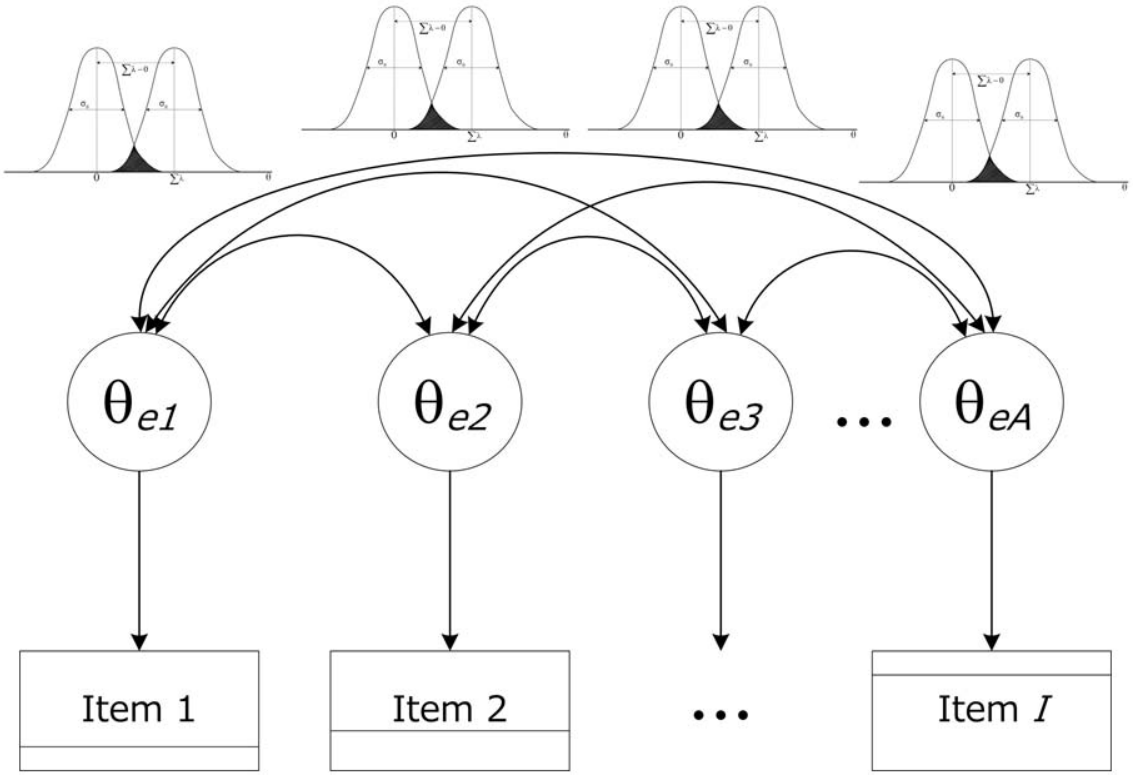
$$P_c \propto \frac{1}{\sigma^2_{\theta_{ea}}}, \tag{50}$$

So that the proportion with correct diagnostic classification, P_c, is hypothesized to be proportional to the common ability variance, $\sigma^2_{\theta_{ea}}$. However, even with zero common variance, classification is not perfect due to random/stochastic variation within DCM (i.e., due to slipping or guessing). Another measure that can be considered is the ratio of the maximum separation in location of any given pair of two ability mixture distributions between complete skill masters and nonmasters to their common standard deviation of ability, given by

$$\Delta = \frac{\sum \lambda_a - 0}{\sqrt{\sigma^2_{\theta_{ea}}}}. \tag{51}$$

This "effect size" measure, Δ, is useful because in relation to the above, as the denominator goes to zero (for consistent nonzero $\sum \lambda_a$) then the MCCIRM generating model collapses to a DCM because the CMIRT part reduces to a constant and can be aggregated into the LCDM-part intercept. As the numerator goes to zero (for consistent nonzero $\sigma^2_{\theta_{ea}}$) then the MCCIRM generating model collapses to a (nonmixture) CMIRT model. Figure 3 relates these scenarios back to a mixture CMIRT model under the case of simple structure:

Figure 3. Mixture MIRT Model with Simple Loading Structure and Mastery Location Offsets

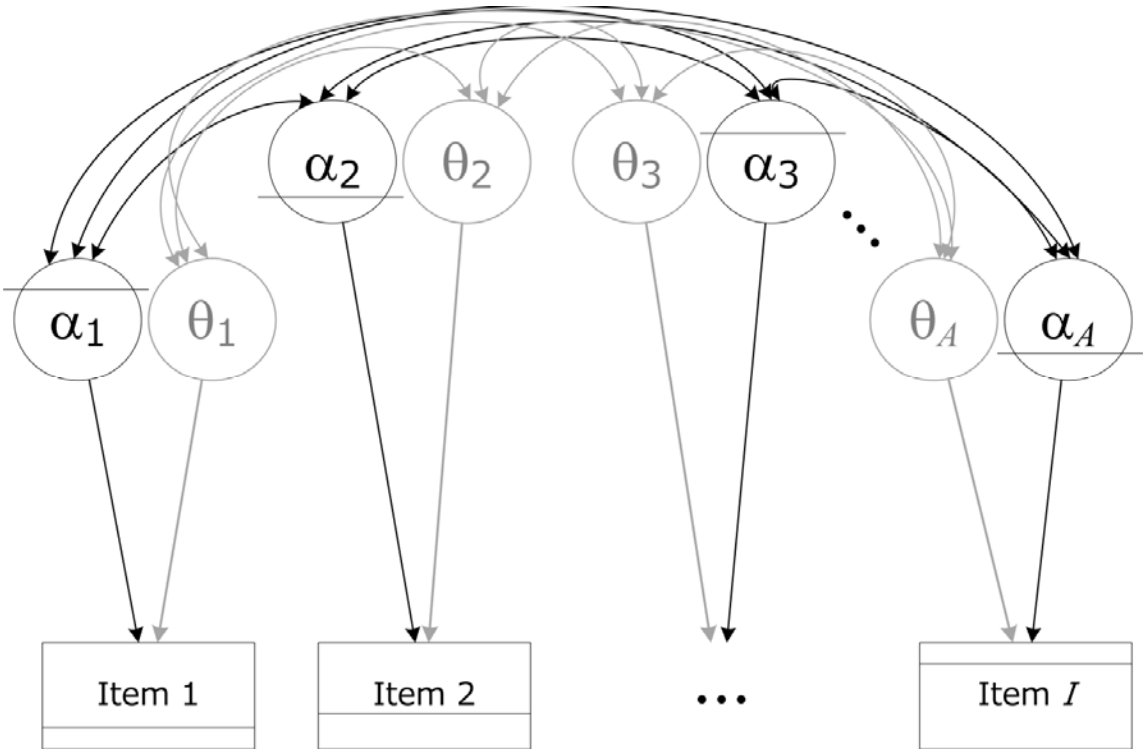


From Figure 3 one can see that as the variance degenerates (i.e., $\sigma^2_{\theta_{ea}} \to 0$), the two bell curves contract until they collapse to infinitesimally narrow bars corresponding to master and nonmaster classes. When variance does not degenerate and as $\sum \lambda_a \to 0$, the location offset decreases until the master and nonmaster classes perfectly overlap. Some additional items are of note. First, the ability mixtures given in Eq. (49) have $\pi_{\mathrm{M}} = 0.50$ for all *a* attributes by study design assumption. Other mixing proportions could possibly be investigated in future research. Another study design choice could also allow for unequal $\sigma^2_{\theta a}$ as specified above in Eq. (51). However, in this study it is further assumed that

$$\sigma^2_{\theta_{e1}} = \sigma^2_{\theta_{e2}} = \ldots = \sigma^2_{\theta_{eA}} \equiv \sigma^2_{\theta_{e\bullet}} \tag{52}$$

So that for each skill/ability pair that there is a common ability variance denoted as $\sigma^2_{\theta_{e\bullet}}$ (i.e., the homogeneous variance $\sigma^2_{\theta a}$ for all values of *a*). Finally, another study design specific choice is in assuming $\sum \lambda \equiv \left(\sum_i \lambda_i \Big/ I \right)$ for defining effect size. This assumption is made so that uniform offset effects could be studied. Future studies could investigate heterogeneous separation between skill masters and nonmasters for different attributes and/or items. Given these, Figure 3 could be also visualized within the MCCIRM framework to be:

Figure 4. Conceptual MCCIRM under Simple Structure



To summarize, the main idea is that when the common ability variance $\sigma^2_{\theta_{e\bullet}}$ goes to zero, then the MCCIRM reduces to a DCM as described above. Likewise, when the logit increment of correct response (sum of the LCDM-like parameters) for skill masters is zero, there is no mixture location offset in ability distributions and they coincide exactly (for nonzero ability variance).

Multiple Abilities: The Complex Structure Case

In the simple structure case above only one ability is considered, θ_{ea}. For the MCCIRM given in Eq. (48), *complex structure* is defined as the joint condition for the *i*-th item where $q_{ia} = 1$ when the *i*-th item requires the *a*-th attribute and at least one other

$q_{ia'} = 1$ for $a \neq a'$ and likewise for c_{ia}. When multiple abilities are required for the *i*-th item for the *e*-th examinee, one way to conceptualize this within the MCCIRM of Eq. (48) is to consider it as a weighted composite, $\ddot{\theta}_e$, of the *A* multiple abilities, and in particular as discussed above, one ability per attribute,

$$\ddot{\theta}_e = \sum_{a=1}^{A} \gamma_{ia} \theta_{ea} c_{ia} \tag{53}$$

This weighted ability composite aspect of the MCCIRM is akin to a "reference" composite (Reckase, 2009) estimated in a unidimensional IRT model when there are actually multiple underlying abilities. In the complex case, when various θ_{ea} are correlated, then the induced effect size is smaller than that in the simple case because the composite variance is larger than just the sum of the separate variances (See *Appendix A). Variance of continuous trait composite in the MCCIRM for complex structure* for more details regarding ability variance under complex structure.

Again, as the effect size shrinks then the mixtures overlap to a greater degree, making classification accuracy hypothetically more difficult to achieve. Thus, under these conditions complex structure is hypothesized to have an even more detrimental effect on diagnostic classification relative to simple structure because of this nature of the composite variance and the covariance introduced by the individual positively correlated abilities. The subsequent methods described in the remainder of the study address how

such effect sizes were investigated in an appropriate way to examine what degree of variance degeneracy leads to impact on classification accuracy.

Compensatory Processes for Complex Structure

Additional comments should be given about the case when loading structure is complex and the response process is compensatory. To illustrate this, consider a hypothetical example for a particular item with complex loading structure and requires two attributes (e.g., addition and multiplication). Further assume that the response process is compensatory such that if an examinee can respond correctly about one attribute that this can somewhat "make up" for a lack of having the other skill. Recall that the formula for the LCDM representation of the CRUM was given in Eq. (21), and for this particular example well represents the logit of correct response to the fifth item on a diagnostic assessment, given by

$$\text{logit}\left(P\left(X_{e5}=1 \mid \underset{\sim}{\alpha}_e, \underset{\sim}{\lambda}_5, \lambda_{5,0}\right)\right) = \lambda_{51}\alpha_{e1}q_{51} + \lambda_{52}\alpha_{e2}q_{52} + \lambda_{5,0} \quad (54)$$

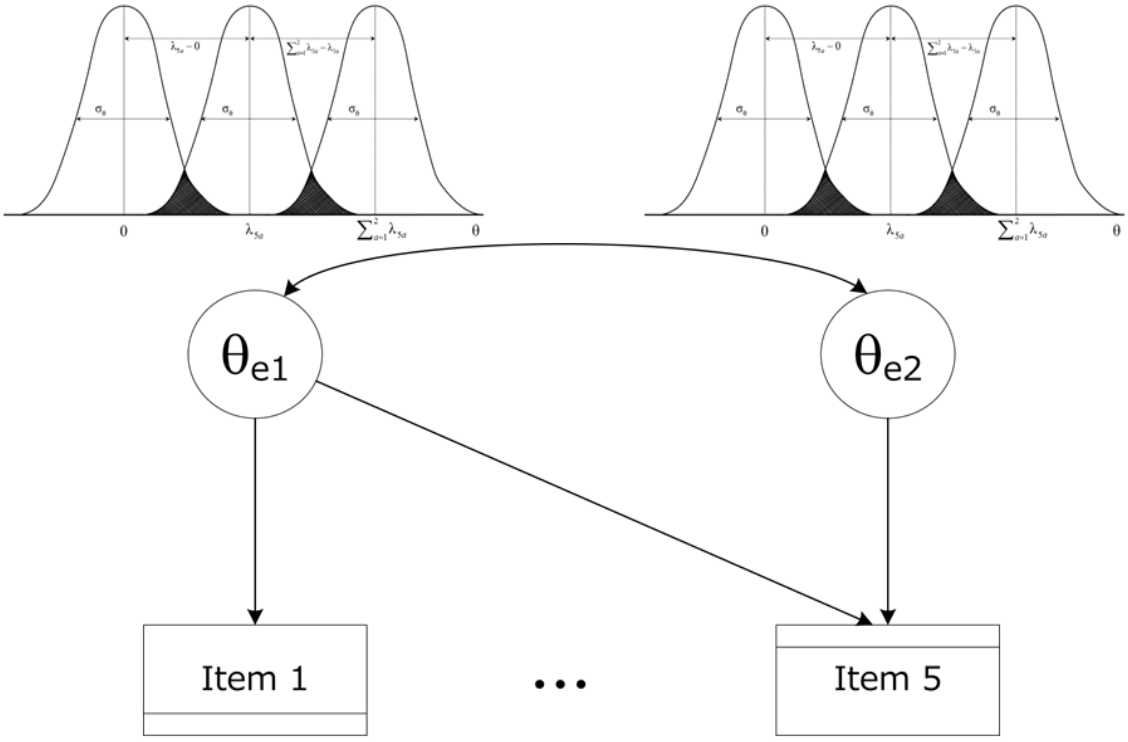
It is apparent from Eq. (54) that different log-odds for nonmasters of both attributes, masters of only one attribute, and masters of both attributes, are implied and are given in the table:

Table 7. LCDM CRUM Weights with Two Attributes, Complex Structure for Example Item

Scenario	Log-odds of correct response
Nonmasters of both attributes	$\lambda_{5,0}$
Masters of only one of α_{e1} or α_{e2}	$\lambda_{51} + \lambda_{5,0}$ if only mastered α_{e1} or $\lambda_{52} + \lambda_{5,0}$ if only mastered α_{e2}
Masters of both α_{e1} and α_{e2}	$\lambda_{51} + \lambda_{52} + \lambda_{5,0}$

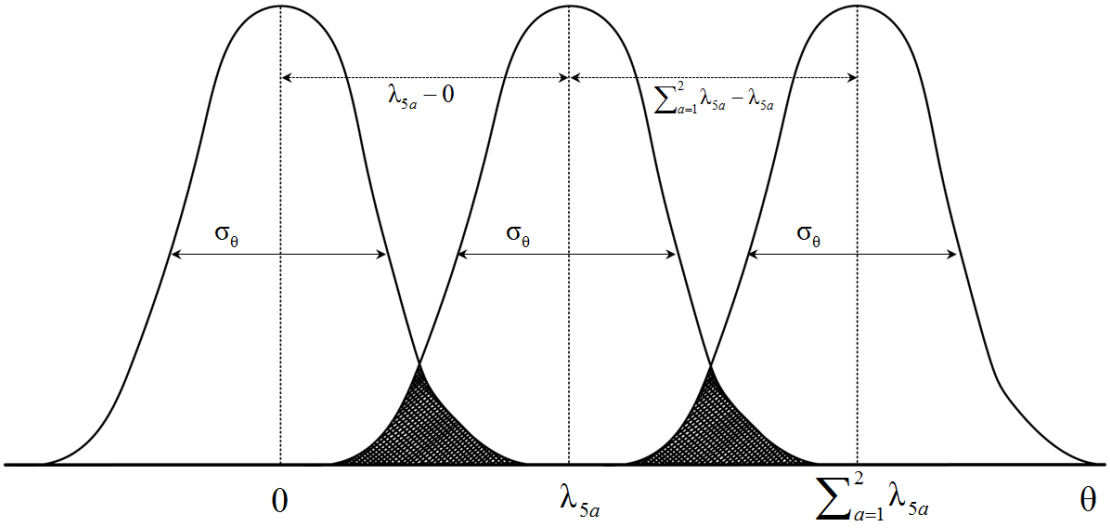
Thus, there is more than just one change in log-odds of correct response when there is item-to-skill loading complexity for compensatory processes. If combinations of mastery of the two skills as location offsets are again conceptualized in a mixture MIRT model, then the previous Figure 3 could be modified to be:

Figure 5. Mixture CMIRT Model with Two Abilities and Complex Structure for Item #5



Although there are two underlying mixtures for item #1 due to its simple loading structure onto θ_{e1}, there are *three* underlying mixtures for item #5 which loads on θ_{e1} and θ_{e2} here, where each mixture and their overlap are illustrated by:

Figure 6. Example of Compensatory Skill Mastery Offset for Complex Structure



Thus, for item #5 three underlying mixtures are induced: (1) for nonmasters of skills 1 and 2 with location equal to 0, (2) masters of only one skill with location equal to either λ_{51} if possessing skill 1 or λ_{52} if possessing skill 2, and (3) masters of both skills with location equal to $\lambda_{51} + \lambda_{52}$.

Because of this, when defining the effect size as in Eq. (51), *the maximal difference between complete nonmasters and total masters was used in data generation*. Here in the example, this would correspond to $\Delta = \left(\sum_{a=1}^{2} \lambda_{5a} - 0 \right) \Big/ \sigma_{\theta_e.}$, which is the difference in location between masters of both skills compared to nonmasters of both skills. Defining the effect size in this way with the maximal difference will create consistency between effect sizes in noncompensatory and compensatory LCDMs under complex structure for minimal and maximal mastery groups. An implication of this is that classification for masters of less than all skills for compensatory models under

complex structure could suffer more, because as can be seen above in the example, overlap with nonmasters would be worse than described by the maximal difference effect size definition. This effect size convention is only for data generation, as classification for profiles of examinees who lack all or less than all skills is examined among the study findings as well as for attribute-wise marginal attainment from estimated DCMs.

For a DCM such as the DINA model that is strictly noncompensatory, there are no "middle" mixtures due to increase in log-odds of correct response from having only mastered a single skill among multiple required skills. Similarly, when there is simple loading structure an item will only load onto one skill, thereby only providing a single increase in the log-odds of response for mastery. This last point regarding simple structure also impacts how total ability variance within the MCCIRM is controlled as discussed above versus other conceptual possibilities. Thus, more consideration regarding these other conceptualizations are attended to first before going further.

