The Performance of Local Dependence Indices with Psychological Data

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy
in the Graduate School of The Ohio State University

By
Carrie R. Houts, M.S.
Graduate Program in Psychology

The Ohio State University
2011

Dissertation Committee:
Michael C. Edwards, advisor
Robert Cudeck
Thomas E. Nygren
Copyrighted by

Carrie R. Houts

2011
Abstract

The violation of the assumption of LI, known as local dependence (LD), when applying Item Response Theory (IRT) models has been shown to have a negative impact on all estimates from the model, specifically item-level, person-level, and test-level values. For this reason, numerous indices and statistics have been proposed to aid analysts in the detection of LD. A large-scale simulation study was conducted to evaluate the relative performance of selected LD measures, as measured by Type I error rate and power, as well as the parameter recovery for items in datasets exhibiting LD. Measures were assessed at differing levels of test length, number of respondents, type of LD present (i.e., none, surface, underlying), fitted IRT model, degree of LD exhibited by item pairs, and number of item pairs exhibiting LD. The levels of the independent variables were chosen to replicate conditions more typical of studies involving psychological assessment, as opposed to educational measurement.

Parameter estimates when LD was present were negatively affected to such a degree as to support the recommendation to use LD indices as part of any preliminary analyses. Previous findings regarding differences between surface and underlying LD and the conformance of selected indices to their theoretical Null distributions were replicated. No LD index displayed the best performance across all conditions; two indices displayed superior performance in terms of both power and Type I error but are only available when the data is strictly dichotomous in nature. Two other LD measures are available
across the two fitted IRT models examined and displayed adequate to good performance in most simulation conditions. The use of these indices in tandem is the final general recommendation for applied researchers. Areas of future investigation, including the observed power differences of LD indices when detecting surface versus underlying local dependence, are identified.
Dedication

To Alice the Cat
Acknowledgments

I would like to thank my advisor, Dr. Michael Edwards, for his time, patience, support and guidance in my graduate studies and this endeavor in particular. I will continue to wear my Team Edwards t-shirt with pride.

My thanks to Dr. Nancy Betz, Dr. Bob Cudeck, and Dr. Tom Nygren for their time and feedback in improving the quality of this work. Additionally, I acknowledge Casey Blalock and Kenny Olson for their assistance in trouble-shooting R code when, after hours of hair-pulling, it still wouldn’t work; their help kept me from having to explain why my monitor was embedded in drywall.
Vita

2002 B.A. ......................................... Psychology, Rhodes College

2006 M.S. ......................................... Psychology, University of Memphis

2006 to present ................................... Graduate Assistant, Department of Psychology, The Ohio State University

Publications


Fields of Study

Major Field: Psychology
Table of Contents

Abstract ........................................................................................................................................ ii

Dedication .................................................................................................................................... iv

Acknowledgments ....................................................................................................................... v

Vita ................................................................................................................................................ vi

Table of Contents .......................................................................................................................... vii

List of Tables ................................................................................................................................. xi

List of Figures ................................................................................................................................ xiii

Chapter 1: Introduction ...................................................................................................................... 1

1.1: Effects of Ignoring Local Dependence ...................................................................................... 6

1.2: IRT Models Relevant to Psychological Assessment .............................................................. 8

1.2.1: Two-parameter logistic model ............................................................................................ 9

1.2.2: Graded response model ....................................................................................................... 9

Chapter 2: Local Dependence Indices ............................................................................................ 11

2.1: IRT-based indices .................................................................................................................... 13

2.1.1: Q3 ......................................................................................................................................... 13

2.1.2: Fisher’s r-to-z transformed Q3 ............................................................................................ 14
2.1.3: Likelihood ratio statistic – $G^2$ ................................................................. 15

2.1.4: Jackknifed slope index. ................................................................................. 17

2.2: Covariance Structure Based Indices ................................................................. 18

2.2.1: NOHARM-based $\chi^2_{D}^2$ ........................................................................ 18

2.2.2: Confirmatory factor analysis modification indices. ..................................... 20

Chapter 3: Method ..................................................................................................... 23

3.2: Latent Construct and Item Parameters ............................................................. 23

3.3: LD type and Degree of LD ................................................................................ 24

3.3.1: Surface local dependence. ............................................................................. 24

3.3.2: Underlying local dependence. ....................................................................... 24

3.4: Data Generation and Validation ..................................................................... 26

3.5: Parameter Estimation ....................................................................................... 27

3.6: Calculation and Validation of LD Indices ....................................................... 27

Chapter 4: Two Parameter Logistic Model Results ...................................................... 29

4.1: Parameter recovery ........................................................................................... 30

4.1.1: No local dependence....................................................................................... 31

4.1.2: Surface local dependence. ............................................................................ 31

4.1.3: Underlying local dependence. ..................................................................... 33

4.1.4: Overall comparisons .................................................................................... 35
List of Tables

Table 1a,b. Example observed and model-predicted contingency tables ......................70
Table 2a,b. Example raw and summed JSI value matrices.............................................71
Table 3. Independent variables manipulated in simulation study..................................72
Table 4. RMSE for Null 2PLM by items and sample size.............................................73
Table 5. RMSE for cells SLD 2PLM, $k = 10$, $N = 250$, $\pi_{LD} = 0.8$, $p = 1, 2, \text{ and } 4$ ....74
Table 6. Group means for the significant items by degree interaction for 2PLM $\alpha$-parameter RMSE..................................................................................................................75
Table 7. RMSE for cells ULD 2PLM, $k = 40$, $N = 250$, slope ratio = 1.5, $p = 1, 2, \text{ and } 4$ ..........................................................76
Table 8. Group means for the significant item by degree interaction for 2PLM $\alpha$-parameter RMSE..................................................................................................................77
Table 9. Type I error rates for Null 2PLM by number of items and sample size ..........78
Table 10. Means and SDs of LD indices using distributional critical values for testing in 2PLM Null cells ..............................................................................................................79
Table 11. Overall power and Type I error rates for SLD and ULD 2PLM cells ..............80
Table 12. Prominent F values for predictors of power and Type I error, in conjunction with the 2PLM, of LD indices with SLD.................................................................81
Table 13. Prominent predictors of power and Type I error of LD indices, in conjunction with the 2PLM and ULD .................................................................82
Table 14. RMSE for Null GRM by items and sample size

Table 15. RMSE for cells SLD GRM, k = 10, N = 250, π_{LD} = 0.8, p = 1, 2, and 4

Table 16. RMSE for cells ULD GRM k = 40, N = 250, slope ratio = 1.5, p = 1, 2 and 4

Table 17. Type I error rates for Null GRM by items and sample size

Table 18. Means and SDs of LD indices using distributional critical values for testing in GRM Null cell

Table 19. Overall power and Type I error rates for SLD and ULD GRM cells

Table 20. Prominent predictors of power and Type I error of LD indices with SLD

Table 21. Prominent predictors of power and Type I error of LD indices with ULD

Table 22. Average polychoric correlations for GRM cells with 1 LD item pair exhibiting the highest level of either type of LD

Table 23. Average mean and standard deviation (from 50 randomly selected replications) of summed JSI values in selected 2PLM cells

Table 24. 2PLM cells with greater than 5% discard rates

Table 25. Overall power and Type I error for LD indices with 2PLM when problem cells are excluded
List of Figures

Figure 1. CFA path diagram for MIs explication ...............................................................95

Figure 2. Alternate (and equivalent) CFA path diagram for MIs explication, highlighting
IRT-based conceptualization .........................................................................................96

Figure 3. Recovered discrimination parameter distributions for LD and Null items in the
2PLM SLD cell with 10 items, N = 250, and π_{LD} = 0.8 ........................................97

Figure 4. Average RMSE values for 2PLM discrimination parameters, varying across
degree of SLD and separated by number of items .................................................98

Figure 5. Recovered discrimination parameter distributions for LD and Null items in the
2PLM ULD cell with 40 items, N = 250, and slope ratio = 1.5.................................99

Figure 6. Average RMSE values for 2PLM discrimination parameters, varying across
degree of LD and number of items, separated by type of LD....................................100

Figure 7. Average power (upper portion) and Type I error rate (lower portion) for LD
indices, in conjunction with the 2PLM and SLD, by number of items ....................101

