EXAMINING THE EFFECTS OF RETROFITTING DIAGNOSTIC CLASSIFICATION MODELS TO ITEM RESPONSE THEORY DATA

by

ALLEN CHRISTOPHER MOORE

(Under the Direction of Matthew Madison)

ABSTRACT

Most recent educational assessments were developed under an item response theory (IRT) framework and were designed to scale examinees on an ability continuum. Because there is an increasing desire for more formative feedback about mastery of specific skills, researchers and educators have recently been exploring diagnostic classification models (DCMs), which have potential to improve educational assessment and research due to the multidimensional nature of their feedback. However, because DCMs have only been introduced in recent years, they are not yet widely implemented in analysis. This study seeks to examine the effects of *retrofitting*, or retroactively fitting a DCM to item responses from an assessment not intended for classification. Using simulations and empirical analysis, I explore item parameter translation and model fit performance within a retrofitting context.

INDEX WORDS: diagnostic classification models, item response theory, retrofitting, item parameters, model fit

EXAMINING THE EFFECTS OF RETROFITTING DIAGNOSTIC CLASSIFICATION MODELS TO ITEM RESPONSE THEORY DATA

by

ALLEN CHRISTOPHER MOORE

BS, The University of Georgia, 2020

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF ARTS

ATHENS, GEORGIA

2025

EXAMINING THE EFFECTS OF RETROFITTING DIAGNOSTIC CLASSIFICATION MODELS TO ITEM RESPONSE THEORY DATA

by

ALLEN CHRISTOPHER MOORE

Major Professor:	Matthew Madison
Committee:	Amanda Ferster
	Shiyu Wang

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2025

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

Item response theory (IRT) is a traditional assessment framework widely used in large-scale educational assessments of knowledge or content understanding. Assessments developed under this framework aim to scale examinees on a relative continuum of ability level. There is an increasing desire for more formative feedback during and alongside instruction about which skills examinees have and have not mastered, in addition to obtaining their traditional total score. Consequently, many researchers are exploring diagnostic models to help assess students' content mastery and inform instructional practices.

Within the past decade or so, diagnostic classification models (DCMs; Rupp et al., 2010) have emerged as a new approach to educational assessment in which examinees can be probabilistically classified into one of two or more proficiency statuses based on their responses to assessment items. Most commonly, DCMs classify examinees as either masters or non-masters of a given attribute. DCMs have promising potential to improve educational assessment and educational research, as they allow for diagnostic and formative feedback about which standards students have mastered, which can then be used to guide instructional practices. However, because most assessments today are designed under the IRT framework, DCMs are not yet widely utilized for providing meaningful formative feedback. The best way to produce meaningful and accurate examinee classifications is to use an assessment that was designed to be diagnostic, but this is not always an option. Creating a diagnostic assessment can be time-

consuming, and there is already a great deal of data for tests developed under the IRT framework.

Calibrating DCMs with item responses from an IRT framework is a possibility that has not yet been thoroughly examined. *Retrofitting* occurs when a DCM is used to retroactively classify examinees as attribute masters or non-masters based on their responses for an assessment not designed to classify examinees (Liu et al., 2017). In an article exploring retrofitting DCMs to IRT data, Liu et al. (2017) express that retrofitting is not recommended as a practical means to obtain examinee classifications, but it is noted that the practice could be useful for learning more about construct domains. The authors suggest that a retrofitted model may be used when it is not feasible to develop a diagnostic test, when the purpose of the analysis is to obtain formative feedback to support learning or improve understanding of a construct, or for some assessments where there are distinct subdomains that items are intended to measure. To that end, the purpose of this study is to examine both the relationship between the item parameters for IRT-generated data and the item parameters that a diagnostic model will estimate, as well as the performance of fit statistics in a retrofitting context.

In the first simulation, I examined the extent to which changes in IRT item parameters in a two-parameter logistic model would affect analogous parameters for DCM items. The ultimate goals of this simulation were to better inform appropriate use cases of retrofitting and to better understand the impacts of retrofitting on examinee classifications. In the first simulation, I examined if harder IRT items translated to harder DCM items, as well as if highly discriminating IRT items translated to highly discriminating DCM items.

In the second simulation, I sought to explore how model fit statistics performed in the retrofitting context. The goal of this study was to gain insight into how well, if at all, the

retrofitted model would pick up on the misfit from the initial data generation process. Understanding how the model fit indices perform could be helpful in later determining thresholds for adequate model fit when retrofitting.

Lastly, I conducted an empirical analysis on the dataset SAT12 included with the *mirt* package (Chalmers, 2012) in R, which was obtained from the TESTFACT (Woods et al., 2003) manual. These data comprise 600 examinees' responses to a multiple-choice, 12th-grade science assessment test (SAT) intended to measure chemistry, biology, and physics. The documentation for the dataset included a scoring key, along with a note that the key for one of the items might be incorrect. A modified scoring key was provided and used in my analysis.

In the following sections, I describe the models and methods that were used to carry out the two simulation-based studies and the empirical data analysis.

CHAPTER 2

LITERATURE REVIEW

This chapter summarizes the frameworks and models used in the methods of these studies, building from the concepts of item response theory and latent class modeling to the newer framework of diagnostic classification models.