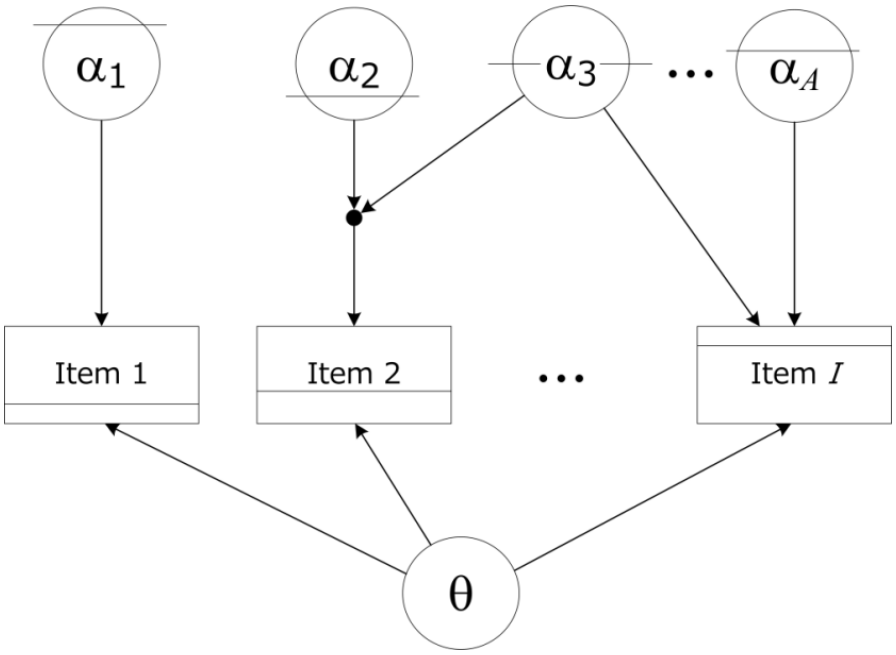
Introducing Systematic Within-Class Variation from Different Mechanisms

Thus far it has been assumed that within-class variation is introduced through underlying mixture distributions with one skill per companion continuous ability. From a DM perspective, this represents variation beyond random noise such as slipping or guessing, whereby each latent skill is by design assumed to be affected equally by introducing continuous ability variation into the modeling for diagnostic classification.

However, there are *other* ways in which such continuous ability variance from underlying mixture MIRT-aspects can be introduced within the MCCIRM framework: (a) a categorical bifactor model with a single general continuous trait and (b) introducing one

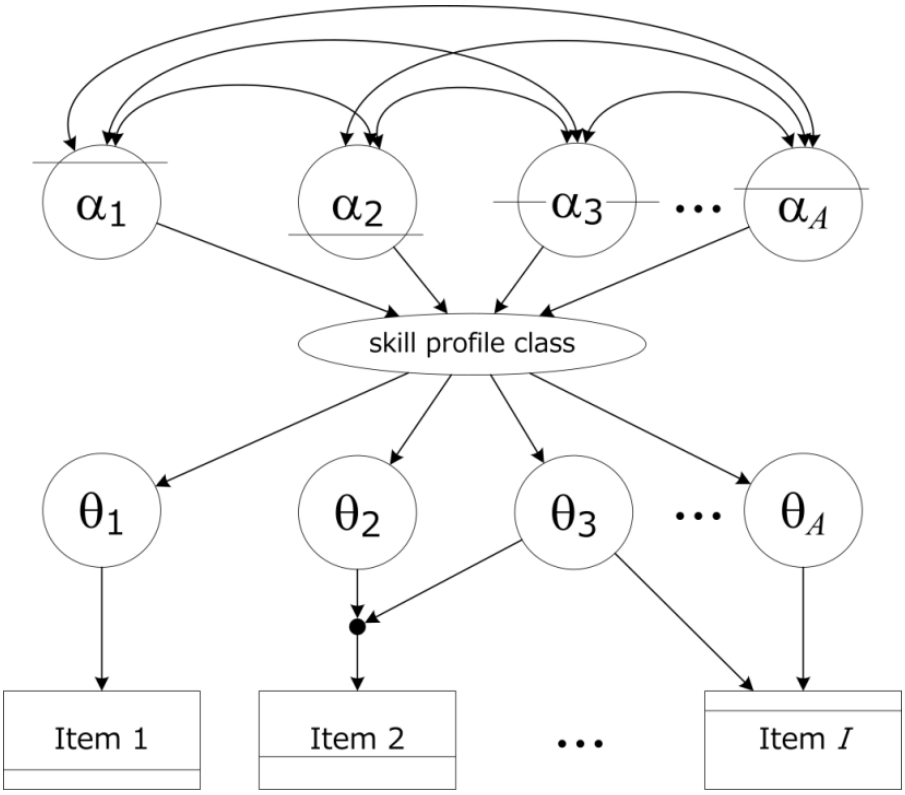
continuous trait per DCM skill profile (von Davier, 2005) where there are unique changes in probability of correct response. An illustration of the former approach using a Categorical Bi-factor model discussed in Henson et al. (2015) is given by:

Figure 7. Categorical Bi-Factor Model



For the latter approach introducing one continuous trait per DCM skill *profile*, this could be possibly visualized as the following:

Figure 8. Conceptual Model Illustrating Continuous Traits per DCM Skill Profile



This conceptualization is more in line with von Davier's GDM, which has been previously used for diagnostic classification where just the latent traits themselves are constrained to -1 and +1. The way systematic variation would be introduced here would be through introducing one ability for nonmasters (located at -1) and additional abilities for every unique increase in response probability when skill profile changes. So, rather than one ability per skill, the skill profile would act like a mediating step here in terms of allowing increasing ability variance to influence diagnostic classification.

Previous Measurement Models with Continuous and Categorical Traits

Thus, from the above there are other ways that could introduce within-class systematic variation. The current study accomplishes this through a particular choice of the MCCIRM, which has both continuous and categorical latent traits. However, there have been other measurement models with both kinds of traits proposed as well: (a) the Full Reparameterized Unified Model (Roussos et al., 2007a), (b) the General Diagnostic Model (von Davier, 2005), and (c) the diagnostic modeling framework from Rupp et al. (2010), for which the categorical bifactor model from Henson et al. (2014) follows. Lee and Sawaki (2009), Jang (2009a), Li (2013) and Zhao (2013) all used the fusion model in their applied studies for diagnostic classification, while Kunina-Habenicht et al. (2009) and Lee and Sawaki (2009) both used the GDM in their applied studies for diagnostic classification.

As seen above in Figure 7, Henson et al.'s (2014) categorical bifactor model is a derivation from the general diagnostic modeling framework presented in Rupp et al. (2010):

Figure 9. The Four Steps of the Diagnostic Modeling Framework

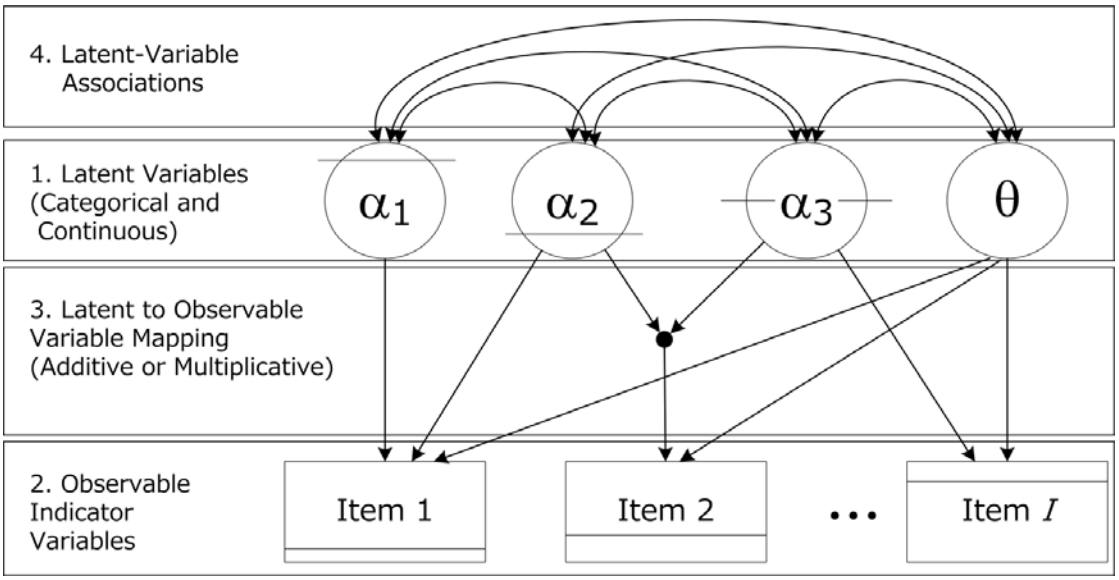


Figure 3.7 from Rupp et al., 2010; adapted with permission; Copyright 2010 with Guilford.

For the categorical bifactor model, there is only one (general) continuous ability assumed to be present, and is uncorrelated with the other latent categorical attributes. Thus, there have been other educational measurement models besides the MCCIRM proposed that incorporate both latent continuous abilities and latent categorical attributes. However, no prior studies to date have examined the impact on skill mastery diagnosis if such a measurement model with this duality were truth when performing diagnostic classification using DCM, where conceptually representing a violation of the LI assumption in DCM.

In sum, when continuous traits are additionally active due to the aforementioned conceptual sources, then diagnostic classification can be impacted. The remainder of the current study is devoted to describing methods and findings from this impact under a variety of conditions identified in the review of the literature within this chapter.

CHAPTER III

METHODOLOGY

Research Questions and Hypotheses

The four study research questions and six hypotheses stated in the first chapter are:

*RQ*1. Does increasing variance of continuous abilities in mixtures of mastery/non-mastery groups cause detectable violations of local independence when performing diagnostic classification?

*H*1. Increasing variance of continuous abilities generated from the MCCIRM is detected by increasingly large Yen's *Q*3 statistics based on results from DCMs without continuous ability.

*RQ*2. Does increasing variance of continuous abilities degrade model fit when performing diagnostic classification?

*H*2. Increasing variance of continuous abilities within the MCCIRM substantially degrades item parameter recovery in LCDM estimates without continuous ability.

*H*3. Increasing variance of continuous abilities within the MCCIRM leads to overestimation of attribute-to-attribute correlations under the LCDM without continuous ability.

*RQ*3. Does increasing variance of continuous abilities lower accuracy of diagnostic classification?

*H*4. Increasing variance of continuous abilities within the MCCIRM substantially degrades proportion of correct diagnostic classification based on estimated LCDMs without continuous ability.

*RQ*4. Are effects of increasing variance of continuous abilities on accuracy of diagnostic classification affected by complexity or compensation?

*H*5. Complex structure substantially degrades proportion of correct diagnostic classification based on the LCDM without continuous ability when variance of continuous abilities within the MCCIRM increases.

*H*6. Compensatory versus noncompensatory processes leads to substantially lower proportion of correct diagnostic classification based on the LCDM without continuous ability when variance of continuous abilities within the MCCIRM increases.

Research Design

A simulation study is conducted to investigate the claims of the study hypotheses. This design was chosen because these questions have never been previously investigated when truth was known and strict control of study conditions is desired. The following *Analysis plan* details conditions and scope of the simulation study in order to conduct the investigation.

Analysis Plan

What DCMs are Studied

The saturated LCDM as well as the following three core DCMs are studied within the LCDM framework: the CRUM, DINO, and DINA models. The LCDM specification

of these core DCMs and the saturated LCDM were detailed in the previous chapter on the *Review of the Literature*.

Simulation Methods and Conditions

Table 8 details the study methods and conditions:

Table 8. Simulation Methods and Conditions

Condition	Levels
Diagnostic classification model	CRUM, DINA, DINO, LCDM
Number of attributes, abilities	3
Average item-to-skill/ability complexity[1]	1, 2
Number of examinees	10,000
Number of items per skill	5, 10
Correlation of attributes/abilities	0.70
Effect size[2], Δ	0.8, 2, 3, True DCM
Probability of complete non-mastery	~*Uniform*(0.05, 0.25)
Probability of complete mastery	~*Uniform*(0.75, 0.95)

**Note*. 1. *Q*-matrix entries were the same for attributes and abilities for each item;
2. Effect size was defined to be ratio of the average sum of LCDM weights above the intercept divided by the standard deviation of ability.

Conditions included: (a) simulated probabilities of complete nonmastery according to *Uniform*(0.05, 0.25) and complete mastery according to *Uniform*(0.75, 0.95), (b) DCMs studied included the CRUM, DINA, DINO, and saturated LCDM, (c) simple and complex structure (target average complexity = 2), (d) test lengths of 15 and 30 items, and (e) effect size, Δ, as defined above targeted to be $\Delta = 0.8$ (large violation in

LI), $\Delta = 2$ or 3 (some violation in LI), and True DCM (no violation in LI / zero ability variance). All scenarios considered three attributes with three companion continuous abilities (up to three each according to generated *Q*-matrices). Population associations among attributes were specified to be 0.70 and likewise among abilities (attributes and abilities were considered uncorrelated with each other). All replications were performed with a sample size of 10,000 simulees and all conditions were based on 100 replications each for approximating sampling distributions.

From Table 8 above, the choice to fix the number of attributes at three facilitates explication of results as this would result in eight skill profiles and is within the range of DCM application studies discussed in the previous chapter. Studying both simple and complex structure gives the study breadth. Although DCMs have most benefit under complex structure, they have been applied to assessments possessing items with simple structure only (e.g., Henson et al., 2015; Jurich & Bradshaw, 2014). Bradshaw and Templin (2014) report an applied study using the LCDM where 19 of 27 items had simple structure and the rest had complexity of 2. Thus, simulation conditions for complexity are within the realm of recent DCM application studies.

The number of examinees was set to be high enough at 10,000 so that sample size effects would not be detrimental (e.g., Kunina-Habenicht, Rupp, & Wilhelm, 2012). Rupp et al. (2010) simulated data to demonstrate estimation using 10,000 examinees as well and remark that "To ensure that the parameter estimates would be estimated accurately, we generated data for 10,000 respondents." For test length, two conditions were chosen where 15 total items will be considered (5 per skill) and 30 total items (10

per skill). These two conditions are within the range of recent DCM applied studies summarized in Table 7, and are intended to reflect a shorter and longer assessment length. For the correlations among attributes or among continuous abilities, these were fixed to be the same value where a magnitude of 0.70 was chosen. Correlations among attributes of 0.70 were chosen by Madison and Bradshaw (2014) for their LCDM *Q*-matrix design simulation study, and represents a non-trivial association between latent traits, but not so high that extreme redundancy is observed (i.e., traits remain distinct). Henson et al. (2015) found a tetrachoric correlation of 0.81 between two attributes after use of the categorical bi-factor model. Jurich and Bradshaw (2014) found a range of attribute tetrachoric correlations between 0.56 and 0.94. Zhao (2013) found attribute correlation ranges of 0.62 to 0.96, 0.02 to 0.95, and 0.47 to 0.90 for three different applications, respectively. Thus, a correlation choice of 0.70 among attributes is within the realm of recent DCM application studies.

The new set of conditions this study investigates are the related to the effect size defined in Eq. (67) from the CMIRT aspects of the MCCIRM generating model. These effect size values were chosen according to the earlier Tables 1 and 2, which describe how the effect sizes correspond to various amounts of overlap (i.e., using the *OVL* percentage). Here, $\Delta \to +\infty$ corresponds to a "true DCM" condition, since there is no overlap for the case. This effect size serves as a benchmark to judge performance against all other scenarios. The other chosen values then represent various decreasing effect sizes due to increasing total continuous ability variance under scenarios for the numerator in Eq. (67) on logit separation between maximal masters and minimal nonmasters

according to previous published studies. The commonly chosen effect size of 0.8 is included among these (Cohen, 1988). The values $\Delta = 2$ and 3 are chosen in order to provide approximately 32% and 13% overlap among underlying continuous ability mixtures (using the *OVL* coefficient).

It was expected for smaller effect sizes (e.g., $\Delta = 0.8$) that non-convergence of DCM estimation could occur. Non-convergence was defined as exceeding the maximum default number of iterations within the E-M algorithm (A. Robitzsch, personal communication, April 7, 2015), and was tallied and reported across repetitions by condition (but no further adjustments made such loosing default convergence criteria in order to aid convergence).

A key idea is that logit separation values of the effect size numerator were chosen according to the current DCM application studies so that probability of complete nonmastery was between 0.05 to 0.25 and complete mastery was between 0.75 and 0.95. Controlling this range allowed the separation between the minimal and maximal mastery groups are similar between different models, even if they use different condensation kernels (e.g., DINA and DINO).

Table 9. Required $\sqrt{Var(\theta_{ea})}$ According to Differences in Probability of Complete Nonmastery versus Complete Mastery

Nonmastery vs. Mastery	True DCM	$\Delta = 3$	$\Delta = 2$	$\Delta = 0.8$
0.05 versus 0.95	0	1.963	2.944	7.361
0.20 versus 0.80	0	0.924	1.386	3.466

**Note*. Values of continuous ability SD according to logit separation corresponding to difference in probability of complete nonmastery versus complete mastery.

From the previous Table 8, the simulation study design has 1×1×1×1×2×1×4×1×1 cells for simple structure case + 4×1×1×1×2×1×4×1×1 cells for complex structure (average complexity = 2) for a total of 40 cells. There was 100 replications per cell for a total of 4,000 replications.

Data Generation

Q-Matrix Generation

For simple structure, *Q*-matrices had 5 items for each attribute for test length of 15 items and 10 items for each attribute for length of 30 items. Items were put in a random order using a random order generator. For the complex condition described below an average item complexity of two was targeted (observed was 1.83 for 15 items and 2.12 for 30 items across 100 replications each). For complex *Q*-matrices, these were constructed based on random draws from a Binomial distribution with a probability of success proportional to item complexity divided by the total number of attributes. All complex *Q*-matrices had two simple structure items per attribute for each test length (Madison & Bradshaw, 2014). No attribute could be measured by all items and each

attribute had to be measured by at least two items. It took 10,249 generated *Q*-matrices to obtain 100 with these characteristics for the 15 item test length condition and 4,664 generated *Q*-matrices to obtain 100 for the 30 item test length condition.

Examinee Attribute Profile and Ability Generation

Attribute patterns were simulated in a manner consistent with Shu, Henson, & Willse (2013) and Willse, Henson, and Templin (2007). Briefly, a multivariate normal distribution was simulated then dichotomized in simulating attributes. This distribution was specified with a matrix defining the tetrachoric correlations among attributes, and simulated based on a population parameter of 0.70 indicating a strong association between attributes. The marginal probability of mastery of each attribute was specified to be 0.50.

Examinee ability was simulated from multivariate Normal distribution such that $\underset{\sim}{\theta} \sim MVN\left(\underset{\sim}{\mu}_{\underset{\sim}{\theta}} = (0,0,0),\ \underset{\sim}{\Sigma}_{\underset{\sim}{\theta}}\right)$ where $\underset{\sim}{\Sigma}_{\underset{\sim}{\theta}}$ had the form based on Eqs. (A1.10) and (A1.11) in *Appendix A* (compound symmetric) with $\rho = Corr\left(\theta_{ea}, \theta_{ea'}\right) = 0.70$ and $\sigma^2_{\theta_{e.}}$ according to Table 9 above (dictated by the targeted effect size for a given condition).