Figure 8. Average power and Type I error for LD indices, in conjunction with the 2PLM
and SLD, by degree of LD .........................................................................................102

Figure 9. Average power for LD indices, in conjunction with the 2PLM and SLD, by
number of items and degree of LD ............................................................................103

Figure 10. Average power for LD indices, in conjunction with the 2PLM and ULD, by
degree of LD .................................................................................................................104
Figure 11. Average Type I error rate for LD indices, in conjunction with the 2PLM and ULD, by degree of LD and number of items .................................................................105

Figure 12. Average power for LD indices from overall analysis, in conjunction with the 2PLM, by type and degree of LD .................................................................................106

Figure 13. Recovered discrimination parameter distributions for LD and Null items in the GRM SLD cell with 10 items, N = 250, and \( \pi_{LD} = 0.8 \) .................................................................107

Figure 14. Average RMSE values for GRM discrimination parameters, varying across degree of SLD and separated by number of items .................................................................108

Figure 15. Recovered discrimination parameter distributions for LD and Null items in the GRM ULD cell with 40 items, N = 250, and slope ratio = 1.5 .........................................................109

Figure 16. Average RMSE values for selected GRM item parameters, varying across degree of LD and separated by type of LD .............................................................................110

Figure 17. Average RMSE values for all GRM item parameters, separated by LD type .................................................................................................................................111

Figure 18. Average power for LD indices, in conjunction with the GRM, by sample size and degree of SLD .........................................................................................................112

Figure 19. Average power for LD indices, in conjunction with the GRM, by number of items and degree of SLD ...........................................................................................................113

Figure 20. Average Type I error for LD indices, in conjunction with the GRM, by number of items and degree of SLD ...........................................................................................................114

Figure 21. Power of LD indices, in conjunction with the GRM, by degree of ULD ......115

xiv
Figure 22. Type I error of LD indices, in conjunction with the GRM and ULD, by number of items and locally dependent pairs .................................................................116

Figure 23. Power of the JSI with the GRM and 2PLM, separated by number of items and number of LD pairs ...........................................................................................................117

Figure 24. Summarized discard rates for the 171 2PLM simulation cells.................118
Chapter 1: Introduction

Item response theory (IRT), a collection of latent variable models, is becoming increasingly popular in the social sciences. Although it is a less familiar family of techniques to psychologists outside of psychometrics, researchers are more frequently using IRT over traditional procedures as they are introduced to it and become aware of the desirable benefits it can provide. However, for the benefits of IRT to hold in practice, strong model assumptions must be satisfied – or at least not violated too badly. As more social scientists, especially those less familiar with IRT, use these models for constructing and assessing scales it becomes increasingly important that the assumptions made by the models are communicated to analysts and the consequences of violating these assumptions are known. Relatedly, the myriad tools for checking such assumptions need to be documented and their strengths and weaknesses well understood, in both an absolute and relative sense.

In detailing the mathematical basis of models that are now known as IRT, Lord and Novick (1968; Chp. 16) discussed two critical assumptions of all IRT models: the chosen mathematical function describes the data well and the item responses exhibit local independence (LI). While the appropriateness of the function is an important assumption, the validity of which should be investigated during any IRT analysis, the main focus of this discussion will cover the assumption of LI among items.
LI requires that item responses be unrelated to one another (i.e., independent) after conditioning on the latent variables. That is, LI, in the case of an $k$-item scale, is when the conditional probability of a given response pattern for the $k$ items is equal to the product of the conditional probability for each of the $k$ individual item responses.

Mathematically, this may be expressed as

\[ P(X_1 = x_1, X_2 = x_2, ..., X_k = x_k \mid \theta) = P(X_1 = x_1 \mid \theta) \times P(X_2 = x_2 \mid \theta) \times ... \times P(X_k = x_k \mid \theta), \]

where $x_i$ is the observed response to the $i^{th}$ item and $\theta$ is a vector containing the latent construct(s). Meeting the assumption of LI implies that the latent variables are able to model all covariance among item responses, that the item responses are related to each other only through the latent variables included in $\theta$.

In psychological research, it is generally desired that a scale require only one latent construct; this goal is so prevalent that unidimensionality, a special case of local independence that occurs when there is only one common factor, is also typically assumed when IRT models are applied to psychological data. In the case of unidimensional models, LI implies that a single latent factor is able to account for all of the covariance among the item responses. If, after fitting a unidimensional model, the item responses conditioned on the primary latent trait level still show any degree of relatedness, the assumption of LI has not been met. To check the assumption of LI given by Equation 1 it would be necessary to examine all possible pairs, triplets, and higher order combinations of items for covariance unaccounted for by the fitted model. The violation of LI, logically termed local dependence (LD), may imply that there is more than one latent variable underlying the observed responses and, therefore, the assumption
of unidimensionality would also be violated. When a model with multiple latent constructs is necessary to meet the assumption of LI, the additional abilities may not be conceptually significant to the researcher and can result from nuisance factors such as a common stimulus passage, speededness, fatigue, or rater effects (Yen, 1993).

The representation of LI given in Equation 1 is often called the strong form of LI (e.g., McDonald, 1981). It is quite restrictive and would never be met in practice, making it an unreasonable assumption to make or empirically test. Several authors (e.g., McDonald, 1994; Stout, 2002) have suggested that rather than using the definition of LI provided by Equation 1, a less stringent definition of LI, called weak LI, be assumed; namely

\[
P(X_i = x_i, X_j = x_j | \theta) \\
= P(X_i = x_i | \theta) \times P(X_j = x_j | \theta) \quad \text{for all } i \neq j, \tag{2}
\]

where all terms are as previously defined. Rather than assuming that every possible higher order set of items is locally independent, only pairs of items (of which there are \( \binom{k}{2} \) pairs for a \( k \) item scale) need be locally independent. This relaxed assumption makes checking for LD among items computationally less expensive and the assumption of LI more likely to hold in practice. Because of the associated computational savings, weak LI has been adopted as the assumption which needs be met and tools developed to check this assumption, typically called LD indices, focus only on item-pairs; further references to LI should be interpreted as referring to weak LI, as defined in Equation 2.

An overview of LI and LD also needs to distinguish between the possible sources of over-relatedness within item pairs. Chen and Thissen (1997) were the first to
distinguish between LD that is a result of additional unmodeled latent variables, which they termed underlying local dependence (ULD), and LD that is due to highly similar content or placement, what they called surface local dependence (SLD).

As described in Chen and Thissen (1997), SLD is created by a deterministic process where respondents answer two questions identically because the items are similar, either in content or location on the instrument. For example, a childhood autism scale has an item concerning the child’s distractibility and a separate item assessing their difficulty in paying attention. Although there may be meaningful clinical distinctions between these items, based on observed response patterns, many parents respond to the two items as if the questions were identical.

The degree of similarity is denoted by \( \pi_{LD} \), the probability that the second item in the LD pair will have a response identical to the first item without regard to level of the latent ability. More specifically, the SLD process for an item pair is:

With probability \( \pi_{LD} \):

\[
\text{Response to Item 2} = \begin{cases} 
1, & \text{if } X_1 = 1 \\
0, & \text{if } X_1 = 0 
\end{cases}
\]

With probability \( 1 - \pi_{LD} \):

\[
\text{Response to Item 2} = \begin{cases} 
1, & \text{with } P(X_2 = 1|\theta) \\
0, & \text{with } P(X_2 = 0|\theta) 
\end{cases}
\]

As seen in the upper section of three (3), with a certain probably the response to the second item of a pair of LD items does not rely on the underlying process implied by the IRT model, which involves \( \theta \). For example, a manifestation of SLD is parents who respond “agree” to the distractibility item of the autism instrument, and then, with probability \( \pi_{LD} \), interpret a “difficulty paying attention” item as a restatement of the
distractibility question and put “agree” as well, in an effort to be consistent across questions, rather than responding “agree” because their child’s actual issues with paying attention warrant an “agree” response. Another situation in which SLD may be observed is with fatigued research participants; if, by the end of an extensive psychological screening session, some proportion of participants fail to fully consider their answer to the final item of a depression scale and simply copy their “Strongly disagree” response for the previous item to finish as quickly as possible, the item responses would be highly related, not because of participants’ innate depression levels, but because those two items happened to appear at the end of the session.