Item Response Theory

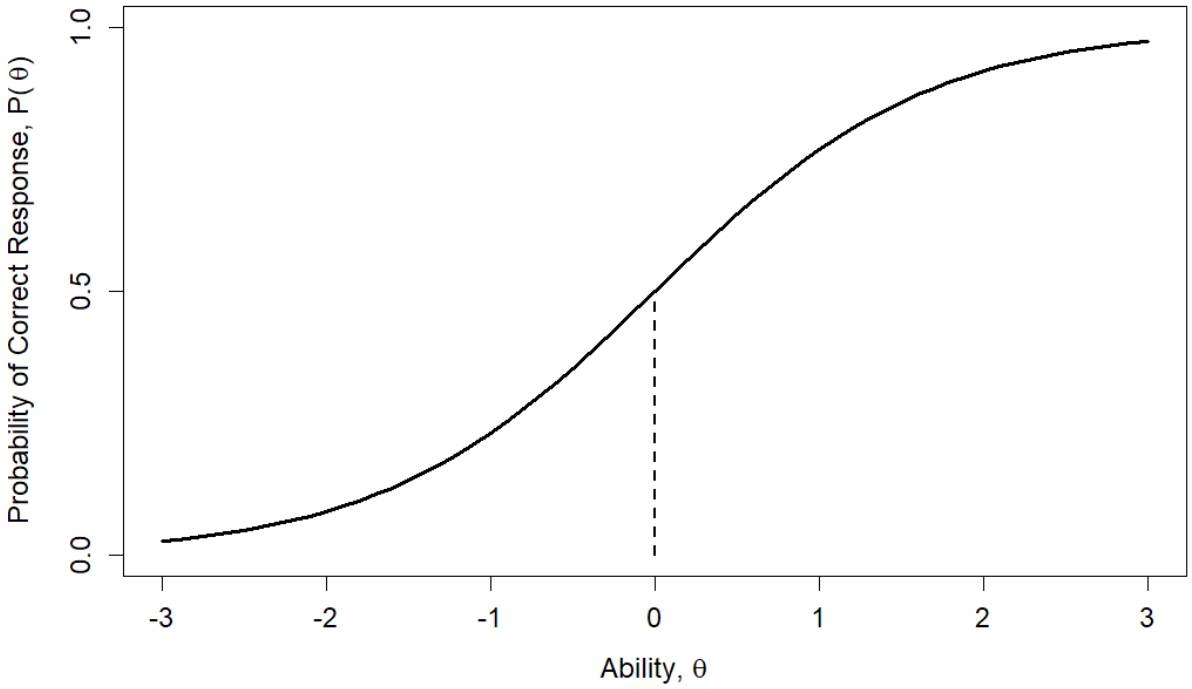
IRT is currently the most ubiquitous framework for analyzing large-scale assessment results, having developed from the concepts of classical test theory (CTT) before it. While CTT was focused on an assessment as a whole and considered overall scores to sets of questions, IRT focuses on each single item, modeling the relationship between a person's latent ability and the probability that they respond correctly to a specific item. IRT models grew in popularity due to their appealing invariance properties, in which item parameter estimates are not dependent on the characteristics of the sample, and ability estimates are not dependent on the specific assessment.

A common IRT model, the two-parameter logistic model (2-PL; Birnbaum, 1968), models probability of correct response $P(X_{ij} = 1)$ to item i for examinee j as a function of two item parameters and continuous examinee ability level θ_j. Item *difficulty*, denoted b_i, is defined as the point along the ability continuum at which an examinee would have a $.50$ chance of answering the item correctly. Item *discrimination,* denoted a_i, quantifies how well the item can discriminate between examinees who know the correct answer and those who do not. The item response function for a 2-PL model can be expressed as

$$P(X_{ij} = 1 \mid \theta_j, a_i, b_i) = \frac{\exp\left(a_i(\theta_j - b_i)\right)}{1 + \exp\left(a_i(\theta_j - b_i)\right)}$$

The probability of correct response to an item under an IRT framework can also be expressed visually using an item characteristic curve (ICC). An example ICC for an item with $a = 1.2$ and $b = 0$ is shown in Figure 1.

Figure 1. *Example Item Characteristic Curve (ICC) for a 2-PL Model Item*



The dashed vertical line at zero represents the difficulty of this item, which demonstrates that an examinee with an ability level of zero will have a .50 chance of answering this item correctly. Although harder to visualize, the slope of this curve at its steepest point represents discrimination.

Latent Class Models

In contrast with the IRT framework, latent class models (LCMs) represent correct response probability as being conditional on categorical latent class membership. These analyses are generally exploratory in nature, as the number of classes is not known before analysis. Instead, researchers must look at the relevant theory and the characteristics of the data through factor analysis to determine the number of latent classes present. Additionally, in traditional latent class analyses, no constraints are placed on the item-attribute relationships *a priori*, meaning it is unknown or unspecified which items measure each attribute. The general latent class formula models the probability of observing a specific latent class as

$$P(X_r = x_r) = \sum_{c=1}^{C} \nu_c \prod_{i=1}^{I} \pi_{ic}^{x_{ir}} (1 - \pi_{ic})^{1-x_{ir}}$$

where ν_c is the probability of membership in latent class c, π_{ic} is the probability of correct response to item i for a member of latent class c, and x_{ir} is the observed response of respondent r to item i.

Diagnostic Classification Models

Diagnostic classification models (DCMs; Rupp et al., 2010) are a special family of latent class models that are relatively recent in psychometric literature and are unique in being both confirmatory and constrained. While a typical latent class model is exploratory, DCMs are confirmatory for two reasons. The first is that the number of latent classes are specified prior to analysis as distinct attribute profiles, or vectors of length A with elements $\alpha_c = [\alpha_1, \alpha_2, \ldots, \alpha_A]$ indicating which attributes the examinee has mastered. Each number in the attribute profile will have a value of 1 for attributes that an examinee has mastered and a value of 0 for attributes that an examinee has not mastered. For example, on an assessment measuring three attributes, an

examinee with the attribute profile $[1, 0, 1]$ will have mastered attributes α_1 and α_3, but not attribute α_2. An examinee with the attribute profile $[0, 0, 0]$ would be referred to as a *complete non-master* of the set of attributes, and an examinee with the attribute profile $[1, 1, 1]$ would be referred to as a *complete master* of the set of attributes.