Item Parameter Generation

The following study conditions were considered in order to introduce effect size. First, the probability of complete non-mastery was sampled from a Uniform distribution based on $P\left(X_{ei} = 1 \mid \underset{\sim}{\alpha} = (0,0,0)\right) \sim Uniform(0.05,\ 0.25)$. Therefore, the corresponding

intercept term λ_0 from a LCDM under three attributes would then be drawn from $\lambda_0 \sim Uniform(-2.944, -1.386)$ since $P(X_{ei}=1 \mid \underset{\sim}{\alpha}=(0,0,0)) = \exp(\lambda_0)/(1+\exp(\lambda_0))$.

Likewise, the probability of complete mastery was sampled from another Uniform distribution based on $P(X_{ei}=1 \mid \underset{\sim}{\alpha}=(1,1,1)) \sim Uniform(0.75,\ 0.95)$. The corresponding sum of all LCDM weights including λ_0 would then be based on a $Uniform(1.099,\ 2.944)$. Subsequently the sum of all LCDM weights *above* λ_0, denoted as $\sum_i \lambda_i$, was computed as the difference between the total sum of all weights and the draw of λ_0. For the four DCMs considered (CRUM, DINA, DINO, LCDM), the $\sum_i \lambda_i$ were allocated to individual weights in play according to the related *Q*-matrix entries for a given item in each replication. Specifically, $\sum_i \lambda_i$ was evenly distributed among individual weights for the CRUM by assigning values of $\sum_i \lambda_i$ divided by the sum of *Q*-matrix row entries for a particular item (i.e., the one weight gets all of $\sum_i \lambda_i$ in simple structure and each weight gets half of $\sum_i \lambda_i$ if complexity = 2). For the DINA model, the weight for the highest order term received the full $\sum_i \lambda_i$ value. Weights in the DINO model were allocated according to that described in Henson et al. (2009) (e.g., for complexity = 3, main effect

weights received $\sum_i \lambda_i$, two-way interaction effects received $-\sum_i \lambda_i$, and the three-way interaction received $\sum_i \lambda_i$).

For the LCDM, weights were all positive and allocated as follows. For simple structure, the single weight received the full $\sum_i \lambda_i$. For item complexity = 3, main effects and the three-way interaction received a value according to $\left(k\sum_i \lambda_i\right) \Big/ \left(\varsigma + 1\right)$, where ς is complexity (in this case $\varsigma = 3$) and $k \sim Uniform\left(0.50,\ 0.66\right)$. The two-way interactions received a value of

$$\frac{\left(1-k\right)\sum_i \lambda_i}{\left(\varsigma\left(\varsigma-1\right)\right)/2}. \tag{55}$$

This way of allocating the $\sum_i \lambda_i$ led to main effect weights of the same magnitude as the three-way interaction whereby both are slightly-to-somewhat larger than weights for the two-way interactions.

Item Response Generation

Item responses were generated as follows. First, item parameters were simulated as described above. Then, if the probability of correct response based on Eq. (9) given item parameter values were less than a random draw from a *Uniform*(0, 1) distribution,

the item response was considered incorrect. Otherwise, the item response was scored as correct. This approach was also consistent with Willse et al. (2007).

Estimation

DCMs were estimated in the R package CDM (Robitzsch et al., 2013). Marginal maximum likelihood (MML) estimation using the Expectation-Maximization (E-M) algorithm based on de la Torre (2009) and de la Torre (2011) was performed. Numbers of iterations for studying convergence were saved and reported on below (1,000 iterations indicates non-convergence).

Characterization of Mixture Distributional Features

Mixture characteristics will be described as discussed in the section *Characteristics of Mixture Distributions* in the previous chapter, *Review of the Literature*.

Empirical Investigation of Local Independence

Differences between observed item responses and model-based expected item responses were used to compute model residuals, given by

$$d_i = X_{ei} - E_{ei}, \tag{56}$$

Where X_{ei} is the *e*-th examinee's scored response to the *i*-th item (1=correct, 0=incorrect) and E_{ei} is the expected response. Expected item responses will be computed in a similar manner to Henson et al. (2015), given as

$$E_{ei} = \sum_{p=1}^{2^A} \nu_{ep} \cdot \hat{P}\left(X_{ei} = 1 \,|\, \underset{\sim}{\alpha}_{ep}, \hat{\underset{\sim}{\lambda}}_i, \hat{\lambda}_{i0} \right) . \tag{57}$$

Here, $\hat{P}\left(X_{ei} = 1 \,|\, \underset{\sim}{\alpha}_{ep}, \hat{\underset{\sim}{\lambda}}_i, \hat{\lambda}_{i0} \right)$ is the estimated probability of correct response using the estimated model parameters for one given skill profile, $\underset{\sim}{\alpha}_{ep}$, among $p = 1,...,2^A$ possible skill profiles. The ν_{ep} term is the probability that the *e*-th examinee is classified into the *p*-th skill mastery profile. The unique number of simulated response strings will determine (match) the patterns of expected responses.

The *Q*3 statistic (Yen, 1984) was estimated to suggest pairs of test items that showing local dependence (Embretson & Reise, 2000). The *Q*3 statistic represents the correlation between items after partialling out the latent trait(s), calculated by correlating residual scores among item pairs i and i' given by (Yen, 1993)

$$Q_{3ii'} = r\left(d_i, d_{i'} \right) . \tag{58}$$

Parameter Recovery

Parameter recover regarding the LCDM-like aspects of the MCCIRM are summarized using bias as defined in Maris (1999) and mean absolute deviation (MAD) as defined in Junker (2007) as:

$$BIAS = \left(\frac{1}{R} \sum_{r=1}^{R} \hat{\phi}^{(r)} \right) - \phi \tag{59}$$

And

$$MAD = \frac{1}{R}\sum_{r=1}^{R}\left|\hat{\phi}^{(r)} - \phi\right|, \tag{60}$$

Where ϕ denotes the true item parameter value and $\hat{\phi}^{(r)}$ denotes the estimated item parameter value for the *r*-th replication, $r = 1,...,R$. In this study, $R = 100$ replications. Mean estimated values along with true parameter values are reported in accompanying these measures. Scatterplots of estimated versus true item parameter values are examined.

Quantifying Classification Performance

Two metrics of classification performance will be reported: (a) proportion of examinee classification into the correct skill mastery profile (pattern-wise CCR) and (b) proportion of examinee classification into the correct skill mastery state for each skill (attribute-wise CCR). The former will be estimated by dividing the number of examinees whose attribute patterns are correctly identified by the total number of examinees (Feng et al., 2014). These two correct classification rates (CCRs) will be reported marginally over conditions by effect size and model and then also condition-specific to the third decimal place. EAP estimates will be used with a posterior predicted probability (*p*) of >0.50 indicating skill mastery. Feng et al. (2014) remark that MAP estimates provide better classification on all *A* attributes while EAP estimates result in examinees classified correctly on more attributes. Other researchers (e.g., Templin & Henson, 2006; Roussos

et al., 2007a; Jang, 2009a) have also previously considered various "zone of indifference" regions when setting mastery probability cut-points where $p < 0.40$ indicates nonmastery, $p > 0.60$ indicates skill mastery, and $0.40 \leq p \leq 0.60$ where mastery classification is indeterminate. Ackerman et al. (2010) alternatively defined the indeterminate zone of indifference region as $0.45 < p < 0.55$. No indifference region is adopted in the current study.

Presentation of Results

Tables, scatterplots, and trellis boxplots of study measures paneled by conditions and models are used to present the study findings. All graphics subscribe to principles of Cleveland (1994) and Tufte (2001).

How Methodology Addresses Research Questions

Table 10 described how research questions and hypotheses were addressed:

Table 10. Methods for Addressing Study Hypotheses

Hypothesis	**Methods**
*H*1. Increasing variance of continuous abilities generated from the MCCIRM is detected by increasingly large Yen's Q3 statistics using results from the LCDM estimation without continuous ability.	Yen's *Q*3 statistics are summarized by model, effect size, and simulation conditions in tables and using boxplots.
*H*2. Increasing variance of continuous abilities within the MCCIRM substantially degrades item parameter recovery in LCDM estimation without continuous ability (of the LCDM-part item parameters of the MCCIRM).	Bias and MAD are summarized by model, effect size, and simulation conditions in tables, scatterplots, and using boxplots.

Hypothesis	Methods
*H*3. Increasing variance of continuous abilities within the MCCIRM leads to overestimation of attribute-to-attribute correlations in the LCDM without continuous ability.	Attribute-to-attribute correlation estimates are summarized by model, effect size, and simulation conditions in tables and using boxplots.
*H*4. Increasing variance of continuous abilities within the MCCIRM substantially degrades proportion of correct diagnostic classification based on the LCDM without continuous ability.	Correct classification rate (CCR) attribute-wise and profile-wise are summarized by model, effect size, and simulation conditions in tables and using boxplots.
*H*5. Complex versus simple structure substantially degrades proportion of correct diagnostic classification based on the LCDM without continuous ability when variance of continuous abilities within the MCCIRM increases.	Methods used in *H*4 are employed to compare Complex versus Simple structure conditions.
*H*6. Compensatory versus noncompensatory processes leads to substantially lower proportion of correct diagnostic classification based on the LCDM without continuous ability when variance of continuous abilities within the MCCIRM increases.	Methods used in *H*4 are employed to compare Compensatory versus Noncompensatory conditions.

Possible Limitations of the Approach

Limitations of the methods include potential inappropriateness when the assumptions made that are discussed in these first three chapters are violated. Such violations would cast doubt on validity of conclusions, but many are made by study design choice so that those are controlled for. However, other limitations of the approach would be if such assumptions are unreasonable or do not relate to real-world practice. Because the selected study methodology and simulation conditions were grounded in a thorough and current literature review, this strengthens the case for their appropriateness

and of the general study approach in addressing research questions and hypotheses. Where alternative choices, conditions, and assumptions could have been made or investigated in future research this has been explicitly noted previously.

CHAPTER IV

RESULTS

This chapter presents results of the study for convergence, local dependence, correct classification, item parameter recovery, and attribute-to-attribute correlations.

Convergence

Figure 10 gives the number of iterations needed by condition. No replications resulted in the default maximum of 1,000 iterations, seemingly implying convergence from this standpoint. The number of iterations was somewhat less for the simple structure condition relative to replications with average complexity of two, more so for the 10 items per attribute condition. Within complex structure for the 10 items per attribute condition (test length = 30 items), increasing violations of local independence through decreasing effect sizes were associated with increased numbers of iterations, especially for the CRUM and LCDM models. The DINA and DINO models featured this much less so, whereby a jump in iterations was observed for any effect size but differences within the Δ = 3, 2, and 0.8 conditions for these models were more modest. Numbers of iterations for these models in either test length were roughly similar to that within simple structure.

Figure 10. Number of Iterations by Simulation Condition

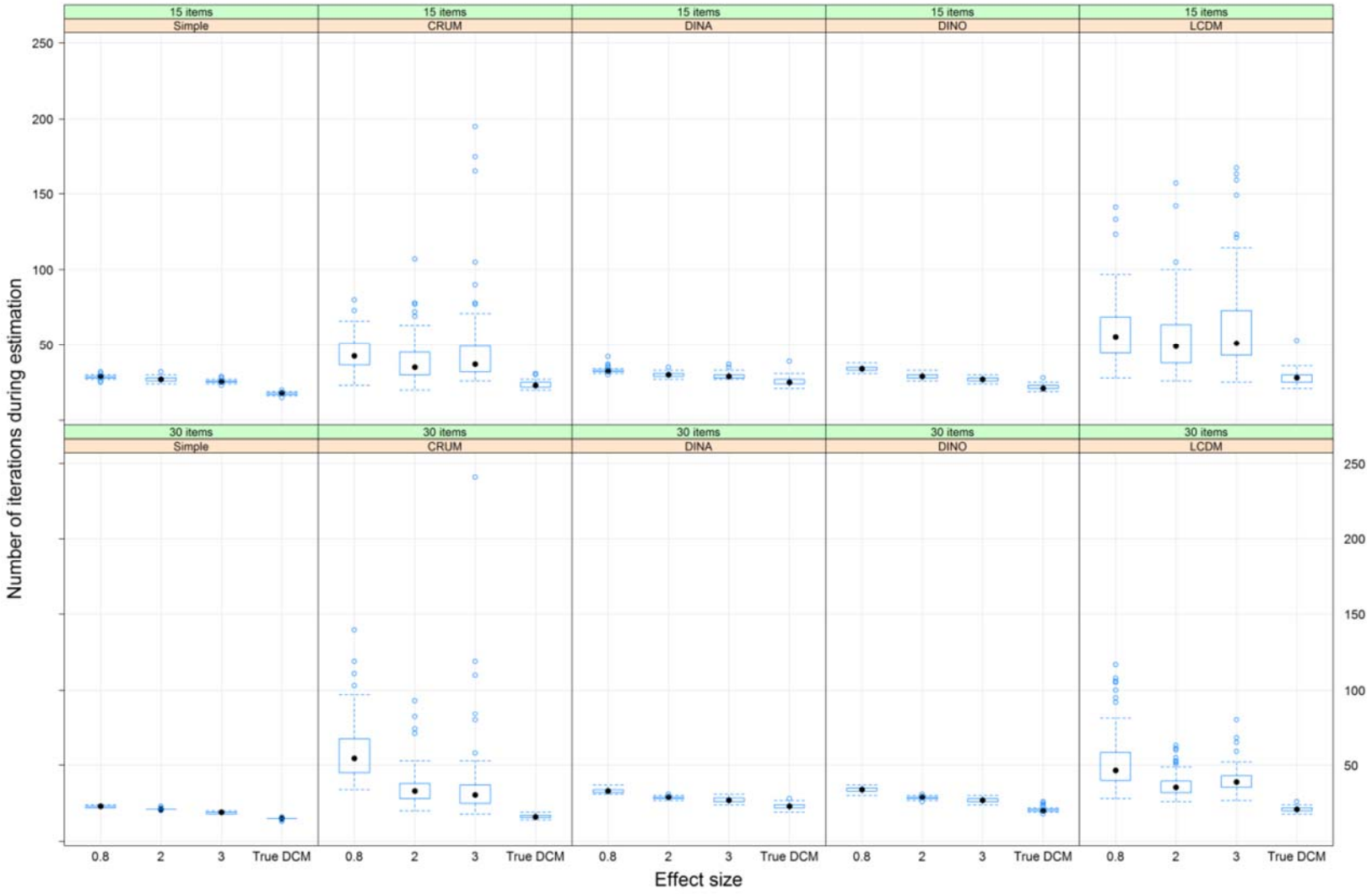


Figure 11. Mean of Yen's *Q3* Statistic

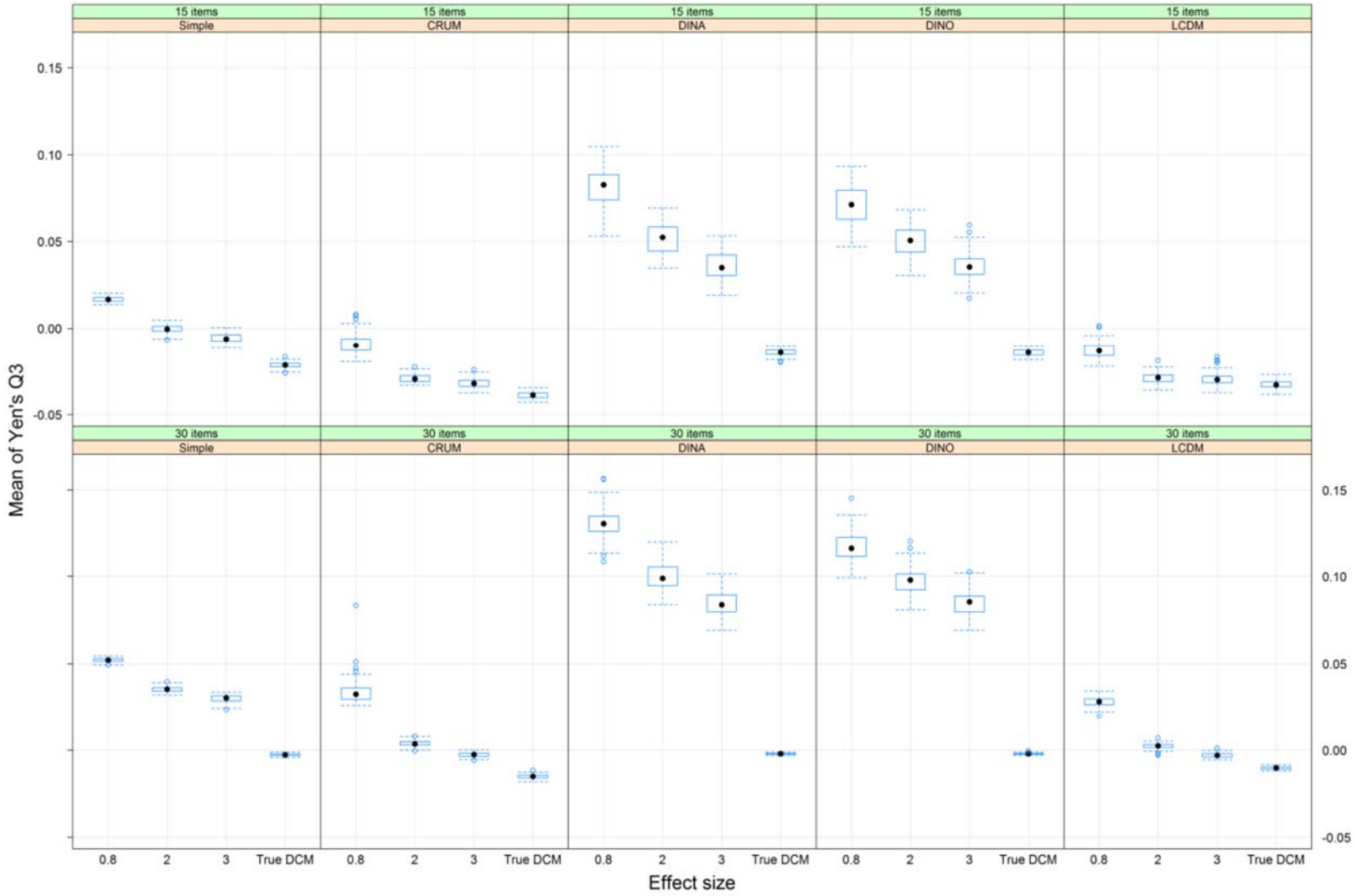


Figure 11 conveys detection of local dependencies by decreasing effect sizes. Also, Figure 12 reports the mean of the absolute value of Yen's *Q*3 statistic. Here for the 30 item condition, decreasing effect size corresponding to greater local independence violation was associated with increasingly larger Yen's *Q*3 values. Interestingly, *Q*3 was much higher in magnitude (more than double for some conditions) for the DINA and DINO models relative to the LCDM and CRUM under complex structure. Results for simple structure were somewhere in between these, but still exhibiting the same pattern of increasing *Q*3 for decreasing effect size.