ULD occurs when too few latent variables are used to model data; in factor analysis verbiage, ULD stems from under-factoring. Unlike SLD, the process implied by an IRT model is still at work in item responses exhibiting ULD and the LD arises because the number of dimensions being modeled is insufficient to account for the common variance among the items. The ULD model, as applied to the weak local dependence definition, posits that within each locally dependent item pair, there is an underlying trait that those two items share that is not common to other items in the scale. Borrowing the terminology commonly applied to the bi-factor model (e.g., Gibbons & Hedeker, 1992), all items from an assessment instrument will load on the general factor (i.e., the construct of primary interest) and item pairs exhibiting ULD will also load on an additional specific latent variable for which those two items are the only indicators. An example of ULD is a physical functioning scale that includes an item about the respondent’s ability to climb one flight of stairs and a separate item regarding their ability to climb several flights of
stairs; in contrast to items with SLD, these “stair climbing” items would not readily be interpreted as soliciting duplicate information but they are clearly related to one another and are most likely more similar to each other than either stair item would be to an additional item regarding some other aspect of general physical functioning.

1.1: Effects of Ignoring Local Dependence

Research regarding the effects of ignoring LD when it is present has revealed negative consequences for the estimates obtained from IRT models (e.g., Ackerman, 1987; Reese, 1995; Wainer, 1995; Wainer & Thissen, 1996). Detrimental effects have been documented for item parameter estimates, individual scale scores, and estimates of the overall information/reliability of a test. Although the relevant research for these various estimates are discussed individually, it should be noted that item parameter estimates are used in the estimation of individual’s level of the latent trait and test information, so any effect ignored LD has on the item-level estimates will also appear in the person- and test-level estimates.

Several studies (e.g., Ackerman & Spray, 1987; Chen & Thissen, 1997; Reese, 1995) have found that ignoring local dependence results in biased item parameter estimates. Both Ackerman (1995) and Reese (1995) found that as the level of LD increases, so does the bias observed in the parameter estimates. Specifically, in both studies, high levels of unaccounted LD resulted in overestimated discrimination values and underestimated severity values for items in locally dependent pairs. In contrast, Chen and Thissen (1997), while also finding that LD affects item parameters estimates, provided simulated data examples in which the bias seen in item parameter estimates had
no predictable direction. Additionally, the effect on item parameter estimates differed depending on whether it was ULD or SLD that was present within the data.

With respect to latent construct estimates, Reese (1995) found that in the presence of high levels of local dependence, as induced by simulating data with a large value on a common LD index, latent trait estimates for low-level respondents were underestimated while overestimation was observed in high-level respondents and also noted that the rank-ordering of respondents may also be affected by extreme LD. In a similar vein, Zenisky, Hambleton, & Sireci (2002) analyzed data from the MCAT and found that individual latent construct estimates from a standard IRT model, which ignored LI, and those from a testlet-based model, which accounts for local dependence among items by using sum scores from groups of locally dependent items, differed by as much as a full standard deviation. As noted by Zenisky, Hambleton, & Sireci (2002), a difference of that size, which is due primarily to modeling choices, could radically impact an individual respondent’s outcome.

Numerous studies (e.g., Thissen, Steinberg, & Moody, 1989; Thompson & Pommerich, 1996; Wainer & Thissen, 1996) have also examined the effect that violating the LI assumption can have on the assessment of the overall quality of a scale. The consensus from such studies is that, regardless of whether one is using a Classical Test Theory measure of reliability (e.g., coefficient alpha; Cronbach, 1947) or the IRT-based information function, the presence of LD among item pairs will inflate either estimate, causing the researcher to find the scores to be more informative or more reliable than it actually is. For example, Sireci, Thissen, & Wainer (1991) found IRT information values
to be overestimated by 10 - 15 % and coefficient alpha to be overestimated by a 0.07 to 0.12 when ULD is present. Although the majority of previous research indicates that the reliability of scores from a scale will be over-estimated, it is of note that these studies used a testlet-based approach to deal with any LD, meaning they summed locally dependent items to deal with the excess covariance, effectively reducing the number of items as well as accounting for excess inter-item dependence. It is known that coefficient alpha decreases as scale length decreases (Schmitt, 1996) and the extent to which the previous findings are informed by this item number decrease is unknown. Research on large scale testing data performed at The Ohio State University found the use of multidimensional IRT models, which retain all scale items and explicitly model LD, resulted in increased reliability estimates.

1.2: IRT Models Relevant to Psychological Assessment

Previous research (e.g., Finch & Habing, 2007) found that the IRT model used to fit data affected the performance of several of the LD indices they studied, making choice of model an important point of consideration when assessing LD diagnostics. Although there are a variety of IRT models that may be used to fit scaling data, the 2-parameter logistic model (2PLM) and the graded response model (GRM) will be the primary focus of this study. These models are selected for their suitability to data obtained from psychological scales, such as a self-esteem measure or depression inventory, where information is most often solicited in the form of dichotomous, “Yes/ No” responses or Likert-style responses, typically with 3 to 7 response options from which a respondent
may choose. Because psychological assessment is primarily self-report data, guessing is assumed to be absent.

1.2.1: Two-parameter logistic model. A common IRT model based on the logistic function estimates the probability of endorsing the $j^{th}$ dichotomously-scored item, conditional on the level of the latent trait, as a function of both a severity and discrimination parameter. The 2PLM may be expressed as

$$P(x_j = 1|\theta) = \frac{1}{1 + \exp \left( -a_j(\theta - b_j) \right)},$$

(4)

where $a_j$ is the discrimination parameter that details how well item $j$ separates people of lower and higher latent trait estimates, $b_j$ is the severity parameter, interpreted as the level of the latent trait at which a respondent has a 50% chance of endorsing item $j$, and $\theta$ is the latent construct parameter typically assumed to follow a standard Normal distribution. The mentioned severity parameter, $b_j$, is commonly referred to as a difficulty parameter when working with achievement constructs, such as math ability. Severity is preferred to highlight the psychological focus of this work, in that the endorsing of statements indicates severity of symptomology, such as for depression. The discrimination parameter is also known as the slope parameter; discrimination parameter and slope will be used interchangeably.

1.2.2: Graded response model. Using the 2PLM as a starting point, Samejima (1969) introduced an IRT model that is able to fit items that have two or more ordered response categories; when the number of categories equals two, the GRM simplifies into the 2PLM. For an item with $C$ categories, the probability of responding in the $c^{th}$ category is
\[ P(u = c) = \frac{1}{1 + \exp\left(-a_j(\theta - b_{jc})\right)} - \frac{1}{1 + \exp\left(-a_j(\theta - b_{j(c+1)})\right)} \]
\[ = P^*(c) - P^*(c + 1), \]  

where \( a_j \) is again the discrimination parameter that details how well the \( j^{th} \) item separates people of lower and higher latent trait levels, \( b_c \) is the threshold for category \( c \), which marks the latent construct level where the probability of scoring in \( c^{th} \) category or higher reaches 50%, and \( \theta \) is defined as before. To facilitate computation, it is specified that \( P^*(0) = 1 \) and \( P^*(C+1) = 0 \). To find the probability of being in the \( c^{th} \) category, the GRM finds the probability of being in that category or higher and subtracts from that the combined probability of being in any of the higher categories. Each item has one discrimination parameter and \( C-1 \) threshold parameters.
Chapter 2: Local Dependence Indices

As has been discussed, there may be severe consequences for failing to address LD during scale construction and evaluation. These failings could be particularly harmful if the scores from the scale will be used to make individual, high-stakes decisions. The importance of identifying items exhibiting LD has been known for some time and numerous indices and statistics have been proposed for detecting LD; a literature search found over twenty indices or statistics for detecting LD (e.g., Chen & Thissen, 1997; Cressie & Reed, 1984; Gesseroli & DeChamplain, 1996; Glas, 1999; Ip, 2001; Rosenbaum, 1984; Stout, 1987; Tsai & Hu, 2005; Yen, 1984, 1993; Zhang & Stout, 1999.