The second reason DCMs are considered confirmatory is that item-attribute relationships are also specified before analysis in a *Q-matrix*, which maps items to the attribute(s) that they measure (Tatsuoka, 1983). A Q-matrix has I rows and A columns, where I is the number of items on the assessment and A is the number of attributes the assessment is intended to measure. An entry $q_{i,a}$ in the Q-matrix will take on a value 1 if item i measures attribute a and a 0 if it does not. A sample three-item, three-attribute Q-matrix is shown in Table 1. In this example, each item measures two of the three attributes.

Table 1. *Sample Three-Item, Three-Attribute Q-Matrix*

Item	α_1	α_2	α_3
1	1	1	0
2	1	0	1
3	0	1	1

Additionally, DCMs are constrained due to the restrictions placed on how attributes impact item responses, as specified in the measurement model. The constraints on the model force attribute mastery to translate to increased correct response probability such that a master of one attribute must have a higher correct response probability than a complete non-master.

The Log-Linear Cognitive Diagnostic Model

Several DCMs have been proposed and explored in the literature surrounding diagnostic measurement, all of which make different assumptions about how attribute mastery impacts item responses. Some of these models are general while others are more constrained and make stricter assumptions. The model used in this study is the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009), which subsumes many other commonly used DCMs. The LCDM item response function for a binary item *i* on an assessment measuring two attributes, α_1 and α_2, models the correct response probability $P(Y_{ei} = 1 \mid \alpha_c)$ for an examinee e in class α_c as

$$P(Y_{ei} = 1 \mid \alpha_c) = \frac{\exp\left(\lambda_{i,0} + \lambda_{i,1,(1)}(\alpha_1) + \lambda_{i,1,(2)}(\alpha_2) + \lambda_{i,2,(1,2)}(\alpha_1 \cdot \alpha_2)\right)}{1 + \exp\left(\lambda_{i,0} + \lambda_{i,1,(1)}(\alpha_1) + \lambda_{i,1,(2)}(\alpha_2) + \lambda_{i,2,(1,2)}(\alpha_1 \cdot \alpha_2)\right)}$$

Where the intercept $\lambda_{i,0}$ is the log-odds of a non-master answering the item correctly; $\lambda_{i,1,(1)}$ is the increase in log-odds of a correct responses for masters of only the first attribute; $\lambda_{i,1,(2)}$ is the increase in log-odds of a correct response for masters of only the second attribute; and $\lambda_{i,2,(1,2)}$ is the increase in log-odds of a correct response for masters of both attributes. An item on an assessment measuring two attributes with the parameters $\lambda_{i,0} = -2.5$, $\lambda_{i,1,(1)} = 1.5$, $\lambda_{i,1,(2)} = 2$, and $\lambda_{i,2,(1,2)} = 1$ would estimate the correct response probability for a complete non-master as

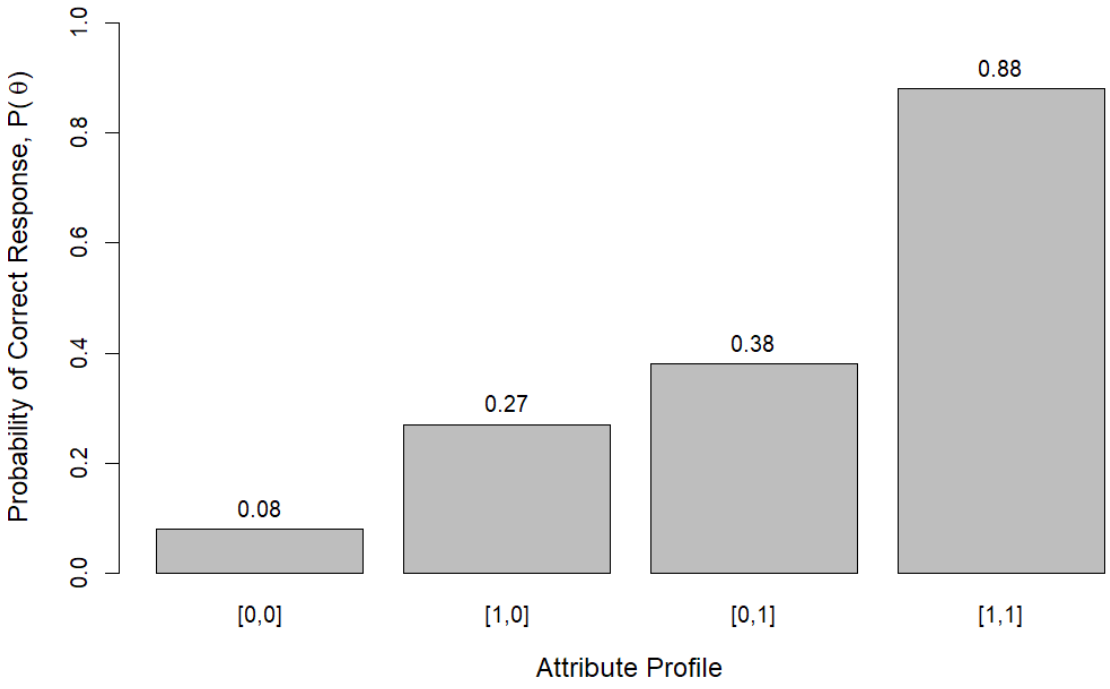
$$P(Y_{ei} = 1 \mid [0,0]) = \frac{\exp\left(-2.5 + 1.5(0) + 2(0) + 1(0 \cdot 0)\right)}{1 + \exp\left(-2.5 + 1.5(0) + 2(0) + 1(0 \cdot 0)\right)} = .076$$

Similarly, the model would estimate the correct response probability for a complete master of both attributes as

$$P(Y_{ei} = 1 \mid [1,1]) = \frac{\exp\left(-2.5 + 1.5(1) + 2(1) + 1(1 \cdot 1)\right)}{1 + \exp\left(-2.5 + 1.5(1) + 2(1) + 1(1 \cdot 1)\right)} = .881$$

The probabilities of correct response for each attribute profile can also be visually represented in an item characteristic bar chart (ICBC). A sample ICBC for the above example item is shown below in Figure 2.