Also observed in Figure 12 was a non-monotonically increasing pattern specifically under the 15 items test length case for just the CRUM and LCDM (but not in their 30 item test length results). This pattern was an artifact of taking the absolute value of the *Q*3 statistics before averaging, as these patterns for this test length and models exhibited the expected increasing pattern in Figure 11 that did not first take absolute values of *Q*3. For these models in the 15 item test length case this phenomenon was due to many values being slightly negative and close to zero. So when the absolute value is first taken (as in Figure 12), the patterns appear to be misleading with respect to true DCM versus increasing effect sizes because the narrow distribution of negative *Q*3 values for true DCM here become positive. Table 11 reports the average of mean *Q*3 values (and absolute value) by model, effect size, and other conditions:

Figure 12. Mean of Absolute Value of Yen's *Q3* Statistic

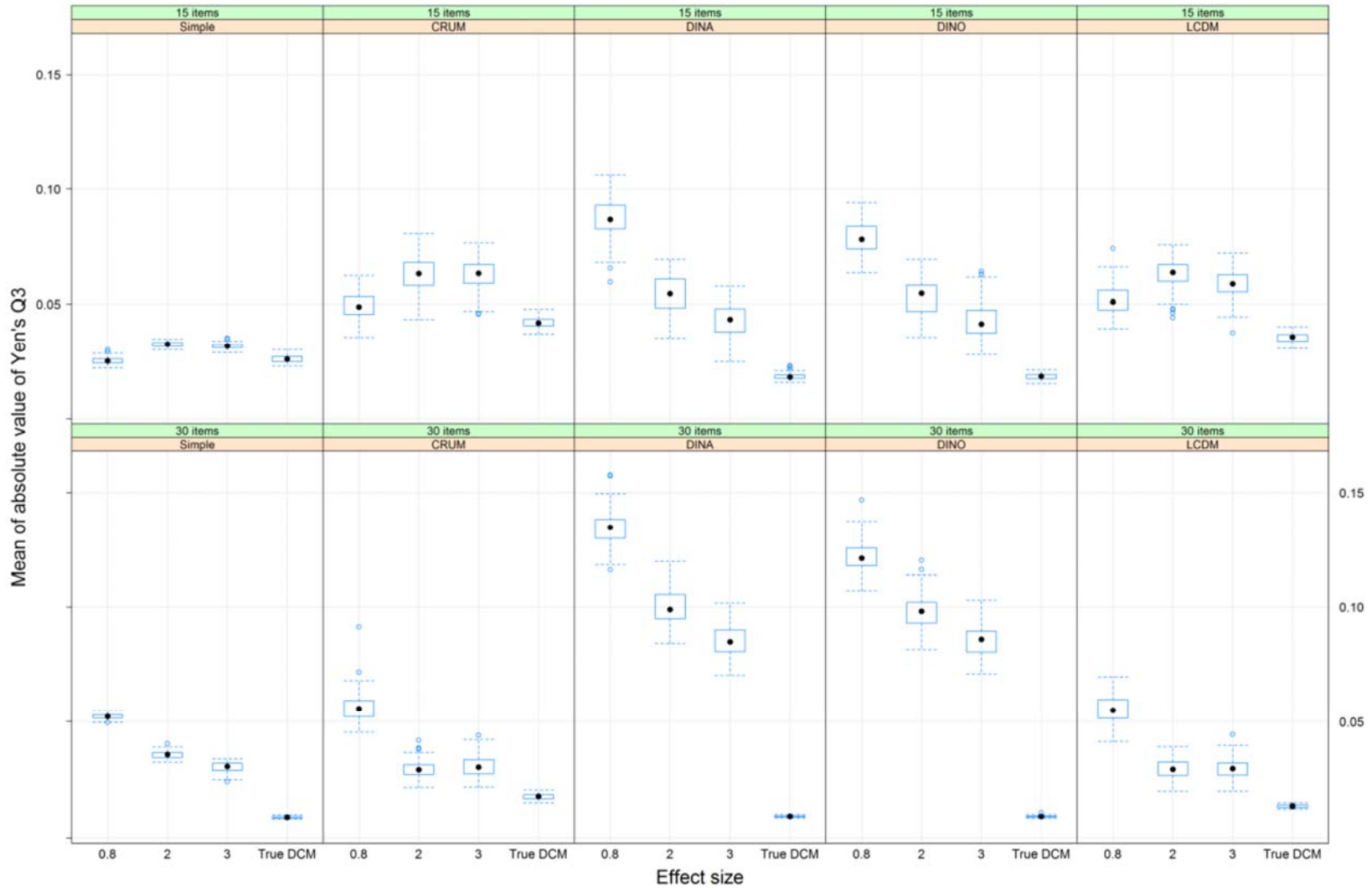


Table 11. Mean Values of Yen's *Q*3 Statistic by Study Conditions

	No. items = 15							
	True DCM		$\Delta = 3$		$\Delta = 2$		$\Delta = 0.8$	
Model	*Q3*	|*Q*3|	*Q3*	|*Q*3|	*Q3*	|*Q*3|	*Q3*	|*Q*3|
Simple	-0.022	0.026	-0.006	0.032	-0.000	0.032	0.017	0.025
CRUM	-0.039	0.042	-0.032	0.063	-0.029	0.063	-0.009	0.049
DINA	-0.014	0.018	0.036	0.043	0.052	0.054	0.082	0.087
DINO	-0.014	0.018	0.036	0.042	0.050	0.043	0.071	0.079
LCDM	-0.033	0.035	-0.030	0.059	-0.029	0.063	-0.013	0.052
	No. items = 30							
	True DCM		$\Delta = 3$		$\Delta = 2$		$\Delta = 0.8$	
Model	*Q3*	|*Q*3|	*Q3*	|*Q*3|	*Q3*	|*Q*3|	*Q3*	|*Q*3|
Simple	-0.003	0.009	0.030	0.030	0.035	0.036	0.052	0.052
CRUM	-0.015	0.018	-0.003	0.030	0.004	0.030	0.034	0.056
DINA	-0.002	0.009	0.085	0.085	0.099	0.100	0.131	0.134
DINO	-0.002	0.009	0.085	0.085	0.097	0.098	0.117	0.122
LCDM	-0.010	0.014	-0.003	0.030	0.002	0.030	0.028	0.055

Interestingly, violations were higher for models that allowed only two levels of probability (DINO and DINO) compared to DCMs that allowed for more than two (CRUM and LCDM) as seen in the above table. Still, *Q*3 increased above and beyond the true DCM condition of no effect size for all models. Thus, it was concluded that Yen's *Q*3 adapted for DCMs as previously described could serve as a potentially useful diagnostic for detecting violations of local independence for diagnostic measurement.

Correct Classification

Overall examinee attribute pattern correct classification rates (CCR) by effect size across other conditions are presented in Table 12.

Table 12. Pattern-wise Correct Classification Rate (CCR) by Effect Size across Other Conditions

	True DCM	$\Delta = 3$	$\Delta = 2$	$\Delta = 0.8$
Mean	0.910	0.591	0.495	0.352
SD	0.040	0.113	0.087	0.027
Min	0.802	0.413	0.358	0.297
Max	0.996	0.818	0.664	0.399
N*	1,000	1,000	1,000	1,000

**Note.* Each one of 1,000 replications had n = 10,000 simulees.

The True DCM condition (no LI violation) had CCR = 0.91 (min = 0.802, max = 0.995), while CCR for effect sizes $\Delta = 3$, 2, and 0.8 were 0.591, 0.495, and 0.352, respectively. However, marginalizing across models, test length, and complexity conditions does not convey the full range of findings. Figure 13 provides condition-specific CCRs.

Again, correct classification was high and around expected magnitude for the True DCM (no effect size) condition for both simple and complex structure. For simple structure, CCRs eroded about the same magnitude for either test length as effect size decreased to $\Delta = 2$ (CCR=0.621 for 15 item length and 0.652 for 30) and then to $\Delta = 0.8$ (0.375 for 15 items and 0.387 for 30 items). Under complex structure, the pattern of how CCR changed according to the specific DCM as effect size decreased is compelling. Specifically, the degradation of CCR was greater for the CRUM and LCDM (comparable

to each other) relative to that observed for the DINA and the DINO models (again comparable to each other). Here, CCR was approximately 0.51 on average for the DINA and DINO models when $\Delta = 2$ but decreased to 0.39 on average for the CRUM as well as for the LCDM. As Δ decreased to 0.8, the average CCR for DINA = 0.37 and for DINO = 0.36 while for the CRUM and LCDM were both at 0.31 on average. These above rates are based on 30 items, whereby results under 15 items were similar but with greater variation observed in CCRs.

While these findings indicate impact on pattern-wise correct classification, marginal attribute-specific correct classification was also examined. Table 13 describes the overall attribute-specific CCR by effect size across other conditions.

Figure 13. Condition-specific Attribute Pattern-wise Correct Classification Rates

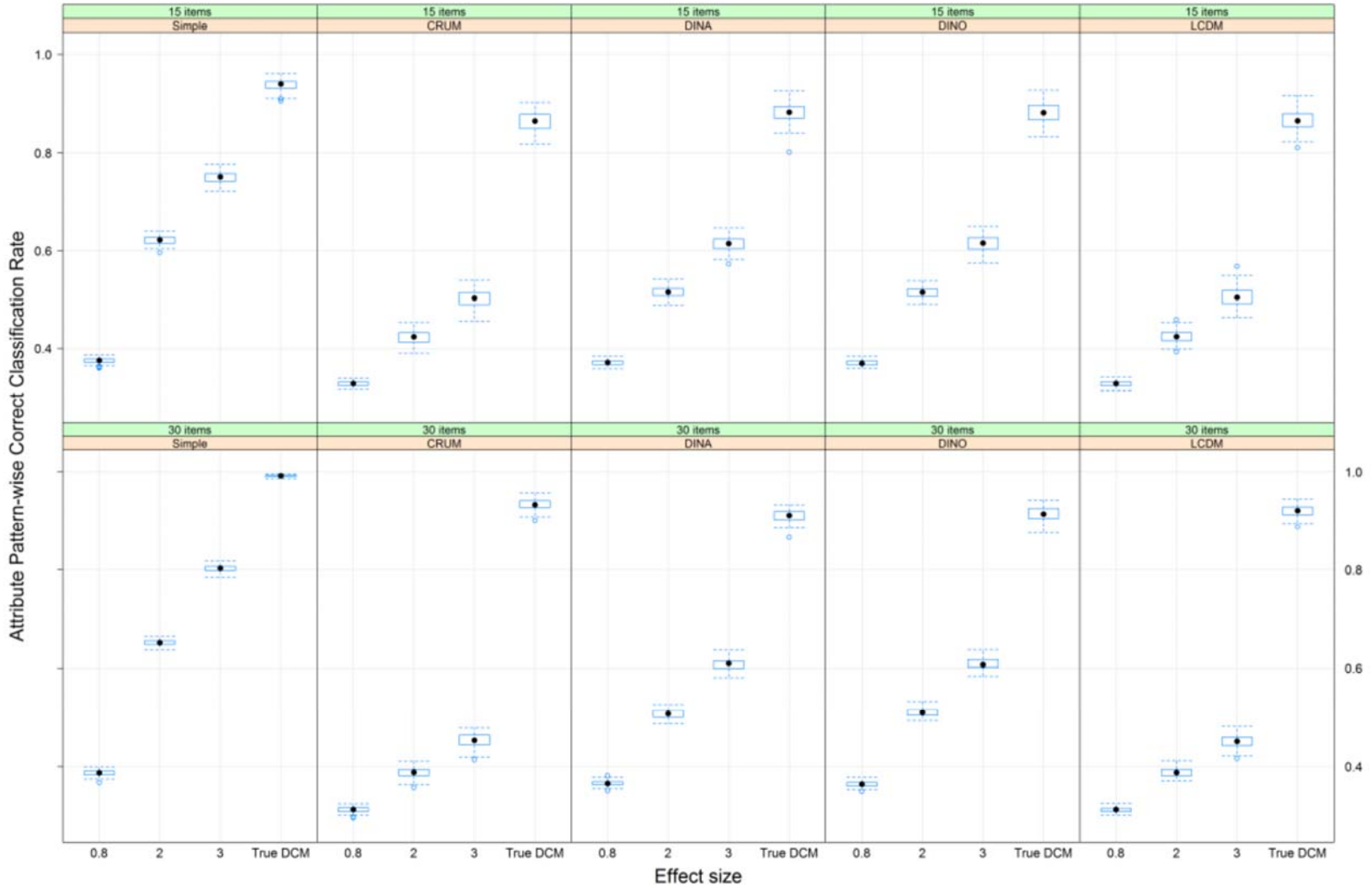


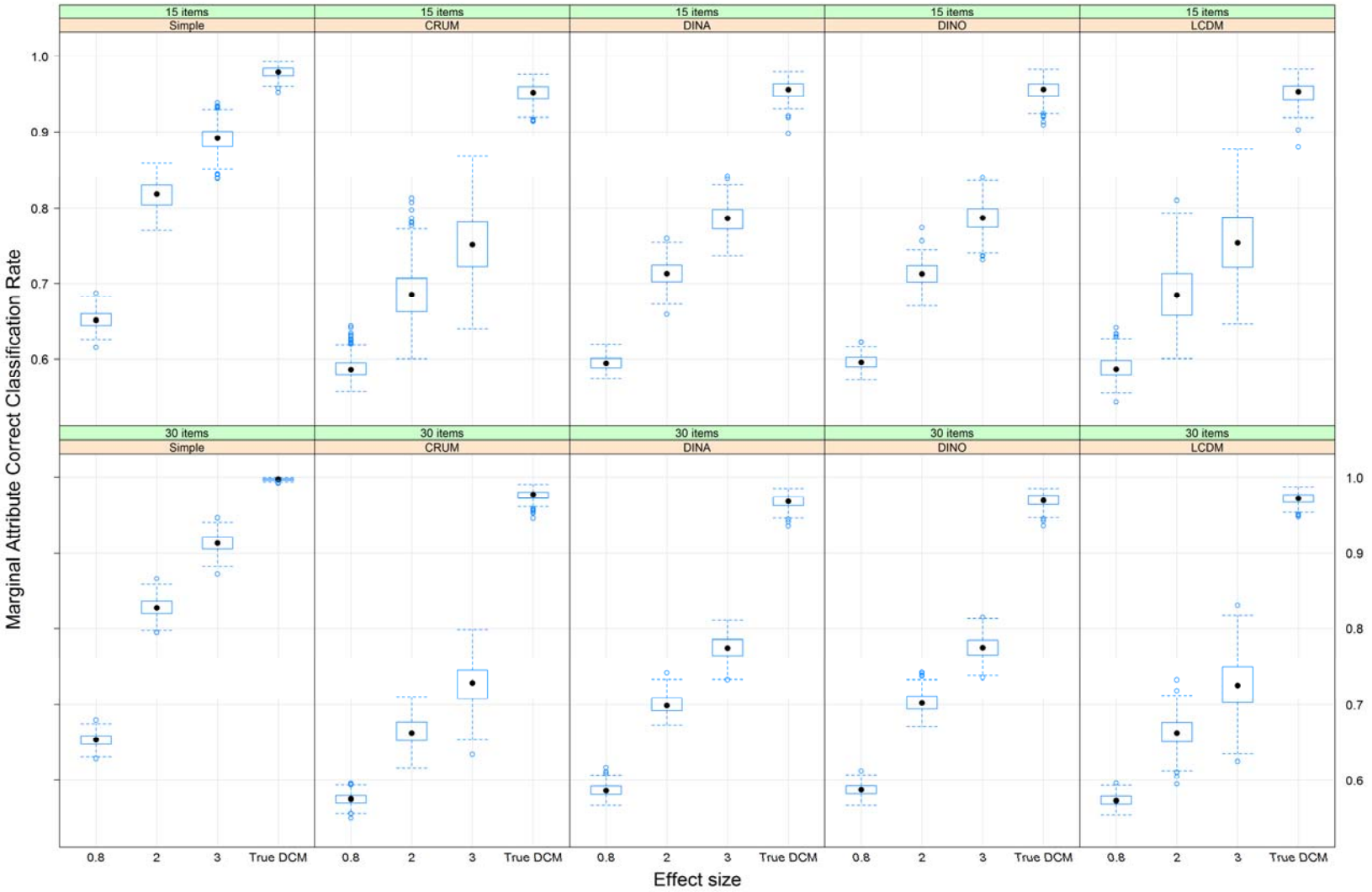
Table 13. Attribute-wise Correct Classification Rate (CCR) by Effect Size across Other Conditions

	True DCM	$\Delta = 3$	$\Delta = 2$	$\Delta = 0.8$
Mean	0.967	0.789	0.718	0.600
SD	0.017	0.066	0.059	0.029
Min	0.880	0.625	0.595	0.544
Max	1.000	0.946	0.867	0.687
N*	3,000	3,000	3,000	3,000

**Note.* Each one of 1,000 replications had n = 10,000 simulees times 3 attributes each.

The True DCM condition (no LI violation) had CCR = 0.967 (min = 0.880), while CCR for effect sizes $\Delta = 3$, 2, and 0.8 were 0.789, 0.718, and 0.600, respectively. Thus, only 60% of examinees were classified correctly on any given skill when the LI violation was large ($\Delta = 0.8$), compared to 96.7% with no violation. However, marginalizing across models, test length, and complexity conditions again does not convey the full range of findings. Figure 14 provides the attribute-specific CCRs by study condition.

Figure 14. Condition-specific Attribute-wise Correct Classification Rates



The findings mirror those for pattern-wise correct classification. Simple structure cases were less impacted by decreasing effect sizes relative to complex structure. Interestingly, the degradation of CCR was not as great for the CRUM and LCDM (comparable to each other) relative to that for the DINA and the DINO models (again comparable to each other). Here, attribute-wise CCR was approximately 0.70 on average for the DINA and DINO models when $\Delta = 2$ but decreased to 0.66 on average for the CRUM and LCDM. As Δ decreased to 0.8, the average CCR for the DINA and DINO models was 0.59 while for the CRUM and LCDM were both at 0.57 on average (based on 30 items). Thus, examining Figure 14 above the difference between the models appears to be at $\Delta = 3$ and 2 but when Δ decreases to 0.8 all models perform similarly poor in classifying examinees on individual attributes.

Correct classification was further examined by grouping examinees into three categories: those with skill patterns of (0, 0, 0), those with skill patterns of (1, 1, 1), and everyone else. The following overall cross-tabulation describes the CCR according to effect size and category:

Table 14. Correct Classification Rate According to Effect Size and Skill Pattern Category

Estimated	*True*		
True DCM	**(0,0,0)**	**All others**	**(1,1,1)**
(0,0,0)	2,967,584 (95.8)	243,592 (6.4)	116 (<0.1)
All others	130,370 (4.2)	3,378,951 (88.9)	92,894 (3.0)
(1,1,1)	137 (<0.1)	178,388 (4.7)	3,007,968 (97.0)
Δ = 3	**(0,0,0)**	**All others**	**(1,1,1)**
(0,0,0)	2,131,660 (68.8)	808,487 (21.3)	276,811 (8.9)
All others	677,504 (21.9)	2,247,472 (59.1)	670,042 (21.6)
(1,1,1)	288,927 (9.3)	744,972 (19.6)	2,154,125 (69.5)
Δ = 2	**(0,0,0)**	**All others**	**(1,1,1)**
(0,0,0)	1,870,126 (60.4)	1,002,752 (26.4)	467,252 (15.1)
All others	744,394 (24.0)	1,838,382 (48.4)	742,027 (23.9)
(1,1,1)	483,571 (15.6)	959,797 (25.3)	1,891,699 (61.0)
Δ = 0.8	**(0,0,0)**	**All others**	**(1,1,1)**
(0,0,0)	1,484,616 (47.9)	1,295,499 (34.1)	831,130 (26.8)
All others	773,937 (25.0)	1,229,258 (32.3)	776,393 (25.0)
(1,1,1)	839,538 (27.1)	1,276,174 (33.6)	1,493,455 (48.2)

**Note*. Numbers reported are frequency and column percentage. Rates are based on 1,000 replications each with 10,000 simulees.