A non-exhaustive list of indices and statistics proposed or adopted as LD measures in the IRT literature includes Yen’s (1984) $Q_1$, $Q_2$, and $Q_3$, Stout’s DIMTEST (1987) and DETECT (Zhang & Stout, 1999) procedures, Chen and Thissen’s (1997) use of Pearson’s $\chi^2$, the likelihood ratio statistic $G^2$, the standardized log-odds ratio difference, and the standardized $\phi$ coefficient difference, Glas’ (1999) proposed use of Lagrangian multipliers, Ip’s (2001) suggestion to use the Mantel-Haenszel test with multiple testing corrections, Tsai and Hsu’s (2005) information entropy based measure AMID, and Gessaroli and de Champlain’s (1996) NOHARM-based $\chi^2$. Although not all LD measures have been subjected to comparative study, some existing recommendations may be found. For instance, Kim (2007) found $Q_3$ to perform well in comparison to nine
other LD measures, Gessaroli and de Champlain (1996) reported performance of their NOHARM-based \( \chi^2 \) that was superior to that of both DIMTEST and DETECT, and Chen and Thissen (1997) recommended, from among the five measures they investigated, the use of \( G^2 \) and \( Q_3 \).

The current study will provide an evaluation of several of these more promising LD indices using simulated data and, in so doing, codify conditions under which the selected indices appear to be especially adept at LD detection. While previous research has comparatively evaluated LD indices (e.g., Chen & Thissen, 1997; Kim, 2007), such projects have typically used conditions involving extremely large samples sizes, large numbers of items that are scored dichotomously, and generating and fitted IRT models that include a guessing parameter; these conditions were constructed to mirror data that is most commonly obtained during educational testing situations. The current evaluation of LD indices will focus on simulating conditions that more closely mirror psychological assessment conditions. While there are obviously cases of psychological instruments collected on large samples of individuals or educational tests with few items, “psychological assessment” is used here as a short-hand for data collected on smaller samples, instruments which are composed of a relatively small number of items, and item responses that do not include guessing and may be assessed on a response scale that include more than two possible response options (e.g., Likert-type scales). As several of these conditions have yet to be addressed in the literature with respect to LD measures, particularly small samples size and use of the GRM, this study will expand our knowledge of LD measures and provide results that are more relatable to common
practice, especially to psychologists who typically employ smaller samples and use items that are scored polytomously.

As noted, there is an abundance of LD diagnostics which could be included in a comprehensive evaluation. To keep this evaluative project a manageable size, only a subset of available LD measures are included. The decision to include a measure is based on the index: (a) being widely used in practice; (b) occurring frequently in the literature (i.e., if an applied researcher searched for LD measures to use, it would likely be encountered); (c) being easily accessible (i.e., readily available in software packages or simply programmed); and (d) providing values specific to each pair of items, as opposed to only an overall test. Computational details and previous research findings regarding the indices selected for inclusion are presented.

2.1: IRT-based indices

2.1.1: Q3. Yen’s Q3 (1984, 1993) compares the observed responses, found the raw data, and the expected responses, those predicted by the fitted IRT model, for each respondent and calculates the deviation between the two values as

\[ d_{ja} = x_{ja} - E_{ja}, \]

where \( x_{ja} \) is the \( a^{th} \) student’s observed response to item \( j \) and \( E_{ja} \) is the student’s model predicted “true score” response to item \( j \), meaning \( E_{ja} \) may take non-integer values. The LD index \( Q_{3ij} \) is then computed as the Pearson product-moment correlation between the deviation scores of items \( i \) and \( j \); specifically,

\[ Q_{3ij} = r_{d小编}, \]
It has been reported that common practice is to use a constant cut-point of .2 to determine the presence of LD (e.g., Chen & Thissen, 1997; Tate, 2003) although it has been suggested that this value results in low power and simulations should be conducted to empirically determine an optimal cut point for a given sample size and test length (Chen & Thissen, 1997). Additionally, Reese (1995) noted that $Q_3$ tended to underestimate the level of local dependence between items that were simulated to have medium to high levels of local dependence. These known issues call for further investigation of the measure, as it is widely used in practice.

2.1.2: Fisher’s r-to-z transformed $Q_3$. To enable normal theory hypothesis testing, $Q_3$ can be subjected to Fisher’s r-to-z formula and standardized using the expected value and variance of

\[ E(z) = .5 \ln \frac{1+E(Q_3)}{1-E(Q_3)} \quad \text{Var}(z) = \frac{1}{N-3}, \]

where $E(Q_3) = -1 / (k-1)$, with $k$ denoting the total number of items and $N$ is the number of respondents (Yen, 1984, 1993).

In a primarily descriptive evaluation of $Q_3$, Yen (1984) found her index to be a useful tool in indentifying locally dependent items when compared to evaluating item content and Kim (2007) concluded the transformed $Q_3$ is a powerful LD index, although it was found to have a large Type I error rate for scales with fewer than 60 items. Several authors (e.g., Chen & Thissen, 1997; Glas & Falcón, 2003; Ip, 2001) have noted theoretical and empirical issues with the assumed null distribution used for testing with
the transformed $Q_3$, calling into question the usefulness of the measure, although it is used frequently in practice.

2.1.3: Likelihood ratio statistic – $G^2$. Prior to the specific formulations of the likelihood ratio statistic, the construction of expected and observed item response tables is detailed, as these values are necessary for the calculation of $G^2$. For each pair of items, it is possible to summarize the responses into a $p \times q$ contingency table, such as Table 1 in which item $i$ has $p = 2$ response categories and item $j$ has $q = 3$ possible responses; entry $O_{ij}$ is the observed number of respondents simultaneously in response category $p$ of item $i$ and response category $q$ of item $j$.

The expected responses, as predicted by the fitted IRT model, may be obtained from

$$E_{pq} = N \int P_{ip}(\theta) P_{jq}(\theta) f(\theta) \, d\theta,$$

where $N$ is the total number of respondents, $f(\theta)$ is the latent trait distribution (typically assumed to be standard Normal), $P_{ip}(\theta)$ is the probability that an examinee with latent trait level $\theta$ is in the $p^{th}$ category of the $i^{th}$ item, and $P_{jq}(\theta)$ is the probability that an examinee with latent trait level $\theta$ is in the $q^{th}$ category of the $j^{th}$ item. In the case of dichotomously scored items, Equation 9 may also be written as

$$E_{pq} = N \int P_i(\theta)^p P_j(\theta)^q [1 - P_i(\theta)]^{(1-p)} [1 - P_j(\theta)]^{(1-q)} f(\theta) \, d\theta.$$
As shown in Table 1b, a contingency table may then be constructed for each item pair with cell values representing expected responses, where $E_{pq}$ is the expected number of respondents simultaneously in response category $p$ of item $i$ and response category $q$ of item $j$.

A comparison of the observed responses to the expected responses is the basis of the likelihood ratio statistic, which has a long history in categorical data analysis and the properties of which have been detailed extensively in Bishop, Feinberg, and Holland (2007). The use of $G^2$ as a method of LD detection was introduced into the IRT literature by Chen and Thissen (1997); in this context, the likelihood ratio statistic will compare the possibly locally dependent observed responses to the expected responses, as predicted by the fitted IRT model which assumes LI.

Using both the observed and expected responses for all items pairs, the likelihood ratio statistic is computed as:

$$G^2 = -2 \sum_{i=1}^{p} \sum_{j=1}^{q} O_{ij} \ln \left( \frac{E_{ij}}{O_{ij}} \right).$$

(11)

$G^2$ is $\chi^2$ distributed with degrees of freedom equal to the number of cells in the contingency table minus the number of log-linear parameters; in the case of dichotomous item pairs, df = 1. Following Chen and Thissen (1997), if an observed cell is empty (e.g., $O_{ij} = 0$), the contribution to $G^2$ from that cell, which from the formula would be undefined, is set to zero. Simulation studies conducted by Chen and Thissen (1997) found $G^2$ to be a reasonable contender, in terms of power and Type I error rate, for detecting
local dependence when compared to several other LD measures. However, more recent research (Kim, 2007) found the null distribution of $G^2$ to be a poor approximation of the theoretical $\chi^2$ null distribution, a problem not encountered by Chen and Thissen (1997), highlighting the need for further investigation in the performance this index.