Figure 2. *Example Item Characteristic Bar Chart (ICBC) for an LCDM Item*



While the 2-PL model directly includes difficulty and discrimination parameters in the item response function, DCMs do not include such parameters. Without the presence of similar item parameters for the LCDM, conceptualizing the characteristics of an item was challenging as the framework is based on examinee class membership rather than examinee ability level. In this retrofitting context, I have defined item *difficulty* as the average correct response probability between masters of an attribute and non-masters of an attribute, which will notably only work with a simple-structure Q-matrix. Although there are multiple ways I could have defined this parameter, I felt that average correct response probability best captured how generally easy or

hard an item would be to different members of a similar sample. Similarly, I have defined *discrimination* in this retrofitting context as the difference between the correct response probabilities for masters and non-masters of an attribute. It is important to note that these definitions would need to be adjusted for studies involving complex items, which measure more than one attribute, to account for partial and complete masters and non-masters of all attributes involved.

Model Misspecification

Research on model misspecification in DCMs has started to appear, much of which examines the effects of Q-matrix misspecification on parameter estimates and model fit. A Q-matrix is considered misspecified when the items in the Q-matrix do not properly align with the attributes they truly measure, which can occur when designing the Q-matrix for an assessment (Madison and Bradshaw, 2014). Careful Q-matrix design, influenced by the opinions of content experts, is essential to any diagnostic assessment design process. Different types of Q-matrix misspecification have been shown to impact both reliability and classification accuracy, severely impacting the validity of any inferences made using the assessment.

For example, Rupp and Templin (2007) observed through a simulation study that while overspecified Q-matrices did not result in strongly decreased classification accuracy, underspecified Q-matrices did decrease classification accuracy in the DINA model. Kunina-Habenicht et al. (2012) found through a simulation study that classification accuracy for log-linear models can significantly decrease with too many incorrect Q-matrix entries, or if assuming the wrong number of dimensions. Madison (2023) used simulations to observe that with longitudinal DCMs, classification accuracy was not strongly impacted when only either the measurement model or structural model was misspecified, but both types of model

misspecification occurring together severely decreased performance in both classification accuracy and reliability.

Other types of model misspecification include measurement non-invariance and item parameter drift. Measurement non-invariance describes violations to the assumption of measurement invariance, which assumes that item parameter estimates are independent of the sample and examinee classifications are independent of the items. One source of measurement non-invariance is differential item functioning (DIF).

DIF is broadly defined as an item functioning differently for examinees who have matched proficiency measures but are from different groups (Angoff, 1993). Because DCMs assume multidimensionality, a more specific definition for DIF in the DCM framework is the difference in probabilities of correct response between two examinees with the same attribute profile but are from different groups (Hou et al., 2014; Li & Wang, 2015). There have been multiple studies examining methods for detecting DIF in the DCM framework, ranging from more traditional methods such as the Mantel-Haenszel test (MH; Mantel & Haenszel, 1959) and the simultaneous item bias test (SIBTEST; Shealy & Stout, 1993) to those which are based specifically on DCM models, including a modified higher-order DINA model (HO-DINA; de la Torre & Douglas, 2004), the Wald test (Hou et al., 2014), and the LCDM-DIF method proposed by Li and Wang (2015). Although functioning differently, these methods all assess the presence of DIF and the degree to which it occurs. Overall, conducting DIF analysis to detect any potential item bias is crucial for ensuring test fairness across multiple groups of examinees (Camilli, 2006).

As an extension of DIF, item parameter drift refers to violations in measurement invariance occurring over time. Madison and Bradshaw (2018) examined item parameter drift in

the context of longitudinal DCMs and found that their proposed model was robust to effects of measurement non-invariance on classification accuracy and reliability, but item parameter estimation was somewhat impacted.

While DCMs can be robust to certain misspecifications, they are prone to suffering decreased classification accuracy with an underfitted Q-matrix, measurement model, and/or structural model. Other types of misspecification seem to worsen model performance only when the degree of misspecification is large or multiple types of misspecification occur simultaneously. There is currently little research available on retrofitting as a type of misspecification.

This study comprises two simulations conducted in R version 4.0.5 (R Core Team, 2021), each of which explored a different aspect of the process and effects of retrofitting, and an empirical analysis of retrofitting on IRT item response data. The first simulation sought to examine how manipulating difficulty and discrimination item parameters might impact the analogous difficulty and discrimination of LCDM items when retrofitting. The second simulation sought to determine the impact of retrofitting on LCDM model and item fit. The empirical analysis used data from the *mirt* package in R with item responses from a 12th-grade science assessment test to empirically examine how item parameters translate and how model fit indices perform in a retrofitting context. The overarching motivation of the three studies combined was to improve how researchers interpret DCM results when retrofitting has occurred. In the following sections, I will detail my methods and results for each of the two simulation studies and the empirical analysis.

CHAPTER 3

SIMULATION 1: ITEM PARAMETERS

Methods

The first simulation examined if item difficulty and/or item discrimination would translate when retrofitting the LCDM to item responses from an assessment developed under the IRT framework. Put another way, this simulation examines the degree to which difficult IRT items translate to difficult DCM items, and the degree to which highly discriminating IRT items translate to high discriminating DCM items. If item difficulty translates, then as IRT item difficulty increases, DCM item difficulty will also increase, and we will observe strong positive correlations between the parameters. Similarly, if discrimination translates, then as IRT item discrimination increases, so will DCM item discrimination, and we will observe strong positive correlations.

Conditions

Models. I used the 2-PL IRT model for data generation and the LCDM for calibration.

Sample Characteristics. The sample consisted of 1,000 generated examinees to be sure that sample size would not affect item parameter recovery in a model with no interaction terms (Sen & Cohen, 2021). These ability levels were generated from a standard normal distribution ($\mu = 0, \sigma = 1$) to mimic ability estimates seen in other IRT contexts.