Thus, as effect size decreases correct classifications obviously degenerate. Interestingly, as effect sizes get smaller the off-diagonal of the above contingency table that increases the most is the misclassification from (0,0,0) to all other patterns besides (1,1,1). Here, this column percentage goes from 6.4% for true DCM to 21.3% for $\Delta = 3$, 26.4% for $\Delta = 2$, to 34.1% for $\Delta = 0.8$. Similar increases in misclassification from (1,1,1) to all others are observed (4.7% for true DCM increasing to 33.6% for $\Delta = 0.8$).

Item Parameter Recovery

Bias in item parameter recovery was examined and the following scatterplots convey estimated intercepts versus true for test length of 15 and 30 items:

Figure 15. Scatterplot of True Value of the Intercept, λ_0, versus Estimated Value for 15 Items

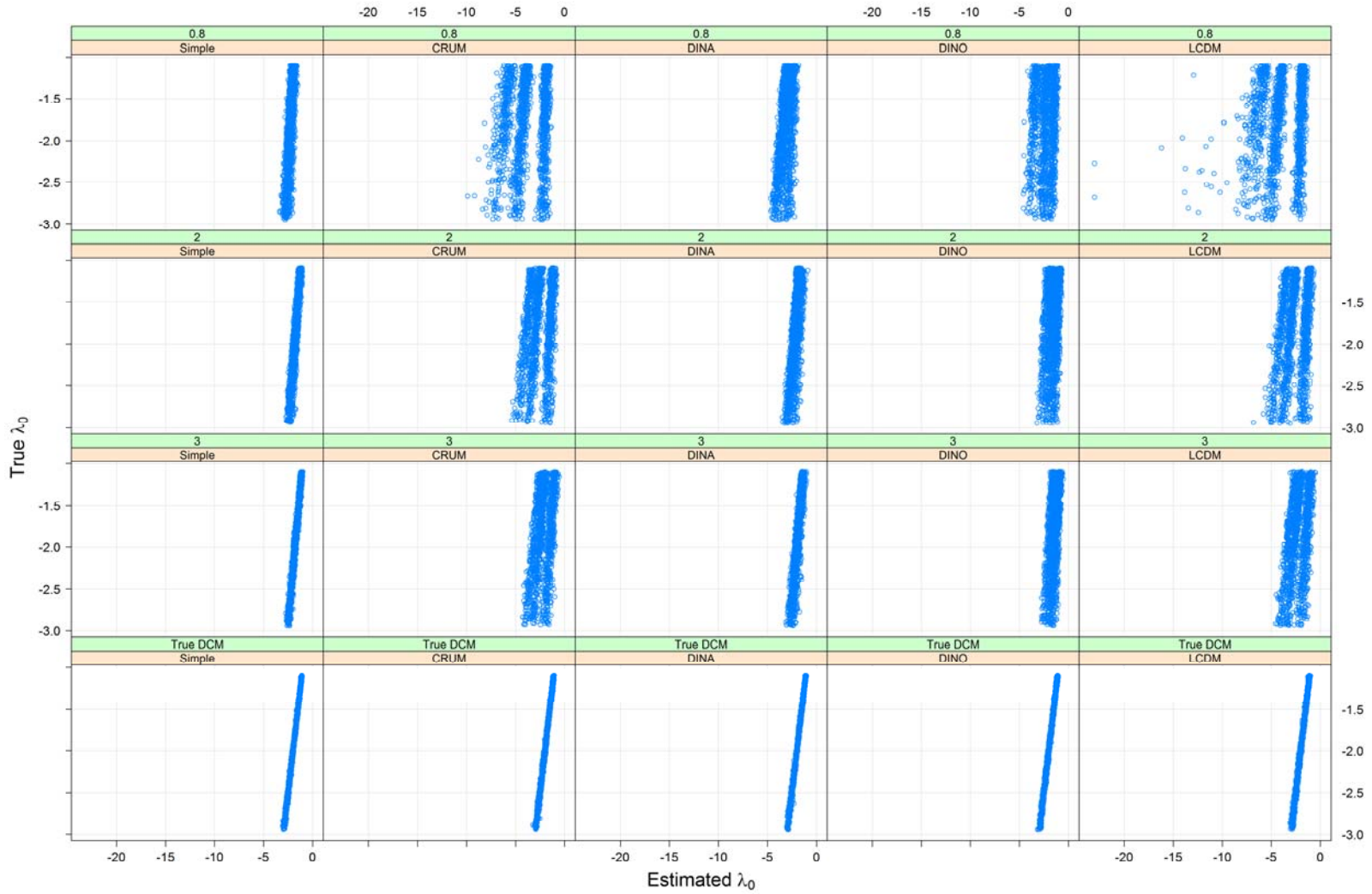
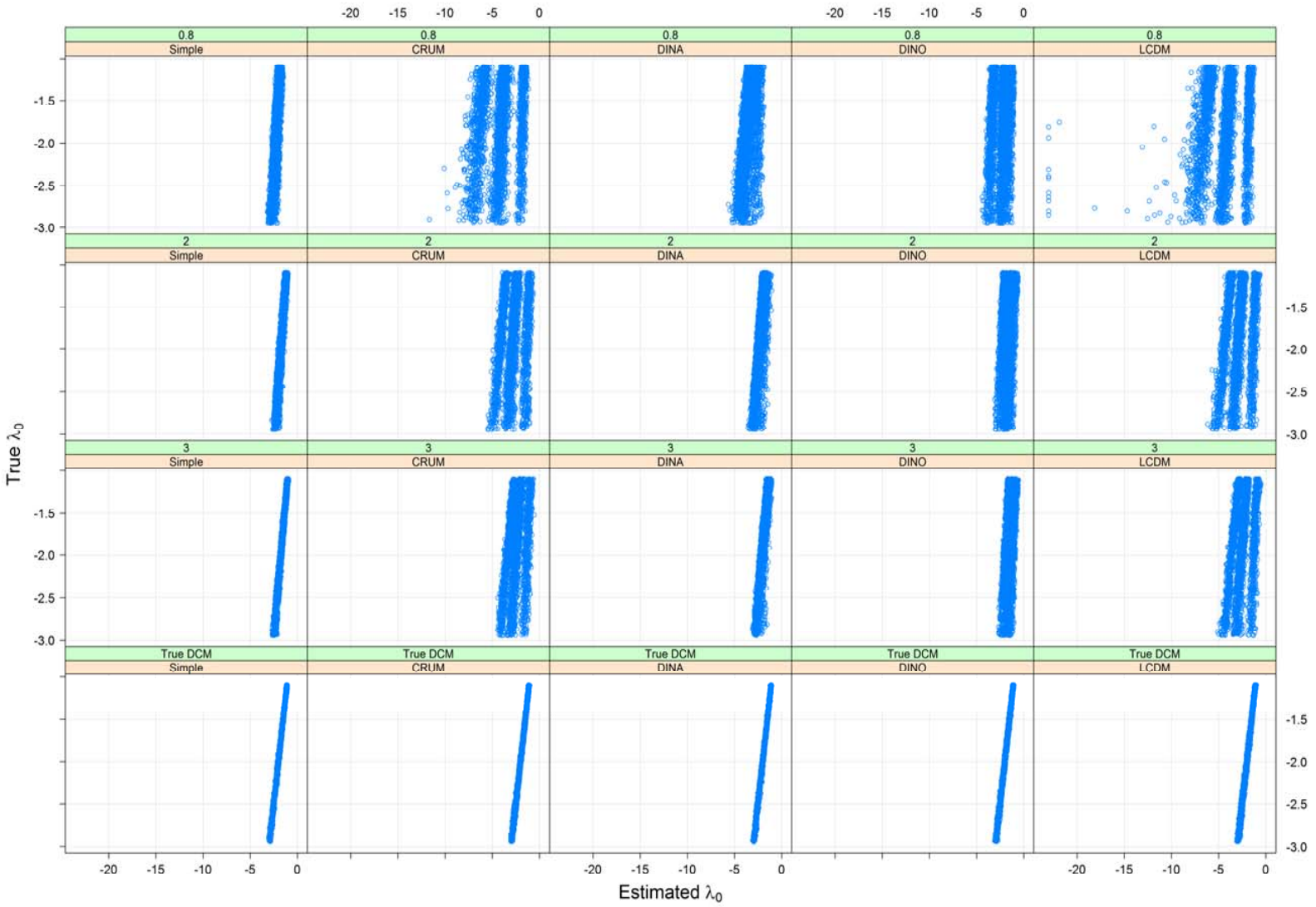


Figure 16. Scatterplot of True Value of the Intercept, λ_0, versus Estimated Value for 30 Items



For simple structure, as effect size decreases the variability in estimated logits of correct response for complete nonmasters become more variable. This is greatly more so for complex structure, where even for an effect size of $\Delta = 3$ intercepts diverge non-trivially from true value for the CRUM and LCDM. By the time Δ decreases to 0.8 (large LI violation), estimation of the intercepts for the CRUM and LCDM are suspect. It is noteworthy that while the same phenomena were observed for the DINO and DINA models, the impact on estimated values appeared to be much less so compared to the CRUM and LCDM which allow for >2 probability levels. This is also be directly observed by examining magnitude of differences using bias and MAD measures which are described next.

Figure 17. Bias in Parameter Recovery of the Intercept, λ_0

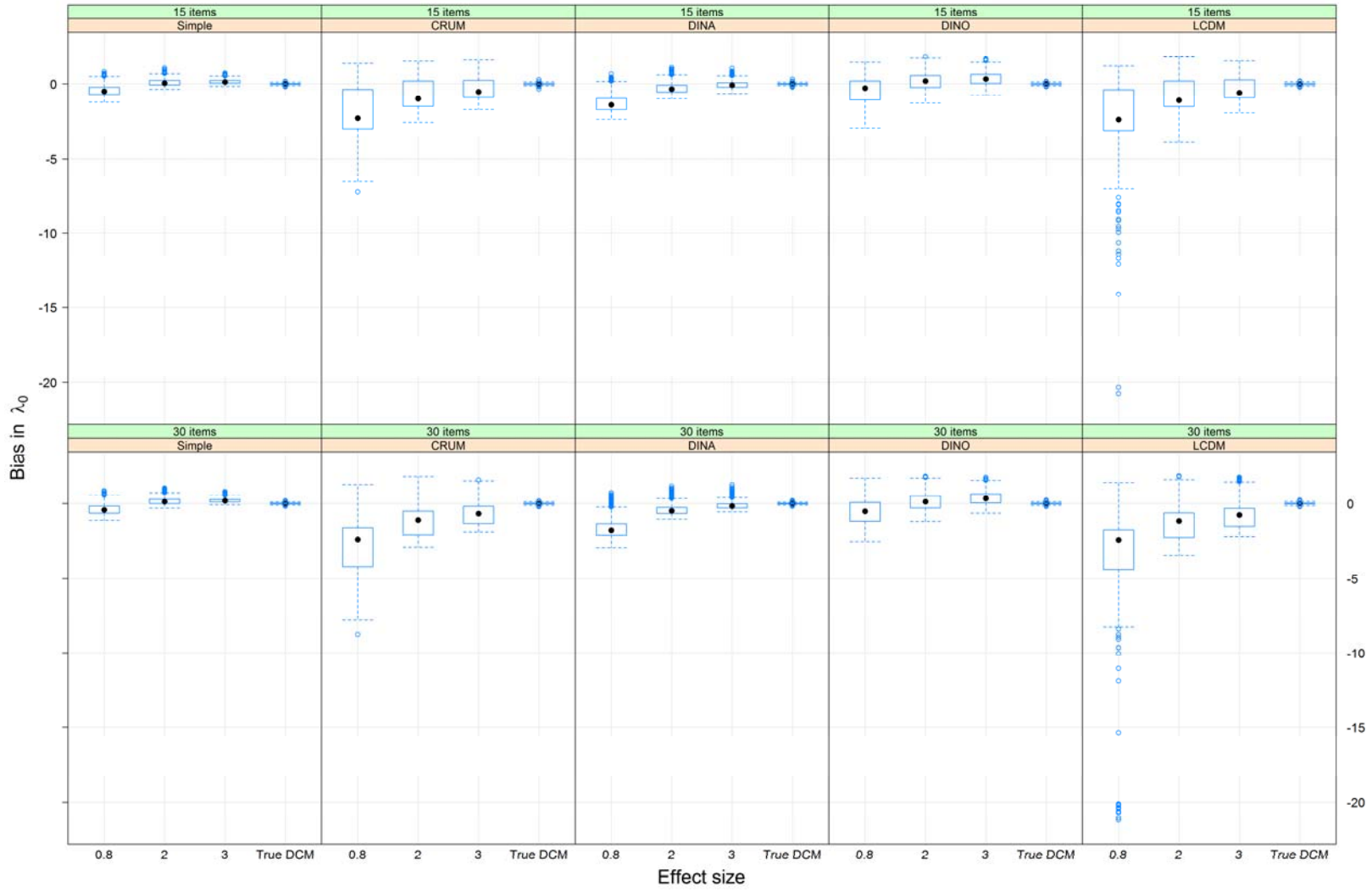
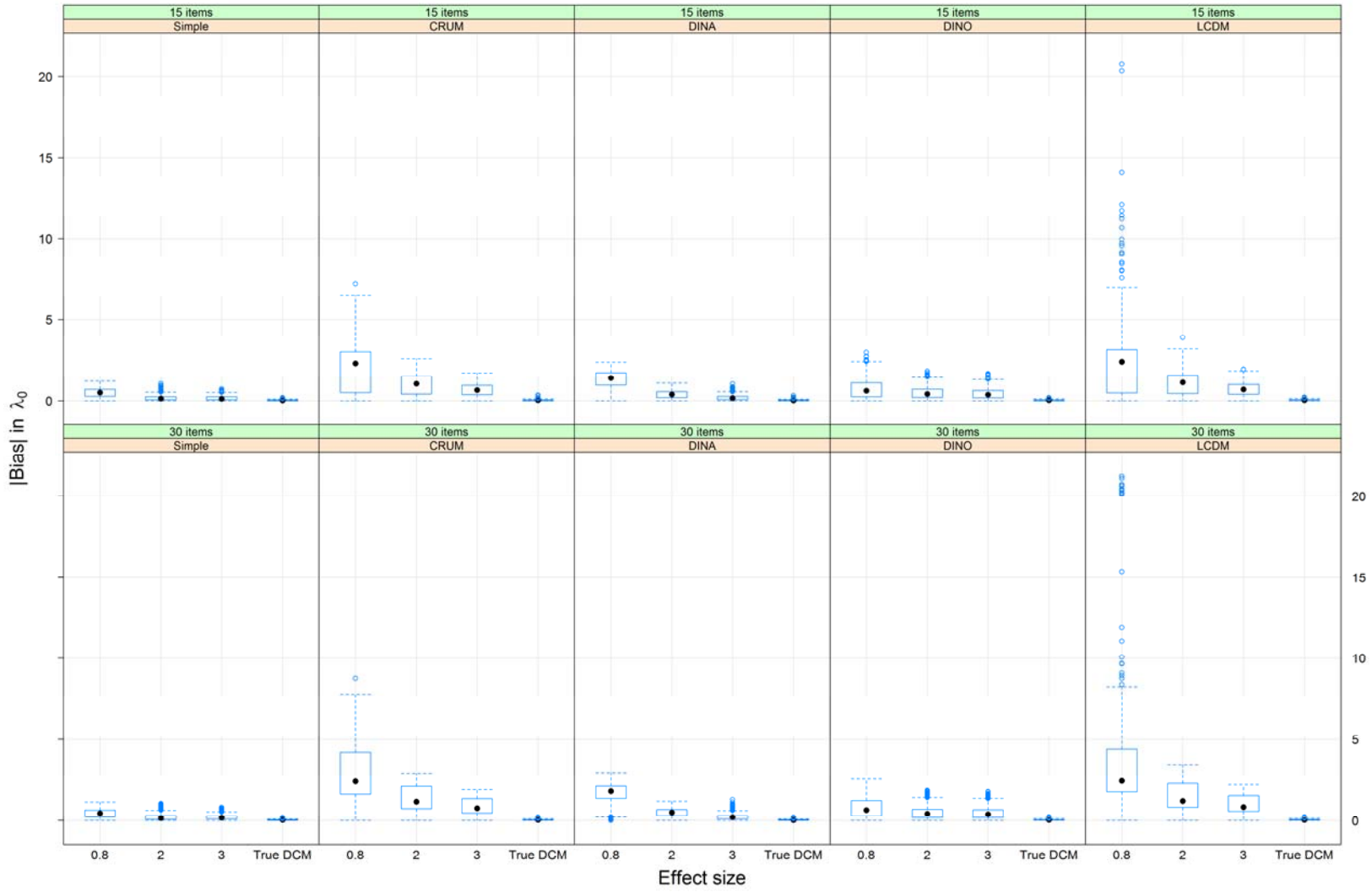


Figure 18. MAD in Parameter Recovery of the Intercept, λ_0

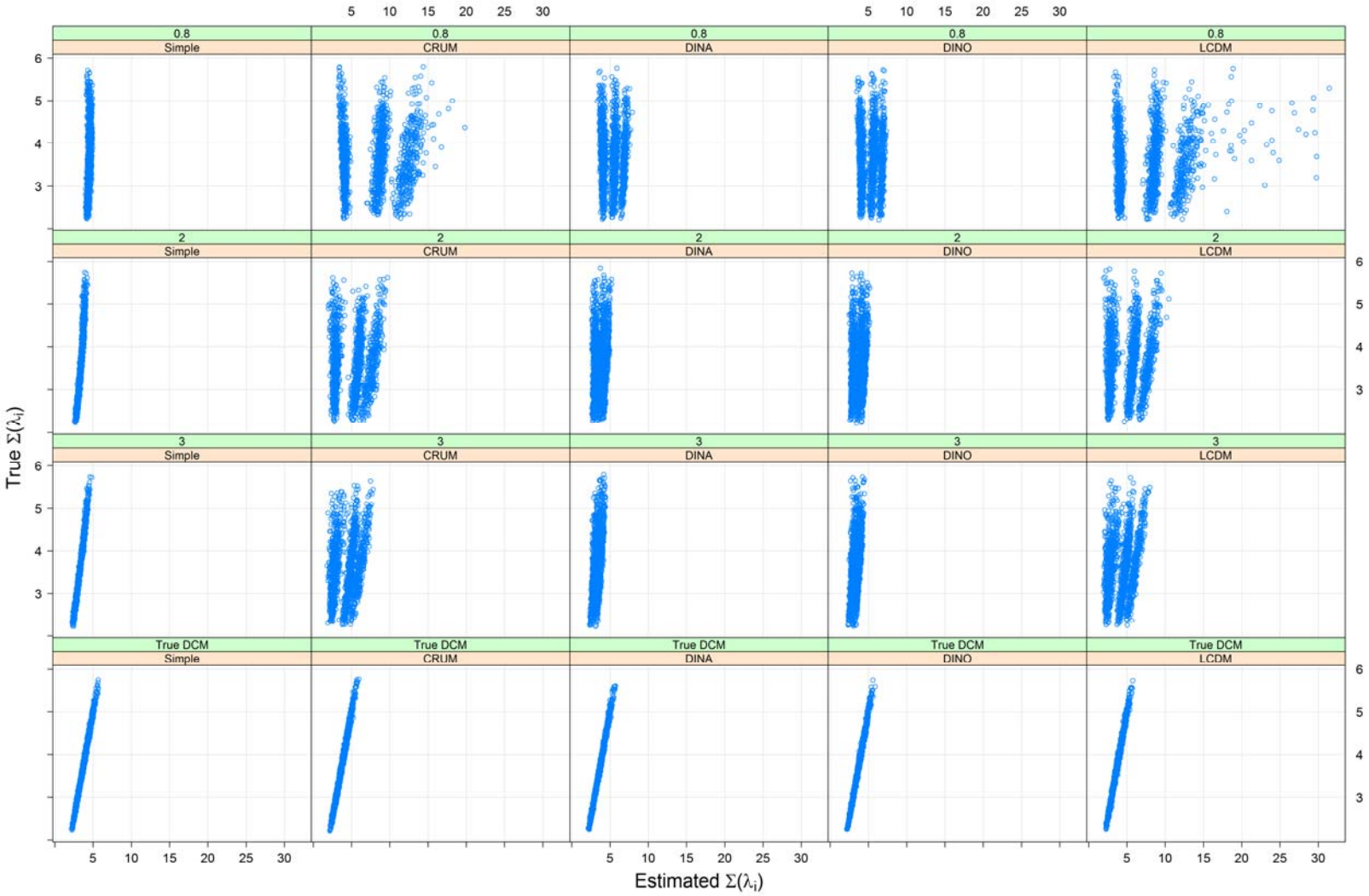


From the above, the true DCM (no effect size/no LI violation) was unbiased for the intercept across complexity, model, and test length, as expected (also see Table 14). For simple structure, impact on the intercept was markedly less than that relative to models in complex structure, although both bias and MAD increased as effect size decreased as anticipated. The boxplots for both bias and MAD demonstrate how much more estimated intercepts were impacted for the LCDM and CRUM compared to the DINA and DINO models. The distribution of bias and MAD became noticeably skewed for the smallest effect size of $\Delta = 0.8$ for the LCDM in particular (e.g., MAD = 2.3 for 15 items and 2.8 for 30 items). This is quantified further in the following table of average values of bias and MAD of the intercept by study conditions.