2.1.4: Jackknifed slope index. A recently developed index for LD detection, introduced by Edwards and Cai (in prep) is based on the observation that locally dependent items often exhibit inflated slope/discrimination parameters. Using this phenomenon as a basis for a jackknife-type procedure, Edwards and Cai suggest obtaining item parameter estimates for a full data set including all items and, in the subsequent steps, removing one item to obtain revised item parameter estimates. More concretely put, once estimates are obtained from a dataset containing all items, the first item is temporarily removed. Parameter estimates from this reduced dataset are obtained and JSI values calculated for all items remaining in the reduced dataset. Item 1 is then returned to the dataset, item 2 is removed and the process begins again, repeating until all possible datasets containing k-1 items have been created and analyzed in turn.

The jackknifed slope index (JSI) for the $j^{th}$ item, as impacted by the removal of the $k^{th}$ item, is then calculated as

$$JSI_{j(k)} = \frac{a_j - a_{j(k)}}{se(a_{j(k)})},$$

(12)

where $a_j$ is the full data slope estimate, $a_{j(k)}$ is the slope estimate of item $j$ with item $k$ removed from the data set and $se(a_{j(k)})$ is the standard error of the re-estimated slope parameter. Preliminary studies have found the JSI detects LD resulting from both SLD
and ULD processes, although it is more adept at SLD detection, and it is effective with even a moderate sample size (N = 300).

The preliminary research regarding the JSI used visual inspection of a $k$ by $k$ matrix (where, as before, $k$ is the number of items) with missing diagonal elements. To automate the assessment of JSI values an alternate procedure was developed. As could be deduced from Equation 12, for each pair of items a JSI value was calculated for the slope change in the first item, induced by removing the second item, as well as the slope change in the second item, induced by removing the first item. Since the purpose of the measure is to identify possibly dependent item pairs, the decision was made to sum these corresponding JSI values, resulting in a lower triangular matrix with $(k)^{(k-1)}$ elements. For clarity, Table 2a and b provide a simple example of this process. These elements were then tested against an ad hoc value of the mean of the lower triangular elements plus 2 times the standard deviation of those same elements, in which positive summed JSI values more extreme than (mean + 2*SD) were flagged as locally dependent and values less than or equal to that value were considered locally independent. In the example in Table 2, the mean summed JSI value is 0.3 and the standard deviation is 1.6 leading to summed JSI values being tested against 3.5 and values greater than this being flagged as a pair of items exhibiting LD. This is an untested cut-off and one goal of the work described in this document is to gauge the effectiveness of this procedure.

2.2: Covariance Structure Based Indices

2.2.1: NOHARM-based $X^2_{G/D}$.  McDonald (e.g., 1967, 1997) developed a non-linear factor analysis method for dichotomous item responses that has been implemented
in the NOHARM program (e.g., Fraser and McDonald, 1988). Gessaroli and DeChamplain (1996) introduced an approximate $\chi^2$ statistic that provides an overall assessment of dimensionality by testing the null hypothesis that all off-diagonal values of the residual correlation matrix are equal to zero. Gessaroli and DeChamplain’s statistic ($\chi^2_{G/D}$) is calculated by finding the residual correlations for all possible pairs of items, transforming those residual correlations into Fisher’s $z$ values, and then a final calculation that includes summing over the $z$ values. Specifically, it is obtained by:

$$r_{ij} = \frac{p_{ij}^r}{\sqrt{p_i (1 - p_i) p_j (1 - p_j)}}$$

(13)

$$z_{ij} = .5 \ln(1 + r_{ij}) - .5 \ln(1 - r_{ij})$$

(14)

$$\chi^2_{G/D} = (N - 3) \sum_{i=1}^{k} \sum_{j=1}^{i-1} z_{ij}^2,$$

(15)

where $p_{ij}^r$ is the residual joint proportion of respondents who endorsed both item $i$ and item $j$ (i.e., for item pair $(i,j)$, $p_{ij}^r$ is the discrepancy between the observed proportion of examinees who endorsed both items and the NOHARM model predicted joint proportion which may be found at off-diagonal entry $(i,j)$ in the NOHARM produced residual matrix) and $p_i$ and $p_j$ are the observed proportion of respondents who endorsed items $i$ and $j$, respectively. $\chi^2_{G/D}$ is approximately distributed as a central $\chi^2$ with df = $.5 k (k-1) - t$, where $k$ is the number of items, as before, and $t$ is the total number of estimated parameters; in fitting a one-dimensional model $t$ will equal the number of items.

Previous research has found $\chi^2_{G/D}$ to have adequate power (typically above 0.80) and acceptable Type I error rates (generally near the nominal level) when used in
conjunction with the 2PLM (e.g., Finch & Habing, 2007; Gessaroli & DeChamplain, 1996) but comparisons to other more commonly used LD measures, such as $Q_3$, have yet to be reported. As NOHARM is limited to applications with dichotomous data, so too is $\chi^2_{\nu/D}$, which is considered a shortcoming of the statistic. Although it has not been mentioned in the literature previously, the intermediate steps of this measure seem quite similar to Yen’s (1984, 1993) $Q_3$ index, particularly the use of residual correlations among item pairs and the Fisher’s r-to-z conversion that may also be applied to $Q_3$. While $\chi^2_{\nu/D}$ was developed to provide a test of overall fit, DeChamplain and Tang (1997) noted that the intermediary z-values from Equation 14, after standardization, provide an index that might be useful in flagging particular item pairs within the data set, much like Yen’s transformed $Q_3$.

2.2.2: Confirmatory factor analysis modification indices. The use of modification indices (MIs) after fitting a confirmatory factor analysis (CFA) model has been suggested as a method for detecting locally dependent items several times (e.g., Steinberg & Thissen, 1996; Swygert, McLeod, & Thissen, 2001). MIs assess the improvement in overall model fit, as measured by $\chi^2$, which would be achieved if an initially restricted parameter was freely estimated. This can be applied to LD detection by obtaining modification index values for residual correlations among all individual item pairs, with significant values indicating the possible presence of LD. To make the process more concrete, Figure 1 depicts a single factor CFA path model with 4 manifest predictors. The initial model is estimated with only paths that are drawn with solid lines and a corresponding $\chi^2$ value of model fit is obtained; a modification index for item pair
X1 and X2 may then be calculated to find the expected decrease in the model $\chi^2$ value if a residual correlation between the two variables of interest (as depicted by the dashed curve) was added to the model. An alternate, and equivalent, way of depicting MIs is presented in Figure 2. Stemming from the IRT paradigm, which mechanically is unable to incorporate residual correlations, Figure 2 replaces the correlation between the error terms of X1 and X2 with an additional latent variable, labeled as “specific LV” in Figure 2, upon which only the two locally dependent variables load with constrained-to-be-equal factor loadings. Additionally, the specific latent factor is uncorrelated with the general factor. After imposing the constraints of equal factor loadings and uncorrelated latent variables, a test on generated data found the degrees of freedom and all fit statistic values for the two conceptualizations of LD (as represented by Figure 1 and Figure 2) to be equal, implying that the models are, in fact, equivalent.

MIs are commonly found, one at a time, for all possible pairs of items. The most common modification index used in CFA is based on a univariate Lagrangian Multiplier Test. These MIs are calculated as

$$MI = \left[ \frac{\partial}{\partial \theta_i} \log L(\theta) \right]^2 \left[ I^{-1}(\hat{\theta}_r) \right]_{ii},$$

(16)

where $L(\theta)$ is the likelihood of the unrestricted model (the model with the additional residual correlation parameter), and $\left[ I^{-1}(\hat{\theta}_r) \right]_{ii}$ is the $i^{th}$ diagonal element of the Fisher information matrix associated with the more restrictive (initial) model, denoted as $\theta_i$ and with respect to which the partial derivative of the log-likelihood of the unrestricted model is taken (Bollen, 1989). Such pairwise MIs are distributed approximately $\chi^2$ with one
degree of freedom; due to the known distributional form, the modification index values can provide a formal test of significance.

Kim (2007) found modification indices to be very successful in detecting locally dependent item pairs over a variety of simulation conditions. In contrast to these promising results, Hill et al. (2007) encountered issues when applying MIs to collected data, such as impossible modification indices values, which call into question the suitability of MIs as an LD detection method.
Chapter 3: Method

As noted earlier, the purpose of this study is to evaluate LD indices using a simulation study, in which data are generated to mirror the properties commonly encountered in psychological assessment data. This focus on psychological assessment informed the scale length and sample size choices, as well as the generating and fitted IRT models. Table 3 details the manipulated variables within the study and the various levels which were used. In total, there were 342 cells in the study design. One thousand replications were run for each cell.