Item Characteristics. Difficulty was categorized into three levels: easy, medium, and hard. Easy difficulty parameters were generated from a random uniform distribution with a range of $[-1.5, -0.5]$; medium difficulty parameters were generated similarly from a range of

$[-0.5, 0.5]$; hard difficulty parameters were generated similarly from a range of $[0.5, 1.5]$. I chose these ranges to be in line with commonly seen values for item difficulty in the 2-PL setting and to ensure I had equally sized intervals.

Similar to item difficulty, discrimination was categorized into three levels: low, medium, and high. Low discrimination parameters were generated from a random uniform distribution with a range of $[0.25, 0.75]$; medium discrimination parameters were generated from a random uniform distribution with a range of $[0.75, 1.25]$; and high discrimination parameters were generated from a random uniform distribution with a range of $[1.25, 1.75]$. I chose these values based on common interpretations, and because discrimination is typically constrained to be positive but usually does not exceed 2. Difficulty levels were crossed with discrimination levels to create nine total conditions under which I ran the simulation.

Test Characteristics. The total length of the test was 10 items. Templin & Bradshaw (2013) used a simulation study to examine the reliability for the assessment with different numbers of attributes and items per attribute, and they found an average reliability for DCMs of .8 with ten items measuring a single attribute.

Procedure

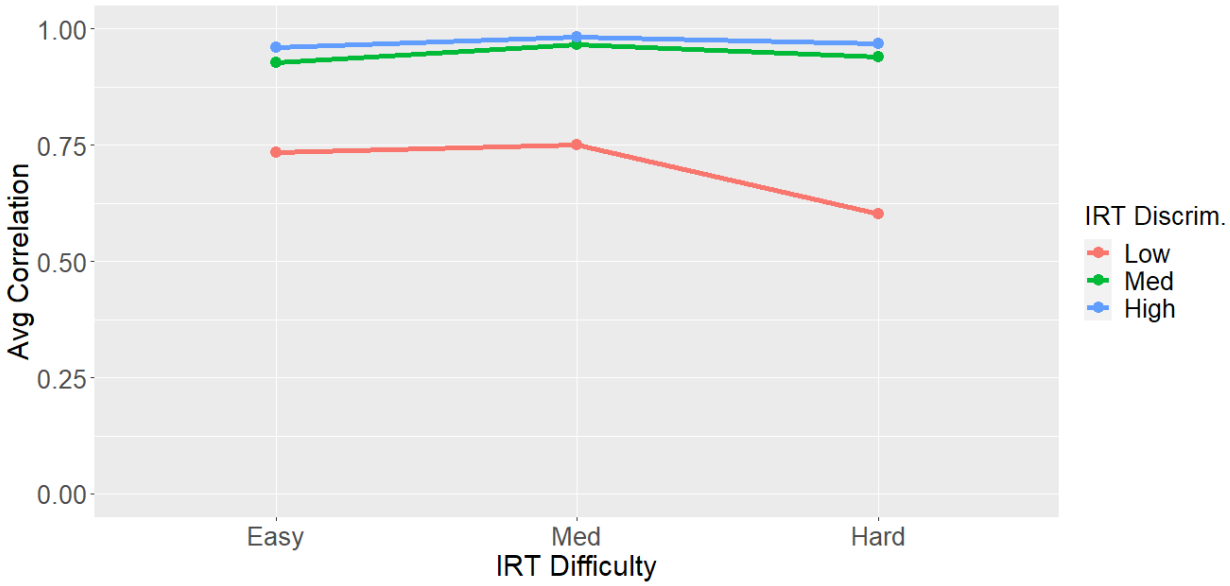
During each replication, 10 pairs of item difficulty and discrimination parameters were randomly generated. I then generated item responses for a 10-item IRT assessment measuring two dimensions, which were used to calibrate the LCDM. I used the item parameters estimated by the LCDM to calculate correct response probabilities for non-masters and masters of the two attributes, which were then used to calculate the analogous item difficulty and discrimination parameters. I ran 50 replications of this simulation for each of the nine conditions crossing difficulty and discrimination, for a total of 450 replications.

After calculating DCM difficulty and discrimination, I correlated these with the initial IRT difficulty and discrimination parameters, respectively. I then averaged these values across replications for each of the nine conditions as well as recorded the proportion of significant correlations for each condition at a significance level of $\alpha = .05$.

Results

The average correlations between IRT and DCM item difficulties are plotted for each combination of the levels of the initial item parameters in Figure 3.

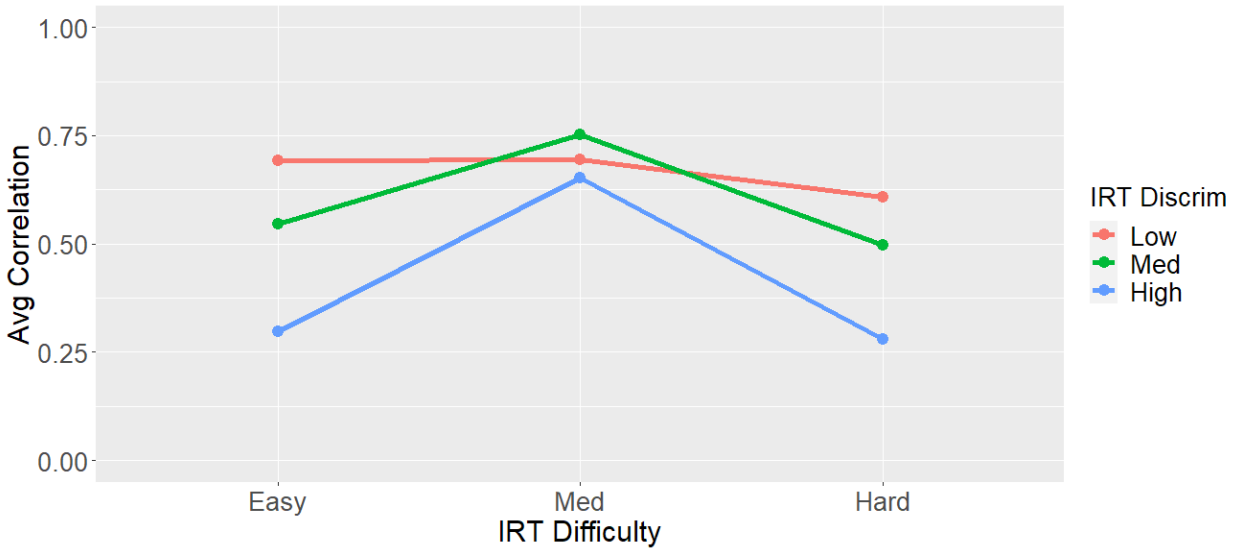
Figure 3. *Average IRT/DCM Difficulty Correlations vs. Initial Parameters*