Table 15. Bias and MAD of the Intercept, λ_0, by Study Conditions

	No. items = 15							
	True DCM		$\Delta = 3$		$\Delta = 2$		$\Delta = 0.8$	
Model	Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD
Simple	-0.001	0.036	0.154	0.166	0.096	0.186	-0.450	0.501
CRUM	-0.004	0.040	-0.358	0.682	-0.746	1.057	-2.020	2.121
DINA	-0.003	0.035	-0.047	0.195	-0.289	0.389	-1.309	1.317
DINO	-0.002	0.041	0.344	0.443	0.165	0.490	-0.470	0.781
LCDM	-0.005	0.045	-0.402	0.738	-0.800	1.128	-2.209	2.292
	No. items = 30							
	True DCM		$\Delta = 3$		$\Delta = 2$		$\Delta = 0.8$	
Model	Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD
Simple	-0.001	0.033	0.202	0.203	0.169	0.210	-0.382	0.446
CRUM	-0.001	0.037	-0.613	0.854	-1.065	1.299	-2.523	2.591
DINA	-0.002	0.032	-0.129	0.218	-0.427	0.485	-1.673	1.680
DINO	0.001	0.039	0.369	0.463	0.128	0.472	-0.562	0.802
LCDM	-0.001	0.044	-0.713	0.968	-1.163	1.383	-2.686	2.759

Figure 19. Scatterplot of True Value of Sum of Weights versus Estimated Value for 15 Items



Because LI violation was introduced in particular through increasing variance according to the effect size in Eq. (51), parameter recovery for the sum of LCDM weights above the intercept was also examined using similar methods. The following scatterplots relate estimated sums of weights versus true sums for test length of 15 and 30 items:

For simple structure, as effect size decreases the variability in estimated increments in the logit of correct response for complete masters appears to attenuate. For complex structure, a great deal of more variability in sums is observed, where even for an effect size of $\Delta = 3$ the sums diverge non-trivially from true value for the CRUM and LCDM. As Δ decreases to 0.8 (large LI violation), estimation of the sums for the CRUM and LCDM is degraded. Interestingly, this pattern for the DINO and DINA models also was reproduced but to a much lesser degree. These relationships are further examined regarding their magnitude of differences using bias and MAD measures described next.

Figure 20. Scatterplot of True Value of Sum of Weights versus Estimated Value for 30 Items

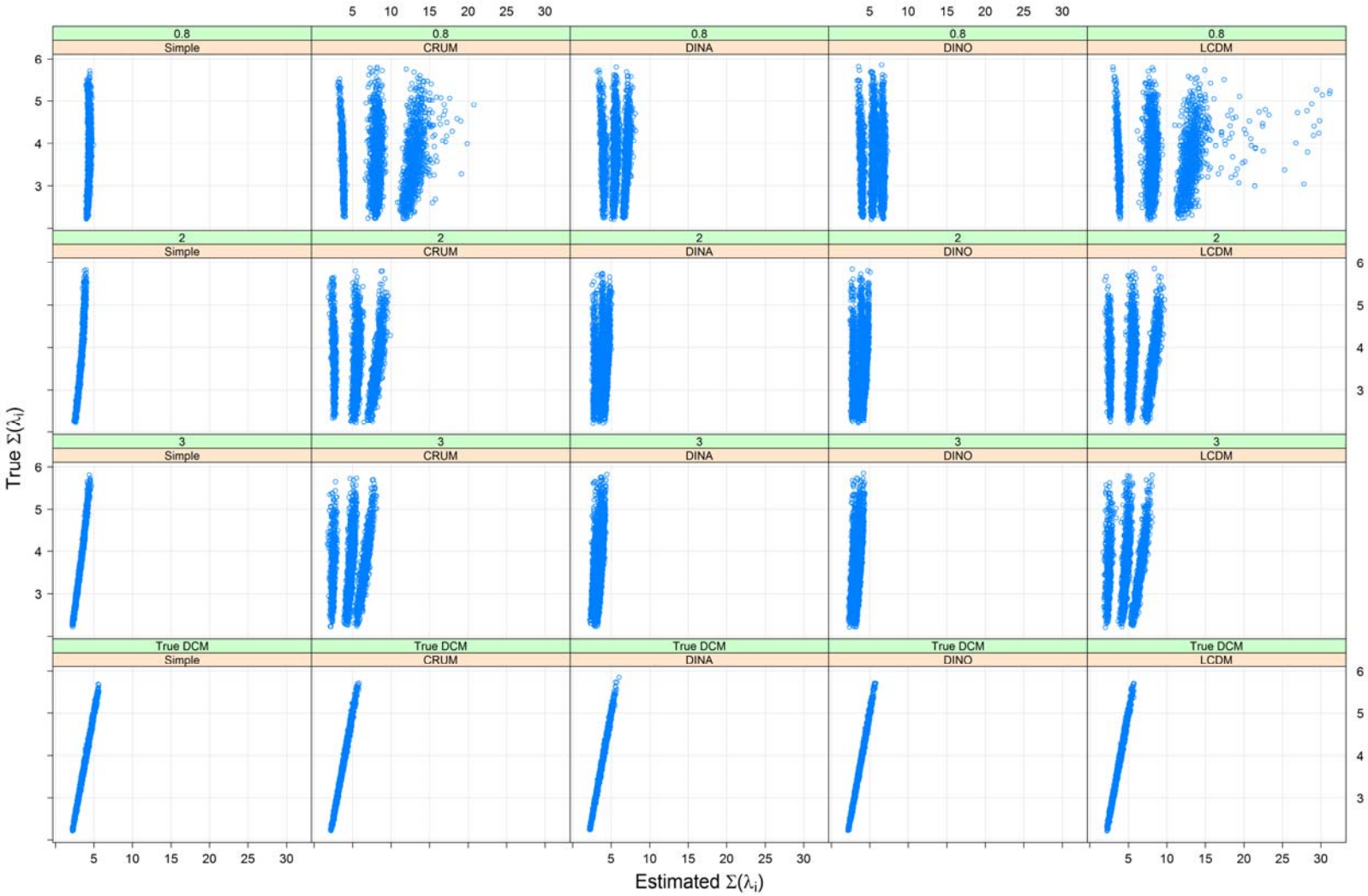


Figure 21. Bias in Parameter Recovery of Sum of Weights above λ_0

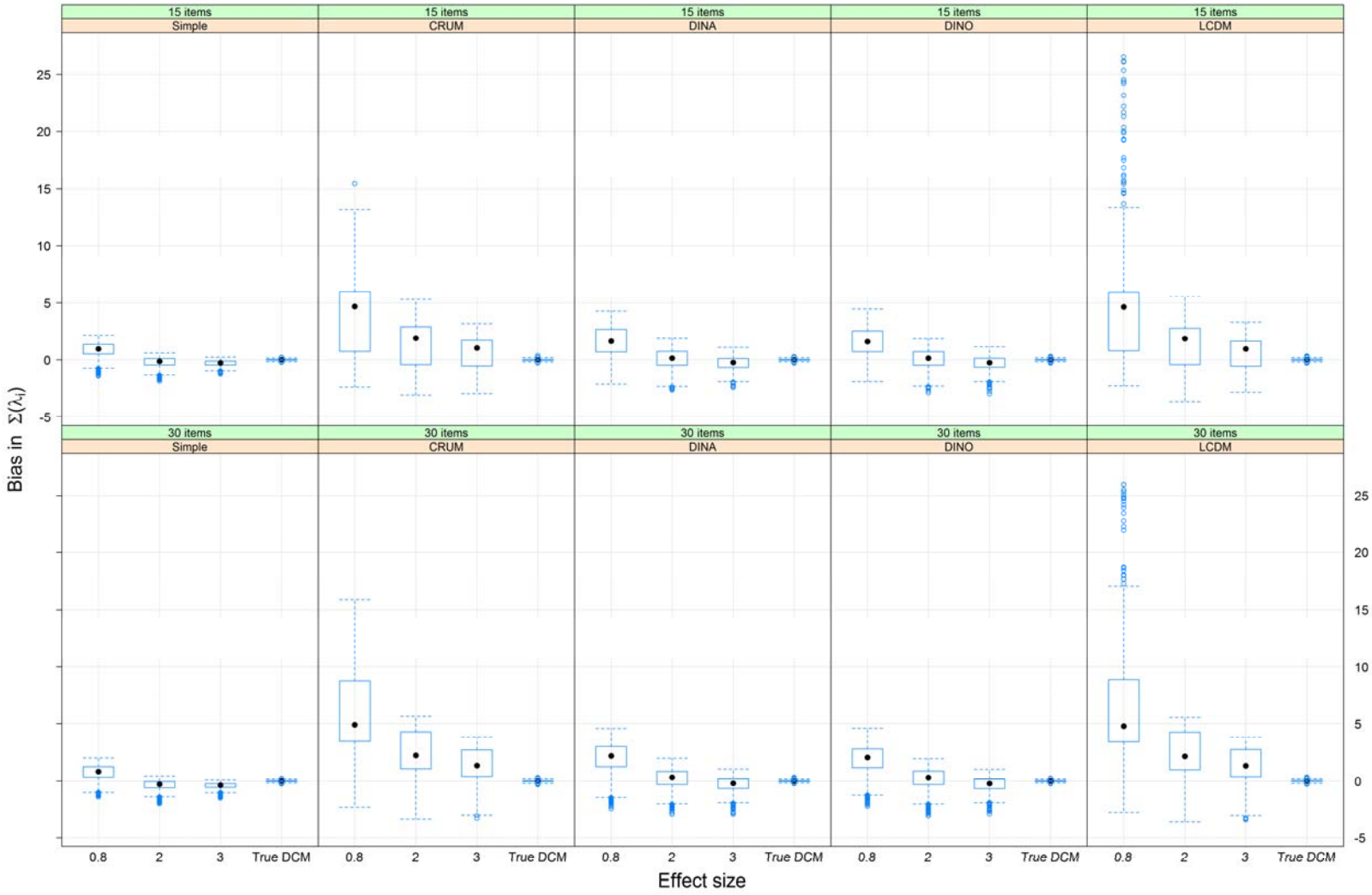


Figure 22. MAD in Parameter Recovery of Sum of Weights above λ_0

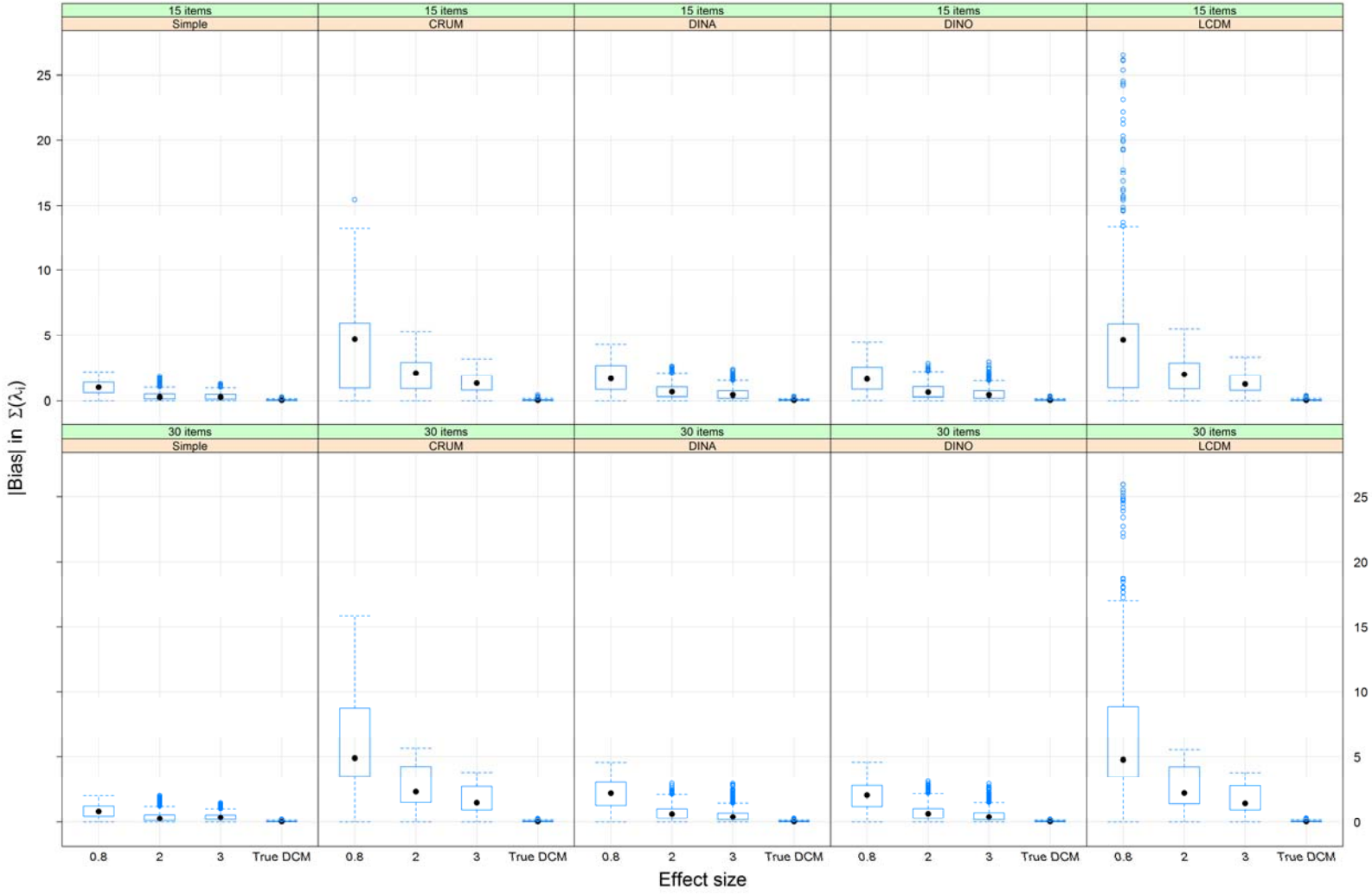


Table 16. Bias and MAD of Sum of Weights above λ_0 by Study Conditions

	No. items = 15							
	True DCM		$\Delta = 3$		$\Delta = 2$		$\Delta = 0.8$	
Model	Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD
Simple	0.001	0.052	-0.301	0.315	-0.207	0.356	0.895	0.980
CRUM	0.003	0.063	0.709	1.357	1.471	2.101	4.103	4.267
DINA	0.002	0.056	-0.306	0.520	0.077	0.715	1.641	1.781
DINO	0.004	0.055	-0.289	0.514	0.090	0.697	1.623	1.747
LCDM	0.005	0.062	0.674	1.331	1.454	2.093	4.354	4.492
	No. items = 30							
	True DCM		$\Delta = 3$		$\Delta = 2$		$\Delta = 0.8$	
Model	Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD
Simple	0.001	0.048	-0.409	0.410	-0.345	0.407	0.744	0.861
CRUM	0.001	0.058	1.249	1.717	2.171	2.630	5.220	5.343
DINA	0.004	0.051	-0.265	0.525	0.236	0.713	2.069	2.172
DINO	0.000	0.051	-0.281	0.530	0.231	0.720	1.951	2.032
LCDM	0.001	0.059	1.239	1.725	2.156	2.564	5.320	5.438