3.2: Latent Construct and Item Parameters

Individual theta values were generated from the standard Normal distribution. For conditions including ULD, theta values were generated from a dimensionally appropriate multivariate Normal distribution with means of zero, unit variances, and covariances of zero between dimensions. Item discrimination parameters were sampled from a Normal distribution with a mean of 1.7 and standard deviation of 0.3; this distribution was selected based on the $a$-parameter distribution found by Hill’s (2004) examination of 15 psychological tests which were fitted with the GRM model. The distribution of severity parameters ($b$s) varied with the number of response categories. For items generated with two response categories, the severity parameter was generated from a Normal (0, 1.0) distribution, similar to the $b$ distributions used in Chen and Thissen (1997), Finch and Habing (2007), and Kim (2007). For items with five categories, the procedure set forth in
Hill (2004) was followed to construct the needed severity parameter values. The first stage of this procedure sampled an initial severity parameter, $b_1$, from a Normal (-1.5, 0.5) distribution. Next, a Normal (1.0, .2) distribution was sampled until a positive value was obtained. The sampled value, termed $d_1$, is the distance between $b_1$ and $b_2$; the sum of $b_1$ and $d_1$ gives the value of the second severity parameter, $b_2$. The process of sampling a $d$ value and adding that $d$ to the previous $b$ value is repeated a total of 3 times to obtain values for the 4 severity parameters necessary for a 5-category GRM. For example, if both the $b_1$ and $d$ distribution were at their expected values for all draws, the resulting severity parameters would be -1.5, -0.5, 0.5, and 1.5.

3.3: LD type and Degree of LD

3.3.1: Surface local dependence. Surface local dependence was detailed earlier as a probabilistic process in which $\pi_{LD}$, the probability that the second item in the LD pair will have a response identical to the first item, denotes the extent of LD present within an item pair. In manipulating degree of LD in item pairs with SLD, $\pi_{LD}$ took on values of 0.3, 0.5, and 0.8, identical to the probabilities originally used in Chen and Thissen (1997).

3.3.2: Underlying local dependence. For the purpose of simulating item pairs with ULD, a compensatory multidimensional IRT model was employed. Compensatory models are those in which the latent traits being estimated are assumed to have an additive model (Bock, Gibbons, & Muraki, 1988). The particular $m$-dimensional compensatory model employed was the multidimensional 2 parameter logistic model (e.g, Reckase, 1985), which may be expressed as
\[ P(x_j = 1|\theta) = \frac{1}{1 + \exp(-a_j(\theta - b_j))}, \quad (17) \]

in which \( a \) and \( \theta \) are multidimensional analogues of \( a \) and \( \theta \) described in Equation 4 and \( b_j \) is now a multidimensional difficulty parameter. A multidimensional GRM may be derived from this formulation, using the same techniques applied to extend the unidimensional 2PLM.

With such a model formulation, it is sufficient to have a high level on only one dimension to receive a high estimated probability of endorsement. These models are typically described as those in which “a deficiency in one ability can be offset by an increase in other abilities” (Bolt & Lall, 2003, p. 395). For instance, using a generic depression scale, the manner in which ability levels are estimated by a compensatory model would allow for a respondent to be well above the mean level in somatic symptoms while at the same time have an latent trait level for affective symptoms that is relatively low and still be predicted, with high probability, to endorse an item such as, “I have experienced changes in my sleep pattern.”

The multidimensional model was specified so, in addition to one general factor onto which all items loaded, each locally dependent item pair contributed an additional latent construct to the model upon which only the items exhibiting ULD loaded, similar to what is depicted in the path diagram of Figure 2. Due to identification issues with latent factors that have only two observed variables, these “specific” factor slopes for item pairs were constrained to be equal and the “specific” latent factors are generated to
be uncorrelated with the general factor. For item pairs with ULD, the degree of local
dependence was defined by the magnitude of the slope of the specific factor in relation to
the slope of the general factor; in the simulation, the ratios used were 0.5, 1.0, and 1.5
times the slope of the general factor. Additionally, rather than a rote process of
multiplying the general latent construct slope to obtain the specific latent construct
slopes, appropriate distributions were sampled. For example, in the 0.5 degree of ULD
condition the general slopes were sampled from a Normal (1.7, 0.3) distribution and
specific slopes were sampled from a Normal (0.85, 0.15) distribution, a distribution with
a mean value half that of the general slope distribution mean and a standard deviation that
allows for a small degree of variation among the sampled values. For completeness, the
distributions from which the specific factor slopes were sampled, corresponding to the
0.5, 1.0, and 1.5 general-specific slope ratios were Normal (0.85, 0.15), Normal (1.7,
0.15), and Normal (2.55, 0.15), respectively.

3.4: Data Generation and Validation

Using the item and latent construct parameters estimates detailed previously,
programs were written in R (version 2.10) (R Development Core Team, 2009) to simulate
data in the Null conditions, the SLD conditions, and the ULD conditions. A dataset from
a Null condition cell in both the 2PLM and GRM conditions (k = 20, N = 1000) was
submitted to MULTILOG (Thissen, 1991) to verify that the data generation was
proceeding properly; item- and latent construct-level parameter estimates from
MULTILOG corresponded to the generating values up to the second decimal place. A
dataset from a ULD condition cell was submitted to IRTPRO (Cai, du Toit, & Thissen, in
press) to verify that the multidimensional data generation was proceeding properly; item parameter estimates corresponded to the generating values to the second decimal place.

3.5: Parameter Estimation

IRT parameter estimation was conducted in MULTILOG and NOHARM (Fraser & McDonald, 1988). MULTILOG employs full information maximum likelihood estimation, implemented with the Bock-Aitkin EM algorithm. NOHARM, as noted previously, is based on the non-linear factor analysis of McDonald (1967) and employs a limited information estimator, the implementation of which is detailed in McDonald (1982).

MULTILOG parameter estimates were used in calculating $Q_3$, Fisher’s r-to-z transformed $Q_3$, and $G^2$. To obtain more precise item parameter estimates from MULTILOG, the default number of quadrature points was increased to 91. No prior distributions were specified for the item parameters estimates and the maximum number of EM cycles was increased from 25 (the default) to 200. Additionally, MULTILOG was also used to obtain expected a posteriori (EAP) latent construct level estimates necessary for the calculation of $Q_3$. NOHARM parameter estimates, and the resulting residual covariance matrix, were used in the calculation of $\chi^2_{G/D}$. All NOHARM estimation procedure default values were left unchanged.

3.6: Calculation and Validation of LD Indices

A program was written in R for the calculation of $G^2$, $Q_3$, Fisher’s r-to-z transformed $Q_3$, and $\chi^2_{G/D}$ (see Appendix A for R code). The program LDIP (Kim, Cohen, & Lin, 2006) was used to verify the accuracy of several of the LD indices calculated in R.
Values for the JSI were obtained from IRTPRO. MIs were calculated in Mplus 6.1, which employed a limited information estimator, specifically a type of diagonally weight least squares (WLSMV) applied to the polychoric correlation matrix, to obtain the needed values (Muthén & Muthén, 1998-2010).
Chapter 4: Two Parameter Logistic Model Results

Preliminary attempts at data simulation and analysis encountered convergence problems within MULTILOG, especially in cells with high degrees of LD, a low number of items, and small sample size. For example, over 40% of the initial replications in the 10 item, N=250, $\pi_{LD} = 0.8$ SLD cells failed to reach convergence by the 200th iteration of the EM algorithm. To prevent such non-converged MULTILOG estimates from affecting the final results, an automated convergence check was incorporated into the data simulation/estimation step, in which the parameters values, and therefore data, for any replication which failed to converge were automatically replaced. Using this convergence check, all replications were ensured to have converged in MULTILOG by the specified 200th iteration.

There were an additional 169 replications, which had converged in MULTILOG, that were discarded due to computational issues encountered during parameter estimation with other programs. Of those replications, 164 of them occurred in a cell that had 10 items and a $\pi_{LD}$ value of 0.8; the number of discarded replications within this group of simulation design cells decreased with increased sample size. Major problems encountered in Mplus for these problematic replications were: 1) items in locally dependent pairs that were almost perfectly predictive of each other (observed polychoric correlations > 0.98) and 2) item pairs with empty cells in the contingency tables. These highly related item pairs also presented problems for IRTPRO, in which the slope
estimate for one or both of the items tended to become unreasonably large (e.g., > 50). These issues resulted in warnings/errors from Mplus and IRTPRO being unable to converge on SE estimates for the noted slope parameters, although the estimation procedure terminated before the maximum number of permitted EM cycles was reached.