Overall, results indicate that IRT item difficulty and DCM item difficulty are strongly related. There was a slight interaction, which indicates that the correlation between IRT and DCM item difficulty was stronger for medium and highly discriminating items (~.95) than for the low discriminating items (~.70). These findings indicate that IRT and DCM difficulty tend to have moderate to strong correlations, which are stronger for items with initial medium or high IRT discrimination. This suggests that a harder IRT item does tend to have harder DCM

difficulty as I defined in the context of retrofitting, especially for medium and high discrimination items.

Figure 4. *Average IRT/DCM Discrimination Correlations vs. Initial Parameters*



The average correlations between IRT and DCM discrimination are also plotted for each combination of the levels of the initial item parameters in Figure 4. Overall, positive correlations indicate that higher quality IRT items translate to higher quality DCM items, with correlations ranging from .30 to .70. However, there is a clear interaction where items with low IRT discrimination and medium difficulty seem to have a weaker relationship between IRT and DCM discrimination. Overall, the IRT/DCM discrimination correlations are weaker than the IRT/DCM difficulty correlations. Items with medium and high IRT discrimination did follow a similar trend to one another, as was the case in the difficulty correlations plot.

The relationship between IRT and DCM discrimination does not follow the trend seen with difficulty parameters. Notably, highly discriminating IRT items produced the weakest correlations between IRT and DCM discrimination, while low discrimination IRT items produced two of the three strongest correlations. Items with medium difficulty and medium

discrimination produce the strongest correlation between IRT and DCM discrimination overall. These results indicate that the relationship between IRT and DCM discrimination as I have defined in the context of retrofitting is not as straightforward as for difficulty. This suggests that a highly discriminating IRT item may not necessarily translate into a highly discriminating DCM item, and that further inspection of initial item parameters is recommended.

CHAPTER 4

SIMULATION 2: MODEL FIT

Methods

The goal of this simulation was to examine how commonly used model fit indices, calculated in R, would perform in the retrofitting context.

Conditions

Models. The first model was correctly fitted such that the LCDM was calibrated with item responses generated from the LCDM. The second model was a misfitted model in which the LCDM was calibrated with item responses generated from a 2-PL IRT framework.

Sample Characteristics. For the LCDM data, 1,000 examinees were generated from a binomial distribution ($p = .50$). For the IRT data, 1,000 examinees were generated from the standard normal distribution ($\mu = 0, \sigma = 1$). These sizes were chosen to represent moderate sized samples attainable in research applications and as sufficient sample sizes for misfit detection (citation of some study looking at model fit inference and sample size).

Item Characteristics. For both sets of data, all item parameters were randomly generated from uniform distributions. For the LCDM data, intercepts were generated from a range of $[-2.5, 2]$ and main effects were generated from a range of $[1, 2.5]$. I chose these ranges based on commonly seen values for these item parameters in empirical and simulation-based studies. For the IRT data, difficulty parameters were generated from a range of $[-1.5, 1.5]$ and discrimination parameters from a range of $[0.25, 1.75]$, as in the first simulation.

Test Characteristics. I generated parameters for 10 items designed to measure one attribute with a base mastery rate of .50. I chose 10 items to reflect a practical classroom scenario with a single-attribute, formative, and diagnostic assessment.

Q-Matrix. All questions were designed to measure only one attribute, creating a Q-matrix with one column in which all entries are 1. As with the first simulation, I chose this Q-matrix structure for simplicity of calculations and interpretations.

Procedure

To create the LCDM data, I generated 1,000 examinees' proficiency statuses, 10 item intercepts, and 10 item main effects. These were used to calculate correct response probabilities, which were then compared to a chance matrix to score the responses as correct or incorrect to create response data that was used to calibrate the LCDM.

The IRT data required a similar process in which I generated item parameters and examinee characteristics with which to create item responses, but these were generated under an IRT framework utilizing continuous examinee ability, as well as item difficulty and item discrimination. I then used the resulting item response data to calibrate the LCDM.

For both models, I recorded AIC, BIC, and mean RMSEA for each of the 1,000 replications, as well as the proportion of replications that produced significant χ^2 and M_2 fit indices at $\alpha = .05$. AIC and BIC are both relative fit indices that weigh the balance between goodness of fit with model complexity to achieve an optimal combination of adequate fit and parsimony. Both include penalties for more complex models, but the penalty is larger for BIC than AIC, so BIC is more likely to prefer a simpler model.

RMSEA, which is an absolute fit index, examines differences between the observed and the predicted item responses while adjusting for sample size to quantify how far a hypothesized

model is from being perfectly able to recreate the data. It's commonly used with sample sizes large enough to produce a statistically significant χ^2 statistic even when there is no model misfit. While RMSEA is not intended for use with DCMs, it can be thought of as an approximation error between the model and the data, adjusting for model complexity.