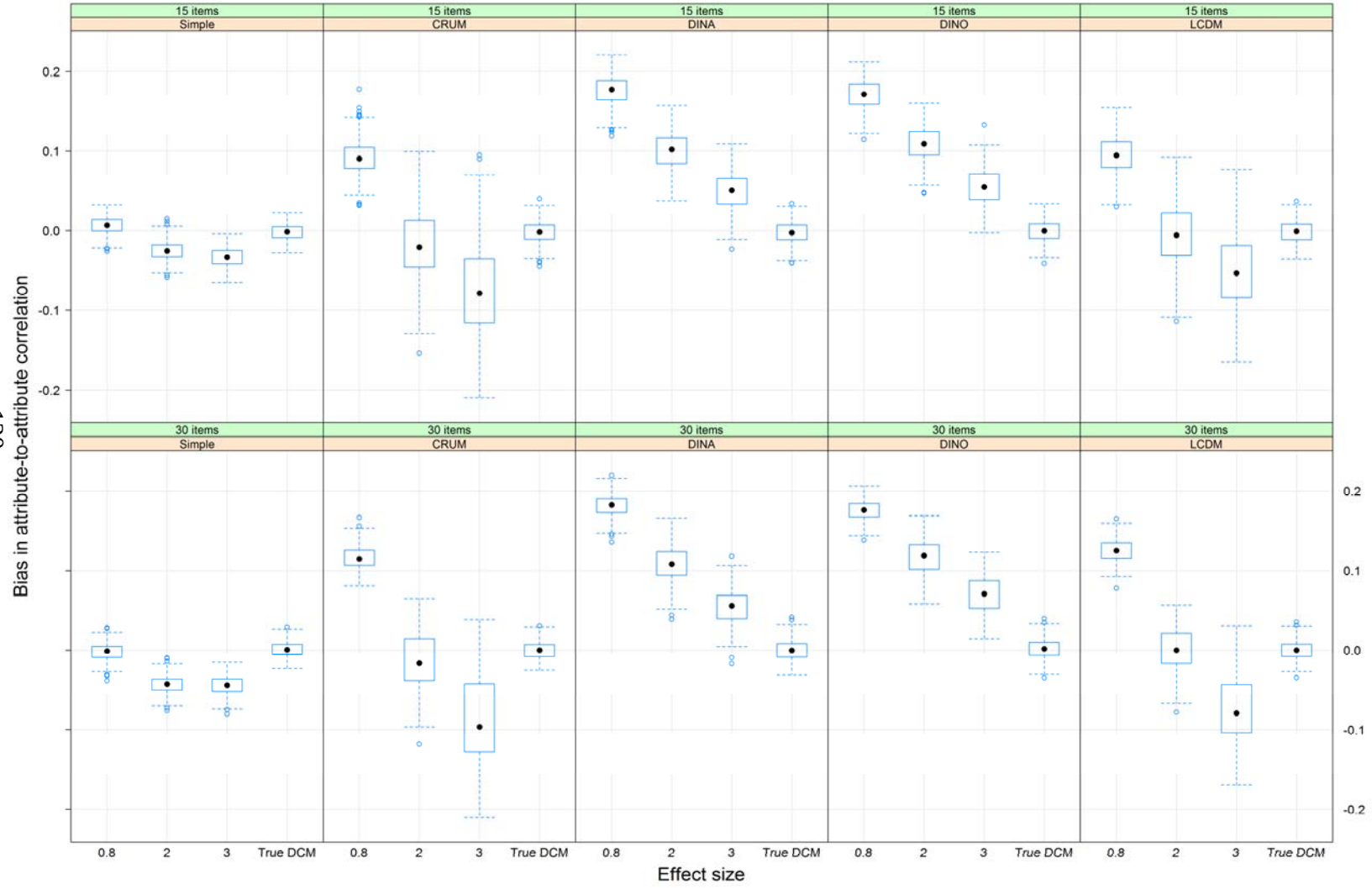
It should be noted again that the effect size was introduced using the average of the sum of weights across items for a given replication while parameter recovery is studied for individual items. Again the true DCM (no effect size) was unbiased for the sum of weights above the intercept across complexity, model, and test length, as expected. For simple structure, impact on sums was notably less than that relative to models in complex structure (e.g., MAD almost half), although both bias and MAD still increased as effect size decreased as anticipated. The boxplots for both bias and MAD again convey increased impacted for the LCDM and CRUM compared to the DINA and DINO models. The distribution of bias and MAD for the sum of weights excluding the intercept was again markedly skewed for the smallest effect size of $\Delta = 0.8$ for the LCDM in particular (e.g., MAD = 4.5 for 15 items and 5.4 for 30 items).

Association among Attributes

Figure 23 provides the comparison of estimated attribute-to-attribute associations relative to the true correlation value of 0.70.

Results from analysis of simple structure items demonstrated lower discrepancies with the true attribute correlation relative to complex structure. For complex structure, as violation of local independence increase (i.e, decreasing effect size), positive bias incrementally increased for the DINA and DINO models. Interestingly, a different pattern was observed for the CRUM and LCDM. Here, differences with true correlation were similar on average for $\Delta = 2$ relative to true DCM (although more variable), but only when $\Delta = 0.8$ did the positive bias in estimated attribute-to-attribute correlations start to increase in magnitude to a greater degree.

Figure 23. Difference between Estimated Attribute-to-Attribute Correlations from True 0.70



CHAPTER V

CONCLUSIONS

Evidence was found that nonzero variance of companion continuous latent traits causes detectable violations of the assumption of local independence in diagnostic measurement, and also degrades ability to recover DCM parameters and perform classification correctly. This finding is consistent with Hansen (2013), although it is interesting how much more impact was observed for the saturated LCDM which was only examined in the current study. The LCDM allows for more than two probability levels to be analyzed (e.g., unlike the DINA) so that overlapping distributions from partial masters appeared to cause further degradation in findings. This was consistent with the other studied model that possesses multiple probability levels, the CRUM.

The current study findings with respect to Yen's $Q3$ suggest several implications for practitioners, as it was found that increasing continuous ability variance generated from the MCCIRM lead to larger values of $Q3$. The mean values (without taking absolute value first) presented in Table 11 in Results provide a basis for preliminary guidelines of empirical evidence indicating local independence violation in DM. That is, because $Q3$ as assessed in this study can be computed when performing diagnostic classification, practitioners can calculate this statistic first and inspect its magnitude before moving on to the interpretation phase of the DM investigation. This evidence should be considered specific to the DCM being implemented, and also potentially

further differentiated by test length. When a diagnostic assessment has 30 items, a Yen's $Q3 > 0.03$ could indicate possible local dependence for the CRUM and LCDM under an average complexity of two. For the same test length and complexity under the DINA and DINO, a Yen's $Q3 > 0.10$ could be indicative of possible local dependence. For an assessment with 15 items, a positive $Q3$ (value > 0) for the LCDM or CRUM assuming average complexity of two could indicate LI violation (as well as for simple structure items). The threshold for the DINA and DINO models is suggested to be raised to $Q3 > 0.06$ for indicating problem levels under similar conditions. A caveat here is that 15 items may be too few to study up to three attributes under average complexity of two for the CRUM or LCDM under the particular LI violations examined in this study. Future studies should examine these models further under similar or even smaller test lengths to increase understanding of patterns in performance. It should be noted that these very preliminary guidelines are somewhat conservative as they denote rules of thumb according to observed $Q3$ values approaching those for the most extreme effect size of $\Delta = 0.8$. Another issue with recommending one-size-fits-all guidelines is that they can suffer from similar pitfalls as those noted by Kline (2011) in his discussion of Hu and Bentler's (1999) recommendations for fit indices in structural equation modeling. In the same way, these guidelines should not be generalized to all DCM situations and extensive future research as noted below could provide refinement of advice for practitioners.

Findings on degeneration of item parameter recovery and diagnostic classification were as expected in the current study, although differences by model and complexity conditions were noted. That is, increasing continuous ability variance introduced through

the MCCIRM lead to substantial poor performance in both item parameter recovery and accuracy of performing diagnostic classification correctly. Overestimation of attribute-to-attribute correlations were also found as continuous ability variance increased, especially for the DINA and DINO models under complex structure (while the correlations under simple structure were relatively impacted less). Results of heightened negative impact for the increasingly parameterized, more flexible DCMs of the CRUM and LCDM are in alignment with the conceptual expectations outlined in Compensatory Processes for Complex Structure described in the Review of the Literature. This could be due in part that with the DINO and DINA DCMs, only one additional weight besides the intercept is estimated. However, in the CRUM and LCDM under complex structure the effective effect size was even more extreme, due to the allowed presence of more than two mastery mixtures even though ability variance was introduced through a simple structure effect size specification. Thus, correct classification was degraded under complex structure and under compensatory processes, and was even more impacted when both of these conditions were examined in conjunction. Triangulating with conclusions regarding *Q*3, the CRUM and LCDM had lower observed *Q*3 values for more extreme LI violation relative to DINA or DINO models, yet were impacted to a greater degree on parameter recovery and especially correct classification under complex structure. Together this underlines the suggestion one *Q*3 guideline for all DCMs should be avoided and motivate model-specific preliminary guidelines for practice.

From the Introduction and Appendix A, it was hypothesized that effect sizes would be effectively smaller under complex structure relative to simple structure if Δ was

introduced into the data through simple structure calculations only and contrasting complete skill masters with nonmasters. This was applied for all items within a given replication, and note that in the complex structure condition that target average item-to-attribute complexity = 2 that some items possessed simple structure. Thus, as the ratio of simple structure items within complex structure condition decreases then so does the effective magnitude of Δ, because of the additional contribution due to correlated abilities. Further, it would be interesting to see how findings improve as correlation among abilities decreases (e.g., $r(\theta, \theta') = 0.35$ instead of 0.70 in this study). Future research could further delineate this relationship on impact.

In the same way additional effect sizes of other magnitudes could be considered for future study (e.g., $\Delta = 0.2, 1.4, 5$). The chosen effect sizes of $\Delta = 3$ corresponded to an overlap coefficient (OVL) of 13.4% (i.e., adjacent Normal distributions overlapped 31.7%), $\Delta = 2$ had OVL = 31.7%, and $\Delta = 0.8$ had OVL = 68.9%. This overlap essentially implies that some skill nonmasters have higher probability of correct item response than masters, which should inhibit ability to distinguish masters of skills from nonmasters. It was indeed observed that results exhibited poor performance for these effect sizes in an increasing fashion. However, the current study only examined the case where each attribute had a single companion continuous latent ability. While Hansen (2013) focused on testlet effects with one, two, and four abilities, future research could consider variants of either the current study or Hansen (2013) and examine if LI violations are extreme for say only some particular attributes. Another possibility for

future study is to introduce ability variance through a different mechanism (e.g., one ability per skill profile), as discussed in Introducing Systematic Within-Class Variation from Different Mechanisms in Review of the Literature.

Another feature of the current study was only examining CMIRT aspects of continuous ability contributions of the particular generating model chosen, the MCCIRM. This generating model choice played a role in impact on outcomes. For example, as delineated in Appendix A, correlated abilities effectively decreased the effect size constructed under simple structure when there more than one ability required for complex structure. However, the original full MCCIRM of Henson et al. (2014) which includes interactions of continuous abilities (akin to the saturated LCDM for attributes) could be studied in future research. Chalmers and Flora (2014) reported that product MIRT models with such interactions provided some similar results to truly NCMIRT models where the entire item response functions are multiplied (but some reservations were given). Hong et al. (2015) have proposed a multiplicative hybrid DINA-NCMIRT diagnostic model as well. Thus, performance in unified models of a noncompensatory nature could be studied in the future as well. Finally, the current study only considered particular choices about Q-matrix design, assigning weights to various DCMs (especially the saturated LCDM) and did not examine all DCMs in extant literature (e.g., NIDO, Reduced RUM). Future investigations could study these models and conditions and their effects on findings.

In sum, increasing continuous ability variance was observed to result in more estimation effort, less ability to recover DCM parameters, lower proportion of attribute

pattern and attribute-specific correct classification, and over-estimation of attribute-to-attribute associations. Further, if such increasing variance is conceptualized as violation of LI, then detection through Yen's *Q*3 statistic appears to be possible and practitioners can use the preliminary guidelines suggested above until future research suggests refinements. Jurich (2014) and Hansen (2013) both report on potential usefulness of other limited information fit statistics for such purposes for DCMs besides *Q*3. Finally, patterns of findings appeared to be condition-specific, such that differential effects were observed for the various DCMs under study and according to simple versus complex structure. Validity of use and interpretation is threatened when measurement assumptions are violated. The local independence assumption for DCM was increasingly untenable when variance of continuous abilities was increasingly introduced. Results from this study suggest estimation and classification are besmirched under such a situation, and therefore estimable unified model approaches such as those proposed by Houts and Cai (2013) and Hong et al. (2015) are potentially promising to consider. Future research can help explicate the relationship among the salient issues further and further promote validity after diagnostic measurement.

REFERENCES

AERA, APA, NCME. (2014). Standards for Educational and Psychological Testing. Washington, DC: AERA.

Ackerman, T., Henson, R., Luecht, R., Templin, J., & Willse, J. (2010, October). Applying Computer Based Assessment Using Cognitive Diagnostic Modeling to Benchmark Tests. Presentation at the Tenth Annual Assessment Conference of the Maryland Assessment Research Center for Education Success, College Park, Maryland.

Ackerman, T. A., & Henson, R. A. (2014). Graphical Representations of Items and Tests that are Measuring Multiple Abilities. (In press).

Ackerman, T., Zhang, W., Henson, R., & Templin, J. (2006, April). Evaluating a third graded science benchmark test using a skills assessment model: Q-matrix evaluation. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.) Second International Symposium on Information Theory (pp. 267-281). Akademiai Kiado, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716-723.

Almond, R., & Shute, V. (2009, April). Calibration of Bayesian Network-Based Diagnostic Assessment. Paper presented in the session, Software for Calibrating Diagnostic Classification Models: An Overview of the Current State-of-the-Art, at the annual meeting of the American Educational Research Association, San Diego, CA.

Birnbaum, A. (1957). Efficient design and use of mental ability for various decision-making problems (Series Report 58-16). Randolph Air Force Base, TX: USAF School of Aviation Medicine.

Birnbaum, A. (1958). On the estimation of mental ability (Series Report No. 15). Randolph Air Force Base, TX: USAF School of Aviation Medicine.

Black, P., & William. D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy, and Practice, 5(1), 7-74.

Bozard, J. L. (2008). Invariance Testing in Diagnostic Classification models. Unpublished master's thesis, Department of Educational Psychology, University of Georgia, Athens, GA.

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing Teacher's Understandings of Rational Numbers: Building a Multidimensional Test Within the Diagnostic Classification Framework. Educational Measurement: Issues and Practice, 33(1), 2-14.

Burnham, K. P., & Anderson, D. R. (2010). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd ed. New York: Springer.

Chalmers, R. P., & Flora, D. B. (2014). Maximum-Likelihood Estimation of Noncompensatory IRT Models With the MH-RM Algorithm. Applied Psychological Measurement, 38(5), 339-358.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and Absolute Fit Evaluation in Cognitive Diagnosis Modeling. Journal of Educational Measurement, 50(2), 123-140.

Chiu, C.-Y. (2013). Statistical Refinement of the Q-Matrix in Cognitive Diagnosis. Applied Psychological Measurement, 37(8), 598-618.

Chiu, C.-Y., & Douglas, J. (2013). A Nonparametric Approach to Cognitive Diagnosis by Proximity to Ideal Response Patterns. Journal of Classification, 30(2), 225-250.

Choi, H.-J. (2010). A Model That Combines Diagnostic Classification Assessment With Mixture Item Response Theory Models. Unpublished doctoral dissertation, The University of Georgia, Athens, Georgia.

Cleveland, W. S. (1994). The Elements of Graphing Data. Hobart.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences, 2nd ed. New Jersey: LEA.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. Journal of Educational Measurement, 42, 133-148.

Dawid, A. P. (1979). Conditional Independence in Statistical Theory. JRSS Series B (Methodological), 41(1), 1-31.

Dawid, A. P. (1980). Conditional Independence for Statistical Operations. The Annals of Statistics, 8(3), 598-617.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. Psychometrika, 69(3), 333-353.

de la Torre, J. (2008a). 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 and Douglas (2008b). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. Psychometrika, 73(4), 595–624.

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

de la Torre, J. (2011). The Generalized DINA Model Framework. Psychometrika, 76(2), 179-199.

de la Torre, J., & Lee, Y.-S. (2013) Evaluating the Wald Test for Item-Level Comparison of Saturated and Reduced Models in Cognitive Diagnosis. Journal of Educational Measurement, 50(4), 355-373.

DeMars, C. E., & Lau, A. (2011). Differential Item Functioning Detection With Latent Classes: How Accurately Can We Detect Who is Responding Differentially? Educational and Psychological Measurement, 71(4), 597-616.

Djouadi, A., Snorrason, O., & Garber, F. D. (1990). The Quality of Training-Sample Estimates of the Bhattacharyya Coefficient. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(1), 92-97.

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, NJ: LEA.

Feng, Y., Habing, B. T., & Huebner, A. (2014). Parameter Estimation of the Reduced RUM Using the EM Algorithm. Applied Psychological Measurement, 38(2), 137-150.

Ferrier, D. E., Lovett, B. J., & Jordan, A. H. (2011). Construct-Irrelevant Variance in Achievement Test Scores: A Social Cognitive Perspective. In L. E. Madsen (Ed.), Achievement tests: Types, interpretations, and uses (pp. 89-108). New York: Nova.

Finch, W. H., French, B. F., & Immekus, J. C. (2014). Applied Psychometrics Using SAS. Charlotte, NC: IAP.

Frick, H., Strobl, C., Leisch, F., & Zeileis, A. (2012). Flexible Rasch Mixture Models with Package psychomix. Journal of Statistical Software, 48(7), 1-24.

Haberman, S. J. (1974). Log-linear models for frequency tables derived by indirect observation: Maximum likelihood equations. Annals of Statistics, 2, 911-924.

Haberman, S. J. (1979). Qualitative data analysis (Vols. 1 & 2). New York, NY: Academic Press.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. Journal of Educational Measurement, 26(4), 301-321.

Hagenaars, J. A. (1993). Loglinear Models with Latent Variables. Newbury Park, CA: SAGE.

Hambleton, R. K., van der Linden, W. J., Wells, C. S. (2010). IRT Models for the Analysis of Polytomously Scored Data: Brief and Selected History of Model Building Advances. In M. L. Nering & R. Ostini (Eds.). Handbook of Polytomous Item Response Theory Models (pp. 21-42). New York, NY: Taylor & Francis.

Hancock, G. R., & Samuelsen, K. M. (Eds.) (2008). Advances in Latent Variable Mixture Models. Charlotte, NC: Information Age Publishing.

Hansen, M. P. (2013). Hierarchical Item Response Models for Cognitive Diagnosis. Unpublished doctoral dissertation, University of California Los Angeles, Los Angeles, California.

Hanson, J.-I., Tippins, N., Wise, L., Drasgow, F., & Sackett, P. (2010). Revision of the Standards for Educational and Psychological Testing. Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

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

Henning, G. (1989). Meanings and implications of the principle of local independence. Language Testing, 6(1), 95-108.

Henson, R. A. (2008). “Functions of Estimating Log-Linear Cognitive Diagnostic Model”, Department of Educational Research Methodology, The University of North Carolina at Greensboro, Greensboro, NC.

Henson, R. A. (2009). Diagnostic Classification Models: Thoughts and Future Directions. Measurement: Interdisciplinary Research and Perspectives, 7(1), 34-36.

Henson, R., & Douglas, J. (2005) Test Construction for Cognitive Diagnosis. Applied Psychological Measurement, 29(4), 262-277.

Henson, R. A., & Templin, J. L. (2007, April). Large-scale language assessment using cognitive diagnosis models. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Henson, R., & Templin, J. (2008, March). Implementation of standards setting for a geometry end-of-course exam. Paper presented at the annual meeting of the American Educational Research Association, New York.

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

Henson, R., Templin, J., Willse, J., & Irwin, P. (2014). Obtaining Diagnostic Information from Large Scale Test. Unpublished manuscript.

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? Educational Measurement: Issues and Practice, 28(3), 24-31.