Although the discard rates in the reported cells are troublingly high, the process of replacing replications that failed to converge is not viewed as a major problem with respect to assessing the effectiveness of the LD indices. Because LD indices may only be calculated when the initial parameter estimation has succeeded, the power and Type I error values should be unaffected by the replacement of unconverged replications. However, it is recognized that the parameter recovery assessments will be downwardly biased by the exclusion of replications that failed to converge. For the interested reader, the overall non-convergence rates for all 2PLM cells, as well as primary analyses excluding any cell with an initial replication discard rate higher than 5%, are presented in Appendix D.

4.1: Parameter recovery

Prior to reporting the performance of the selected LD indices the ability of MULTILOG to recover the generating parameter values will be detailed. This is done in an effort to gauge the importance of the performance of the LD indices; if parameter recovery is in a reasonable range even with large amounts of LD or other independent variable manipulations, the actual performance of the LD measures may be of less
importance in applied settings. Parameter recovery is assessed using bias and root mean square error (RMSE), which are, respectively, calculated as

\[
\text{bias}_q = \frac{\text{est. parameter value}_q - \text{gen. parameter value}_q}{\text{est. parameter value}_q} \quad (18)
\]

\[
\text{RMSE} = \sqrt{\sum \text{bias}^2} \quad (19)
\]

4.1.1: No local dependence. Average RMSE values for cells exhibiting no local dependence, broken down by number of items and sample size are presented in Table 4 to provide a frame of reference for values from cells with LD. Parameter recovery for the discrimination and severity parameter improved with an increase in the number of items and, more substantially, with increased sample size. The largest improvement in discrimination parameter recovery resulted from the increase from \(N = 250\) to \(N = 500\) and, with respect to number of items, a steady improvement was seen from 10 to 20 to 40 items; changes from \(N = 500\) to \(N = 1000\), while resulting in reduced RMSE values, were less pronounced. The severity parameter showed its greatest improvement in recovery as sample size increased but these changes were less dramatic than those seen in the discrimination parameter RMSE values. The observed patterns in both the discrimination and severity parameters generally replicate previous findings regarding the performance of MULTILOG in relation to the 2PLM (Stone, 1992).

4.1.2: Surface local dependence. In discussing the parameter recovery of MULTILOG when SLD was present, results will focus on the RMSE values for those items exhibiting LD, as well as for items conforming to the LI assumption. Table 5 presents the average RMSE values for the set of SLD simulation cells displaying the poorest parameter recovery for locally dependent items, in which the number of scale
items is 10, the sample size is 250, and $\pi_{LD}$ took its most extreme value of 0.8. In relation to the comparable Null cell (Table 4; items = 10, $N = 250$), the discrimination parameter RMSE values for items without LD in Table 5 (“Null items only”) are comparable, if not improved, to that seen in the Null cell while the severity parameter RMSE values are 1.5 to 1.7 times that of the RMSE observed in the Null cell.

Focusing on the “LD Items only” lines of Table 5, the slope parameter RMSEs for items in an SLD pair were quite high (e.g., 7.804). As reported previously, the distribution from which the generating discrimination parameters were drawn is distributed $N(1.7, 0.3)$, which serves to highlight how discrepant the RMSE values actually are; for example, the RMSE value of 8.487 corresponding to the noted simulation condition with one locally dependent pair implies the average estimated discrimination parameter would be over 28 standard deviations away from the generating mean value. To approach these findings in another way, Figure 3 presents the distributions of recovered discrimination parameters for items with no LD and items with LD; a brief visual inspection shows the recovered slope parameters of items constructed to be locally independent approximate the generating distribution (for two LD pairs, the mean slope value is 1.47 and the standard deviation is 0.43); the estimated slopes for LD items are poorly recovered and extremely varied, ranging from 0.17 to 89.85 with a mean of 5.93 and standard deviation of 6.62.

ANOVA$s predicting the LD-items-only RMSE for both parameters ($a$, $b_1$) using the simulation design variables and their interactions (up to 3-way interactions) were conducted. The overall analyses of the discrimination and severity parameter RMSE
values, both of which accounted for 99.9% of all variance in the RMSE values, were statistically significant. The most noteworthy predictors for the discrimination RMSE values were degree of LD, $F(2,16) = 4656.17, p < .001$, number of items, $F(2,16) = 4313.49, p < .001$, and the interaction of degree of LD with number of items, $F(4, 16) = 3117.74, p < .001$. Figure 4 provides a graphical display of the statistically significant interaction. Probing the interaction with Tukey’s HSD post-hoc test, the most prominent finding was that the cells with 10 and 20 items and $\pi_{LD} = 0.8$ (RMSE = 7.08 and 1.27, respectively) were higher than all other item and degree combinations (all adjusted $ps < .001$); for completeness, all statistically significant differences are indicated in Table 6.

With respect to the severity parameter, the only prominent predictor was degree of SLD. Post-hoc tests found the RMSE values for the three levels of degree, 0.3 ($M = 0.27, SD = 0.03$), 0.5 ($M = 0.43, SD = 0.02$), and 0.8 ($M = 0.68, SD = 0.04$), to significantly differ from one another (all adjusted $ps < 0.001$); these values indicate that as the degree of SLD increases, the accuracy of the recovered severity parameters suffers.

4.1.3: Underlying local dependence. The presentation of results for the parameter recovery of MULTILOG when ULD was present will, as with SLD, focus on the RMSE values for only those items exhibiting LD as well as the overall performance. Table 7 presents the average RMSE values of the set of ULD simulation cells displaying the poorest parameter recovery for LD items, in which there are 40 items, the sample size is 250, and the slope ratio took its most extreme value of 1.5. Comparing the RMSE values of locally independent items, found on the “Null items only” lines in Table 7, to that of the Null cell recovery (Table 4; items = 40, $N = 250$), the discrimination recovery
is similar and the average severity parameter RMSE value is somewhat elevated when one or two pairs of locally dependent items are present and is comparable to the Null value when there are four pairs of items exhibiting LD.

Figure 5 presents the distributions of the estimated $\alpha$-parameters for items with LD compared to items not involved in a locally dependent pair. As can be seen, the distribution of the non-LD items from the 1000 replications approximately mirrors the $N(1.7, 0.3)$ generating distribution; for example, the distribution from the cell with four locally dependent item pairs has an observed mean of 1.80 and a standard deviation of 0.51. Although the distribution of the recovered discrimination parameters of the locally dependent items is less severely comprised than the similar SLD distribution from Figure 3, the $\alpha$-parameter estimates for these items are clearly biased; specifically, they are underestimated compared to the generating values. For the noted 40 item, $N = 250$, $\alpha$-ratio = 1.5 cell with $p = 4$ locally dependent items pairs, the LD item distribution has a mean $\alpha$-parameter value of only 0.99 and a standard deviation of 0.27.

ANOVA s using the RMSE of the two item parameters as the dependent variable and the simulation design variables and their interactions (up to 3-way interactions) as predictors were conducted. The overall F-test for the severity parameter was non-significant; only results from the discrimination parameter ANOVA are reported. The most noteworthy predictor of the slope parameters were degree of LD, F(2,16) = 11761.60, $p < .001$. Post-hocs of the main effect of degree, controlling for Type I error using Tukey’s HSD, found the RMSE values of all levels of degree of ULD, specifically
0.5 (M = 0.29, SD = 0.07), 1.0 (0.48, SD = 0.04) and 1.5 (M = 0.71, SD = 0.06), to significantly differ from each other, all adjusted ps < 0.01.