I also recorded the results of two hypothesis tests. Pearson's χ^2 test is used to assess local dependence between pairs of items, but very large sample sizes often lead to significant χ^2 statistics even with adequate model fit. M_2 (Liu et al., 2016) is a limited-information fit index best used when there is sparseness in the contingency table of response patterns. This is because M_2 compares marginal response probabilities for small combinations of items instead of full response patterns, which are often sparse for an assessment with too many items or too few examinees. Statistically significant values of both statistics indicate poor model fit.

Results

AIC and BIC consistently preferred the correctly fitted model, indicating robustness of these statistics to this type of model misfit, each preferring the correct model 96% of the time. Mean RMSEA values were very small for both models, with the correctly fitted model showing a slightly lower distribution of values. Below is a table summarizing mean RMSEA distribution percentiles across all replications for both models.

Table 2. *Mean Item RMSEA Percentiles for Misfit and Correctly Specified Models*

Model	10%	25%	50%	75%	90%
Misfit	.0014	.0015	.0018	.0020	.0023
Correct	.0013	.0014	.0016	.0018	.0021

The difference in performance for the misfitted model and the correctly fitted model appears marginal, suggesting a weaker ability to detect this type of misfit.

Finally, I recorded the proportion of the 1,000 replications that resulted in rejections of Pearson's χ^2 test of model fit, as well as the M_2 statistic. The table below shows that the misfitted model produced significant statistics more frequently than the correct model for both indices. Additionally, M_2 rejected more often than χ^2 for both models. M_2 is more sensitive to local misfit, meaning that small violations of local independence may push M_2 past the rejection threshold but not χ^2.

Table 3. *Significance Rates for χ^2 and M_2 Goodness of Fit Tests*

Fit Index	Model	Significant	Non-Significant
χ^2	Misfit	.099	.901
	Correct	.003	.997
M_2	Misfit	.122	.878
	Correct	.037	.963

These results suggest that the misfitted model, which is not appropriately calibrated to IRT data, may not have adequately represented patterns and relationships in the item response data and did not properly model the underlying structure of the data. This is notable because researchers may not realize retrofitting has occurred if they did not develop the items on their own or otherwise do not know enough about the assessment framework. Those who are unaware of the initial structure of the data may be able to notice these poor model fit indices in a situation where retrofitting has occurred and change course if needed. Additionally, researchers who are intentionally retrofitting may want to pay close attention to fit indices and make note in their studies that the LCDM does not do very well with reproducing the structure of IRT data.

CHAPTER 5

EMPIRICAL ANALYSIS

Methods

Dataset

The final portion of this study was an empirical analysis of retrofitting conducted on a sample of item responses from a 12th-grade science assessment test (SAT), provided for use in the *mirt* package in R. These data comprise responses to 32 multiple-choice items from 600 12th-grade examinees, along with two scoring keys. The two keys were nearly identical except for item 32, which was changed after nominal response model analyses revealed that the initial key may have been incorrect. While the assessment intended to measure topics in chemistry, biology, and physics, an exploratory factor analysis (EFA) conducted in Mplus Version 8 (Muthén & Muthén, 2017) resulted in a two-factor model having better fit, so I used a two-factor model in the rest of the analysis.

Procedure

This analysis included both fitting a multidimensional 2-PL model and calibrating the LCDM with this sample of item responses to examine and compare the item parameters and the performance of model fit indices. For the LCDM, a Q-matrix was created based on maximum factor loadings from the two-factor EFA in Mplus. The Q-matrix used is shown in Table 2.

Table 4. *Q-Matrix Design for Empirical Analysis*

Item	α_1	α_2	Item	α_1	α_2
1	0	1	17	1	0
2	1	0	18	0	1
3	0	1	19	0	1
4	0	1	20	1	0
5	1	0	21	1	0
6	0	1	22	1	0
7	1	0	23	0	1
8	0	1	24	1	0
9	1	0	25	0	1
10	0	1	26	0	1
11	1	0	27	1	0
12	0	1	28	0	1
13	1	0	29	0	1
14	0	1	30	1	0
15	1	0	31	1	0
16	0	1	32	0	1

Item parameter estimates and examinee ability estimates were gathered from the MIRT model summary. Similarly, class membership estimates were gathered from the LCDM summary and then used to calculate correct response probabilities for masters and non-masters of each attribute. I used these probabilities to calculate DCM difficulty and discrimination, according to the definitions provided. Item difficulty estimates from the MIRT analysis were correlated with the parameters calculated in the LCDM analysis, and discrimination parameter estimates were similarly correlated between the two analyses. Lastly, I recorded relevant fit statistics to compare with the results of the second simulation.

Results

The IRT and DCM difficulty estimates had a very strong negative correlation ($r = -.948$), while the IRT and DCM discrimination estimates had a moderately weak positive correlation ($r = .279$). These results show a similar degree of item parameter recovery as the first simulation. Compared with the results of the simulation, the observed difficulty correlation is in line with what I might expect for items with any difficulty and medium or high discrimination, and the discrimination correlation is also what I might expect for highly discriminating items. After examining the item parameter estimates from the MIRT analysis, I noticed that most items from the assessment had medium to high initial discrimination estimates (mean discrimination 1.10).

For the second portion of the simulation, the MIRT model had better fit than the LCDM according to the same metrics used in the second simulation. AIC and BIC both preferred the correctly fitted model, and the mean RMSEA value for the MIRT model (.0251) was marginally smaller than for the LCDM (.0282).

Table 5. *Fit Indices for Correctly Specified and Misfitted Models*

Fit Index	Correct (MIRT)	Misfit (LCDM)
AIC	19280.81	19462.45
BIC	19566.61	19757.04
Mean RMSEA	.0251	.0282

Additionally, the M_2 test of model fit was significant at $\alpha = .05$, $M_2(463) = 629.43$, $p < .001$. Overall, these results further suggest that the measured fit indices are able to detect this type of model misfit.

CHAPTER 6

DISCUSSION

From the first simulation study, I concluded that the LCDM does a better job recovering item difficulty parameters than item discrimination but is not an adequate recreation of the data structures in an IRT model. The second simulation, comparing model fit indices for both a correctly fitted model and a misfitted model, showed AIC and BIC to be the best index of those examined for detecting this type of model misfit, while RMSEA was not consistently useful.