Houts, C. R., & Cai, L. (2013). flexMIRT® user’s manual version 2: Flexible multilevel multidimensional item analysis and test scoring. Chapel Hill, NC: Vector Psychometric Group.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Sensitivity to under-parameterized model misspecification. Psychological Methods, 3(4), 424-453.

Huff, K., & Goodman, D. P. (2007). The Demand for Cognitive Diagnostic Assessment. In J. P. Leighton & M. J. Gierl (Eds.). Cognitive diagnostic assessment for education: Theory and applications (pp. 19-60). Cambridge, UK: Cambridge University Press.

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), Towards adaptive CALL: Natural language processing for diagnostic language assessment (pp. 117-131). Ames, IA: Iowa State University.

Jang, E. E. (2009a). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. Language Testing, 26(1), 31-73.

Jang, E. E. (2009b). Demystifying a Q-Matrix for Making Diagnostic Inferences About L2 Reading Skills. Language Assessment Quarterly, 6, 210-238.

Jang, E. E. (2010). Demystifying a Q-Matrix for Making Diagnostic Inferences About L2 Reading Skills: The Author Responds. Language Assessment Quarterly, 7, 116-117.

Junker, B. W. (2007, July). Some issues and applications in cognitive diagnosis and educational data mining. Presentation at the International Meeting of the Psychometric Society, Tokyo, Japan.

Junker, B. W., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. Applied Psychological Measurement, 25(3), 258-272.

Jurich, D. P. (2014). Assessing Model Fit of Multidimensional Item Response Theory and Diagnostic Classification Models using Limited-Information Statistics. Unpublished doctoral dissertation, Department of Graduate Psychology, James Madison University, Harrisonburg, VA.

Jurich, D. P., & Bradshaw, L. P. (2014). An Illustration of Diagnostic Classification Modeling in Student Learning Outcomes Assessment. International Journal of Testing, 14, 49-72.

Kim, Y.-H. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified Model. Language Testing, 28(4), 509-541.

Kline, R. B. (2011). Principles and Practice of Structural Equation Modeling, 3rd ed. New York, NY: Guilford.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. Studies in Educational Evaluation, 35(2-3), 64-70.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. Journal of Educational Measurement, 49(1), 59-81.

Lazarsfeld, P. F., & Henry, N. W. (1968). Latent structure analysis. Boston, MA: Houghton Mifflin.

Lee, Y., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. Language Assessment Quarterly, 6(3), 239-263.

Li, H., & Suen, H. K. (2013). Constructing and Validating a Q-Matrix for Cognitive Diagnostic Analyses of a Reading Test. Educational Assessment, 18, 1-25.

Linacre, J. M. (1996). Overlapping Normal Distributions. Rasch Measurement Transactions, 10(1), 487-488.

Lord, F. M., & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Addison-Wesley.

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

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 2, 99-120.

Madison, M. J., & Bradshaw, L. P. (2014). The Effects of Q-matrix Design on Classification Accuracy in the Log-Linear Cognitive Diagnosis Model. Educational and Psychological Measurement, (in press).

Maris, E. (1995). Psychometric latent response models. Psychometrika, 60(4), 523-547.

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

Markon, K. E., & Kruger, R. F. (2006). Information-Theoretic Latent Distribution Modeling: Distinguishing Discrete and Continuous Latent Variable Models. Psychological Methods, 11(3), 228-243.

Martella, R. C., Nelson, J. R., Morgan, R. L., & Marchand-Martella, N. E. (2013). Understanding and Interpreting Educational Research. New York, NY: Guilford.

Muthén, L. K., & Muthén, B. O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Perie, M., Marion, S., & Gong, B. (2009). Moving Toward a Comprehensive Assessment System: A Framework for Considering Interim Assessments. Educational Measurement: Issues and Practice, 28(3), 5-13.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut.

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

Reiser, B., & Faraggi, D. (1999). Confidence intervals for the overlapping coefficient: the normal equal variance case. Journal of the Royal Statistical Society, 48(3), 413-418.

Rencher, A. C. (2002). Methods of Multivariate Analysis, 2nd ed. New York, NY: Wiley.

Roberts, M. R., & Gierl, M. J. (2010). Developing Score Reports for Cognitive Diagnostic Assessments. Educational Measurement: Issues and Practice, 29(3), 25-38.

Robitzsch, A., Keifer, T., George, A. C., & Uenlue, A. (2013). 'CDM': Cognitive Diagnosis Modeling. R package version 2.3-15.

Rosenbaum, P. R. (1988). Items Bundles. Psychometrika, 53(3), 349-359.

Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis. Applied Psychological Measurement, 14, 271-282.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. H. (2007a). The Fusion Model Skills Diagnosis System. In J. P. Leighton & M. J. Gierl (Eds.), Cognitively Diagnostic Assessment for Education: Theory and Practice (pp. 275-318). Thousand Oaks, CA: Sage.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007b). Skills diagnosis using IRT-based latent class models. Journal of Educational Measurement, 44, 293-311.

Rupp, A. A., & Templin, J. L. (2008a). Unique Characteristics of Diagnostic Classification Models: A Comprehensive Review of the Current State-of-the-Art. Measurement: Interdisciplinary Research and Perspectives, 6(4), 219-262.

Rupp, A. A., & Templin, J. (2008b). The Effects of Q-matrix Misspecification on Parameter Estimates and Classification Accuracy in the DINA Model. Educational and Psychological Measurement, 68(1), 78-96.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). Diagnostic Measurement: Theory, Methods, and Applications. New York: Guilford Press.

Schwartz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

Shepard, L. A. (2009). Commentary: Evaluating the Validity of Formative and Interim Assessment. Educational Measurement: Issues and Practice, 28(3), 32-37.

Shu, Z., Henson, R. A., & Willse, J. T. (2013). Using Neural Network Analysis to Define Methods of DINA Model Estimation for Small Sample Sizes. Journal of Classification, 30(2), 173-194.

Sinharay, S., & Almond, R. G. (2007). Assessing Fit of Cognitive Diagnostic Models: A Case Study. Educational and Psychological Measurement, 67(2), 239-257.

Skrondal, A., & Rabe-Hesketh, S. (2004). Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. Boca Raton, FL: CRC Press.

Stout, W. (1987). A Nonparametric Approach for Assessing Latent Trait Unidimensionality. Psychometrika, 52(4), 1987.

Stout, W. F. (1990). A New Item Response Theory Modeling Approach with Applications to Unidimensionality Assessment and Ability Estimation. Psychometrika, 55(2), 293-325.

Tatsuoka, K. K. (1985). A Probabilistic Model for Diagnosing Misconceptions in the Pattern Classification Approach. Journal of Educational Statistics, 10(1), 55-73.

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.

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

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

Templin, J., Cohen, A., & Henson, R. (2008). Constructing tests for optimal classification in standard setting. Manuscript under review.

Templin, J., Poggio, A., Irwin, P., & Henson, R. (2008). Combining borderline and contrasting group methods: A latent class analysis approach to standard setting. Manuscript under review.

Templin, J., & Henson, R. (2009). Extracting Diagnostic Information from Existing Large Scale Tests. Invited talk, Department of Psychology, Georgia Institute of Technology. Retrieved from: http://jonathantemplin.com/files/presentations/jtemplin_gatech2009.pdf.

Templin, J., & Bradshaw, L. (2013). Measuring the Reliability of Diagnostic Classification Model Examinee Estimates. Journal of Classification, 30, 251-273.

Templin, J., & Hoffman, L. (2013). Obtaining Diagnostic Classification Model Estimates Using Mplus. Educational Measurement: Issues and Practice, 32(2), 37-50.

Tufte, E. R. (2001). The Visual Display of Quantitative Information, 2nd ed. Cheshire, CT: Graphics Press.

von Davier, M. (2005). A General Diagnostic Model Applied to Language Testing Data. ETS Research Report RR-05-16. Princeton, NJ: Educational Testing Service.

von Davier, M., & Rost, J. (2006). Mixture distribution item response models. In C. R. Rao and S. Sinharay (Eds.), Handbook of statistics (vol. 26): Psychometrics. Amsterdam: Elsevier.

von Davier, M., & Xu, X. (2009). Software for Multidimensional Discrete Latent Trait Models: mdltm.

Wang, S., & Douglas, J. (2013). Consistency of Nonparametric Classification in Cognitive Diagnosis. Psychometrika, (in press).

Wang, Y.-C. (2009). Factor Analytic Models and Cognitive Diagnostic Models: How Comparable Are They? A Comparison of R-RUM and Compensatory MIRT Model with Respect to Cognitive Feedback. Unpublished doctoral dissertation, Department of Educational Research Methodology, University of North Carolina at Greensboro, Greensboro, NC.

Willse, J. T., Henson, R. A., & Templin, J. L. (2007). Using Sum Scores or IRT in Place of Cognitive Diagnosis Models: Can Existing or More Familiar Models Do the Job? Presented at the National Council on Measurement in Education annual meeting, Chicago, Illinois.

Willse, J. T. (2011). Mixture Rasch Models With Joint Maximum Likelihood Estimation. Educational and Psychological Measurement, 71(1), 5-19.

Xu, X., & von Davier, M. (2008). Cognitive Diagnosis for NAEP Proficiency Data. ETS Research Report RR-06-08. Princeton, NJ: Educational Testing Service.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187-213.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8(2), 125-145.

Zhao, F. (2013). Using Noncompensatory Models in Cognitive Diagnostic Mathematics Assessments: An Evaluation Based on Empirical Data. Unpublished doctoral dissertation, Department of Psychology and Research in Education, The University of Kansas, Lawrence, KS.

Zheng, Y., Chiu, C.-Y., & Douglas, J. A. (2013). The NPCD Package: Nonparametric Methods for Cognitive Diagnosis, R package, http://cran.r-project.org/web/packages/NPCD/index.html.

APPENDIX A

VARIANCE OF CONTINUOUS TRAIT COMPOSITE FOR COMPLEX STRUCTURE

When multiple abilities are required for the *i*-th item for the *e*-th examinee (complex structure), one way to conceptualize this within the MCCIRM of Eq. (48) is to consider it as a weighted composite, $\ddot{\theta}_e$, of the *A* multiple abilities, and in particular as discussed above, one ability per *a* attributes described by

$$\ddot{\theta}_e = \sum_{a=1}^{A} \gamma_{ia} \theta_{ea} c_{ia} \,. \tag{A1}$$

This weighted ability composite aspect of the MCCIRM is akin to a "reference" composite (Reckase, 2009) estimated in a unidimensional IRT model when there are actually multiple underlying abilities. If complex structure is present (i.e., ç > 1), the variance of the weighted composite, $Var\left(\ddot{\theta}_e\right)$, can be expressed as

$$Var\left(\ddot{\theta}_e\right) = \sigma^2_{\ddot{\theta}_e} = Var\left(\sum_{a=1}^{A} \gamma_{ia} \theta_{ea} c_{ia}\right). \tag{A2}$$

This can be re-written with only those required abilities (i.e., where all $c_{ia} = 1$) and re-enumerating starting at $a = 1$ as

$$Var\left(\ddot{\theta}_e\right) = Var\left(\sum_{a=1}^{\varsigma} \gamma_{ia}\theta_{ea}\right). \tag{A3}$$

Now, an additional constraint in the current study by assuming γ_{ia} is set to a fixed value for generation (rather than estimate it as in the C-MIRT model), where:

$$\gamma_{i1}^2 = \gamma_{i2}^2 = \ldots = \gamma_{iA}^2 \equiv \gamma_{i.}^2. \tag{A4}$$

Given a common fixed $\gamma_{i.}$ is assumed in the above from Eq. (A3), then this can be rewritten as

$$Var\left(\ddot{\theta}_e\right) = \gamma_{i.}^2\left[Var\left(\sum_{a=1}^{\varsigma} \theta_{ea}\right)\right]. \tag{A5}$$

Because the multiple θ_{ea} could be drawn from a multivariate Normal distribution with possibly nonzero correlations among abilities, the variance of this weighted sum is given by

$$Var\left(\ddot{\theta}_e\right) = \gamma_{i.}^2\left[\sum_{a=1}^{ç} Var\left(\theta_{ea}\right) + 2\sum_{1\le a\ <a'\le ç}\sum Cov\left(\theta_{ea},\theta_{ea'}\right)\right]. \qquad \text{(A6)}$$

For simplicity (and discussed later below) it is further assumed that $\gamma_{i.} = 1$. Then,

$$Var\left(\ddot{\theta}_e\right) = \sum_{a=1}^{ç} Var\left(\theta_{ea}\right) + 2\sum_{1\le a\ <a'\le ç}\sum Cov\left(\theta_{ea},\theta_{ea'}\right). \qquad \text{(A7)}$$

So, to degenerate the total ability variance of $\ddot{\theta}_e$, not only must the attribute-specific variance of θ_{ea} degenerate, but also the covariance of θ_{ea} with the other abilities must be considered. Again, this only practically influences the total variance of the weighted sum if there is a complex loading structure among involved items for abilities. Because $Var\left(\theta_{ea}\right) = \sigma^2_{\theta_{e.}}$, an assumed common variance for all companion latent traits to attributes, then

$$Var\left(\ddot{\theta}_e\right) = ç\sigma^2_{\theta_{e.}} + 2\sum_{1\le a\ <a'\le ç}\sum Cov\left(\theta_{ea},\theta_{ea'}\right). \qquad \text{(A8)}$$

Now, the correlation is related to covariance by

$$Cov\left(\theta_{ea},\theta_{ea'}\right)=Corr\left(\theta_{ea},\theta_{ea'}\right)\times\sqrt{Var\left(\theta_{ea}\right)}\times\sqrt{Var\left(\theta_{ea'}\right)}\ . \quad \text{(A9)}$$

Again, since a common ability variance $\sigma^2_{\theta_{e\bullet}}$ is assumed, then

$$Cov\left(\theta_{ea},\theta_{ea'}\right)=Corr\left(\theta_{ea},\theta_{ea'}\right)\times\sigma^2_{\theta_{e\bullet}}\ . \quad \text{(A10)}$$

Assume $\rho=Corr\left(\theta_{ea},\theta_{ea'}\right)$ for all $a\neq a'$, a common correlation between pairs of traits. Here, the matrix of all such correlations among required abilities has the form

$$\underset{\sim}{\mathrm{P}}_{ç\times ç}=\begin{bmatrix}1 & \rho & \cdots & \rho\\ \rho & \ddots & & \vdots\\ \vdots & & \ddots & \rho\\ \rho & \cdots & \rho & 1\end{bmatrix}=\left(1-\rho\right)\underset{\sim}{\mathrm{I}}_{ç\times ç}+\rho\cdot\underset{\sim}{\mathrm{J}}_{ç\times ç}\ , \quad \text{(A11)}$$

Where $\underset{\sim}{\mathrm{I}}$ is a $ç \times ç$ identity matrix and $\underset{\sim}{\mathrm{J}}$ is a $ç \times ç$ matrix of ones. Thus, $\underset{\sim}{\mathrm{P}}$ is assumed to be compound symmetric (Rencher, 2002) in this study, although other possibilities (e.g., unstructured correlation matrix) exist and can be studied in future research. This then implies

$$Cov\left(\theta_{ea},\theta_{ea'}\right)=\rho\sigma^2_{\theta_{e\bullet}}\ . \quad \text{(A12)}$$

Given that there is $\left[ç(ç-1)\right]/2$ terms in the double-sum over covariances in Eq. (A8), the variance of the ability weighted composite can be re-expressed as

$$Var\left(\ddot{\theta}_{e}\right) = ç\sigma^{2}_{\theta_{e}} + ç\left(ç-1\right)\rho\sigma^{2}_{\theta_{e.}}. \tag{A13}$$

Let this variance of the composite in the complex case be denoted as $\sigma^{2}_{\theta_{e.},C}$. We similarly denote the variance in simple structure case to be $\sigma^{2}_{\theta_{e.},S}$. In simple structure, the composite variance reduces to just the common variance assumed for an ability, $\sigma^{2}_{\theta_{e.},S} = \sigma^{2}_{\theta_{e.}}$, because for simple structure

$$Var\left(\sum_{a=1}^{A}\gamma_{ia}\theta_{ea}\right) = Var\left(\theta_{ea}\right) = \sigma^{2}_{\theta_{e.}}. \tag{A14}$$

As an example, assume complexity = 2 and common correlation among traits = 0.70 leads to the following based on Eq. (A13):

$$Var\left(\ddot{\theta}_{e}\right) = 2\sigma^{2}_{\theta_{e.}} + 2\left(2-1\right)0.7\sigma^{2}_{\theta_{e.}} = 3.4\sigma^{2}_{\theta_{e.}}. \tag{A15}$$

Thus, that $\sigma^{2}_{\theta_{e.},C} > \sigma^{2}_{\theta_{e.},S}$ as would be expected from the above. So now a question is motivated: What scenarios would lead to $\sigma^{2}_{\theta_{e.},C} = \sigma^{2}_{\theta_{e.},S}$? This means that trait variance

is the same for complex and simple structure. Under the previously mentioned assumptions, this would occur when

$$\gamma_{i.}^2\left[\varsigma\sigma_{\theta_{e.}}^2+\varsigma\left(\varsigma-1\right)\rho\sigma_{\theta_{e.}}^2\right]=\sigma_{\theta_{e.}}^2. \tag{A16}$$

Assuming nonzero common variance and dividing through by results in

$$\gamma_{i.}^2\left[\varsigma+\varsigma\left(\varsigma-1\right)\rho\right]=1. \tag{A17}$$

Thus, choosing values of $\gamma_{i.}$ in the following way gives the equality

$$\gamma_{i.}^2=\frac{1}{\varsigma+\varsigma\left(\varsigma-1\right)\rho}. \tag{A18}$$

An example is considered next. If $\sigma_{\theta_{e.}}^2=1,\varsigma=2,$ and $\rho=0.70$, then choosing $\gamma_{i.}^2=1/3.4\approx0.2941$ (and thus $\gamma_{i.}\approx0.5423$) results in equality. Therefore, whenever

$$\gamma_{i.}^2>\frac{1}{\varsigma+\varsigma\left(\varsigma-1\right)\rho}, \tag{A19}$$

And under the previous assumptions, then $\sigma_{\theta_{e.},\mathrm{C}}^2>\sigma_{\theta_{e.},\mathrm{S}}^2$ which means that the total ability variance of the weight composite is greater in the complex case than in simple structure. When this occurs, the effect size, Δ, from Eq. (51) will be smaller because its

denominator will be larger whenever there is complex structure. *Overall, choosing* $\gamma_{i.}=1$ *is made throughout the current study*, which allows easy implementation but guarantees that $\sigma^2_{\theta_{e\square},C} > \sigma^2_{\theta_{e\bullet},S}$ as would be expected from the above. Thus, under these conditions complex structure is hypothesized to have a detrimental effect on diagnostic classification relative to simple structure because of this nature of the composite variance and the covariance introduced by the individual positively correlated abilities. However, it is notable that from Eq. (A6) that as the $Var(\theta_{ea}) \to 0$, then so must the $Cov(\theta_{ea}, \theta_{ea'}) \to 0$ (because of Eq. (A13) under the assumptions). Thus, when ability variance degenerates the total composite ability variance does also, even with positively correlated traits.