I also conducted an empirical analysis using an IRT-based dataset of science aptitude test responses. The first part of the empirical analysis showed similar results to those of the first simulation, producing a strong correlation between item difficulty parameters and a weak/moderate correlation between item discrimination parameters. The second part of the empirical analysis examined the model fit when both an IRT model and a DCM were fit to the same response data, ultimately supporting the results of the second simulation that showed AIC, BIC, and M_2 as useful detectors of this misfit.

Some potential limitations of this study include the simple-structure Q-matrix used in the simulations and the source of the data for the empirical analysis. Using a simple Q-matrix, in which each item measures only one attribute, means the study did not account for assessments with complex items. Complex items are much more likely to appear in assessments in practice, as one item rarely measures a single attribute of interest. The interpretation of the results,

including the nature and strength of the item parameter relationships, should not be generalized to assessments with complex items.

For the empirical analysis, the data originally came from the 2003 TESTFACT manual and were found in the *mirt* package in R. I could not locate detailed descriptions of the items, the examinees, or the assessment they came from, so I cannot guarantee that these data represent an assessment with proper design and implementation. Based on the context of the original data, and knowing that other researchers (Chalmers, 2012) have used it in the context of the *mirt* package as sample data, I felt comfortable moving forward with the analysis using this dataset.

Further research is needed in both contexts to establish any relationships or guideline criteria. I would like to see further research into the interaction observed between the IRT and DCM item discrimination parameters when the initial IRT parameters meet certain conditions. It appeared that there was a difference in the relationship of discrimination parameters with items initially low in difficulty and with medium discrimination, when compared with the other combinations of initial item parameters. Closer analysis may reveal more information about the nature or source of that difference. Additionally, more research into the performance of fit statistics in this retrofitting context would shed light on which indices are useful and why, as well as how they may perform in more realistic conditions. Further research is especially needed to justify any guidelines or thresholds for making model fit decisions.

Ultimately, my intent with these three studies in combination was to begin to improve the application and interpretation of DCMs in a retrofitting context. The overall goal was to contribute to the limited research on retrofitted DCMs and inspire other researchers to further examine the outcomes of this practice, which I believe could lead to a more widespread implementation of DCMs in educational research and assessment. If thoroughly understood and

properly conducted, the practice of retrofitting analysis may allow researchers to better understand behavior and properties of DCMs, such as how they classify examinees, while still using the widely available non-diagnostic item response data. Similarly, educators looking to use diagnostic models in their practice could benefit from the ability to use items banks that were not intended for diagnosing examinees. For example, formative feedback is highly desired among educators and researchers, and DCMs have the capability to provide that. The implementation of retrofitting DCMs to IRT data could make DCMs accessible to more individuals, companies, and schools, and I hope to meaningfully contribute to current literature surrounding this unique process.

REFERENCES

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland and H. Wainer (Eds.), *Differential Item Functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical Theories of Mental Test Scores*.

Bradshaw, L. P., & Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *International Journal of Testing*, *16*(2), 99–118.

Camilli, G. (2006). Test Fairness. In R. L. Linn (Ed.), *Educational measurement* (pp. 221–256). Westport, CT: Praeger Publishers.

Chalmers, R. Philip (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1-29. https://doi:10.18637/jss.v048.i06

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

George, A. C., Robitzsch, A., Kiefer, T., Gross, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*, *74*(2), 1–24. https://doi.org/10.18637/jss.v074.i02

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.

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in Cognitive Diagnostic Modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*(1), 98–125. https://doi.org/10.1111/jedm.12036

Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y.-H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing*, *16*(2), 119–141. https://doi.org/10.1080/15305058.2015.1133627

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. https://doi.org/10.1111/j.1745-3984.2011.00160.x

Li, X., & Wang, W. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement*, *52*(1), 28–54. https://doi.org/10.1111/jedm.12061

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2017). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement*, *78*(3), 357–383.

Liu, Y., Tian, W., & Xin, T. (2016). An Application of M2 Statistic to Evaluate the Fit of Cognitive Diagnostic Models. *Journal of Educational and Behavioral Statistics*, *41*(1), 3–26. https://doi.org/10.3102/1076998615621293

Ma W, de la Torre J (2020). GDINA: An R Package for Cognitive Diagnosis Modeling. *Journal of Statistical Software*, *93*(14), 1-26. https://doi.org/10.18637/jss.v093.i14

Madison, M. J. (2023, April 12-15). *The Effects of Measurement and Structural Model Misspecifications in Longitudinal Diagnostic Classification Models* [Paper session]. National Conference on Measurement in Education Annual Meeting, Chicago, IL.

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*, *75*(3), 491–511. https://doi.org/10.1177/0013164414539162

Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika*, *83*(4), 963–990. https://doi.org/10.1007/s11336-018-9638-5

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI: Journal of the National Cancer Institute*, *22*, 719–748. https://doi.org/10.1093/jnci/22.4.719

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

Muthén, L.K. and Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rupp, A. A., & Templin, J. (2007). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*(1), 78–96. https://doi.org/10.1177/0013164407301545

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

Sen, S., & Cohen, A. S. (2021). Sample size requirements for applying diagnostic classification models. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.621251

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates True Bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. https://doi.org/10.1007/bf02294572

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

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of Diagnostic Classification Model Examinee estimates. *Journal of Classification*, *30*(2), 251–275. https://doi.org/10.1007/s00357-013-9129-4

Woods, R., Wilson, D. T., Gibbons, R. D., Schilling, S. G., Muraki, E., & Bock, R. D. (2003). TESTFACT 4 for Windows: Test Scoring, Item Statistics, and Full-information Item Factor Analysis [Computer software]. Lincolnwood, IL: Scientific Software International

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, 125-145.