4.1.4: Overall comparisons. To facilitate broader comparisons, ANOVAs were also conducted on a combined SLD/ULD dataset. As anticipated from both the SLD and ULD condition ANOVAs, degree of LD, which was re-coded to represent relative magnitude (e.g., both πLD = 0.3 and a-ratio = 0.5 were coded as “low”), was the predictor of discrimination parameter RMSEs with the highest F-value, F(2,64) = 573.75, p < .001. Type of LD, F(2,64) = 495.11, p < .001 and the interaction of LD type and number of items, F(2,64) = 429.42, p < .001, both had F values substantially larger than those of other predictors, as did the three-way interaction between type of LD, number of items, and degree of LD. With respect to the statistically significant three-way interaction, presented in Figure 6, after controlling for multiple comparisons, the most striking finding was that the combination of a 10 item scale, with SLD induced at its highest level, was found to have inflated RMSE values compared to all other possible combinations of scale length, degree of LD, and LD type (all adjusted ps < .001). Additional significant differences are indicated in Table 8.

4.2: Performance of LD indices

In assessing the performance of the LD indices, two concepts are of primary importance: power and Type I error rate. Power is defined as the correct identification of locally dependent pairs by an LD index and is reported as the average proportion of correctly flagged pairs out of the possible 1000 replications. Type I error rate is defined as the incorrect classification of a non-locally dependent pair as being locally dependent;
similar to power, it will be reported as the average proportion of incorrectly flagged pairs out of the possible 1000 replications.

**4.2.1: No local dependence.** As there are no locally dependent pairs to be identified in the Null cells, only Type I error rates are discussed. Table 9 presents the average Type I error rates for the seven LD indices available for use with the 2PLM in the Null condition cells, both overall and decomposed by number of items and sample size. The overall Type I error rates for the indices tended to be low, with $\chi^2_{G/D}$ displaying the lowest overall error rate of 0 and the r-to-z transformed $Q_3$ returning the highest overall error rate, as well as the highest rate in an individual cell (Type I error = 0.817; Items = 40, N = 1000). While error rates for most of the indices tend to increase with larger sample size, across all indices no broad patterns of note appear within the Null condition error rates.

The indices $G^2$, MIs, $\chi^2_{G/D}$, and the r-to-z transformed $Q_3$ are assumed to follow known distributions, which provide the means of formal statistical testing; these indices are, therefore, expected to display Type I error rates at the nominal level, which was set at 0.05 for all indices. The departure of an index from its assumed distributional form is theoretically problematic and requires further examination to determine if an alternate distribution is available, with which the associated critical values could bring Type I error to nominal rates, or if distributional tests should be abandoned and a simple cut point should be chosen that both minimizes Type I error and maximizes power in LD detection. Table 10 presents the observed means and standard deviations for the noted indices. Both $G^2$ and MIs are expected to approximately follow a 1 (or less) degree of freedom $\chi^2$
distribution, which has an expected value of 1 (E(x) = df) and standard deviation of 1.41 (= \sqrt{(2*df)}). The degrees of freedom for the \( \chi^2_{G/D} \) change with number of items and for the 10, 20, and 40 item lengths are 35, 170, and 740, respectively; these values results in expected values of 35, 170, and 740 and expected standard deviations of 8.37, 18.44, and 38.47. The r-to-z transformed Q\textsubscript{3} is expected to approximate a standard Normal distribution. In general, the G\textsuperscript{2} values and MIs tend to mirror the expected distributions, although the MI s standard deviations are somewhat inflated in all N = 1000 conditions. The transformed Q\textsubscript{3} displays noticeable departures from the anticipated values, with extremely inflated mean and standard deviation values in the all but the 10 item cells. The most problematic index, though, is \( \chi^2_{G/D} \) which has observed means approximately half that of the expected value and standard deviations that are also well below the values anticipated by the theoretical distribution.

4.2.2: Surface local dependence. The first and second columns of Table 11 present the means and standard deviations for overall power and Type I error rates for the selected indices across all SLD cells. Overall, the power of the indices for detecting SLD ranged from adequate to extremely high, with the exception of \( \chi^2_{G/D} \). Of the other six indices, Q\textsubscript{3} displayed the lowest average power of 0.77 and both the intermediary residual z values from \( \chi^2_{G/D} \) calculations and MIs exhibiting power very close to 1. However, high power by itself is not informative of the performance of the indices, as an index may be erroneously flagging almost every pair, which would result in high power (and also high Type I error). Taking both power and Type I error together, it is obvious that the MIs are experiencing difficulties and have a statistical testing criterion that is overly liberal, as
seen in the conjunction of both high power and a high error rate. In terms of overall
performance, among the three indices using formal statistical tests, $G^2$ displays the most
desirable performance, with power above 0.90 and a Type I error rate that is closest to the
nominal level of 0.05. For the JSI, $Q_3$, and residual $z$ values, which use cut-values as
opposed to tests based on assumed distributional forms, the residual $z$ values have the
highest power and an error rate comparable to that of JSI. From this overall table, it is
also of note that the power of two of the three the cut-point indices (i.e., JSI and $Q_3$) is
much more variable (as evidenced in the larger standard deviations) than those indices
using distribution critical values.

To further examine the differences among the indices, ANOVAs were conducted
using both Type I error and power as dependent variables; simulation design variables
and their interactions (up to 3-way interactions) were included as predictors. Table 12
provides the ordering of the most prominent F values from these analyses. Of the four
design variables and all possible three way interactions among the variables, number of
items and degree of SLD are the most prominent predictors for almost all indices, with
respect to both power and Type I error rates. The top portion of Figure 7 depicts the main
effect of scale length for power. For all indices but $\chi^2_{G/D}$, as number of items increases,
power also increases; power rates range from an unacceptable 0.46 to 0.99 in the 10 item
cells but within the 20 and 40 items cells, all indices (excepting $\chi^2_{G/D}$) display power rates
above 0.80. With respect to number of items and Type I error, the lower portion of Figure
7 shows that as number of items increases, Type I error rates tend to decrease; the most
noticeable exception to this trend is the inflated Type I error rate for the r-to-$z$
transformed $Q_3$ when applied to 40 item scales. With respect to the main effect of degree of SLD, as seen in Figure 8, both $Q_3$ and $\chi^2_{LD}$ show substantial increases in power as degree of LD increases from 0.3 to 0.5. This trend continues for $\chi^2_{LD}$ as degree of SLD becomes even more pronounced (i.e., $\pi_{LD} = 0.8$) while $Q_3$ power remains relatively stable. While degree of SLD was the most important predictor for power of the MIs, it appears that the changes being predicted were not practically meaningful, as the probability of correctly detecting LD pairs remained above 0.95 for all $\pi_{LD}$ values.

The interaction of number of items and degree of SLD also appeared as a prominent predictor for several indices in power portion of Table 12, making the interpretation of the main effects dubious for $G^2$, JSI, $Q_3$, and $r$-to-$z$ transformed $Q_3$. The interaction is presented in Figure 9; inspection reveals that for all indices in which the interaction was a prominent predictor there was a noticeable decrease in power when moving from $\pi_{LD} = 0.5$ to $\pi_{LD} = 0.8$, but only in conjunction with a 10 item scale.

4.2.3: Underlying local dependence. The right portion of Table 11 presents the means and standard deviations for overall power and Type I error rates for the seven selected indices across all ULD cells. In general, the power of the indices for detecting ULD was lower than would be generally considered adequate with values ranging from a high of 0.87 ($r$-to-$z$ $Q_3$) to a low of 0.26 ($\chi^2_{LD}$). Considering both power and Type I error, there is no index that is performing well; the index with the highest power has a mean Type I error well above the nominal level and the index with the lowest Type I error also exhibits has low power. The indices which are closest to performing at acceptable levels are the likelihood ratio and the intermediate residual $z$ values from $\chi^2_{LD}$ calculations.
These general trends mirror previous findings that ULD is much harder to accurately identify than SLD (e.g., Chen & Thissen, 1997).

Power and Type I error rates for the ULD cells were predicted in ANOVAs by the simulation design variables; prominent factors are reported in Table 13. Degree of ULD was the most noteworthy predictor for all indices and, as seen in Figure 10, the general trend is for a large increase in power to occur as the slope ratio goes from 0.5 to 1.0 and a less pronounced increase from 1.0 to 1.5. Exceptions to this are the $Q_3$ and residual $z$ values, both of which display a more consistent rate of increase across the three levels of degree of ULD. With respect to Type I error rates, the degree of ULD, number of items, and the interaction of those two design variables account for the whole of the Table 13 entries. Figure 11 displays the interaction of degree of ULD and number of items for each index. Although the trends are for Type I error to decrease with an increased number of items and increasing degree of ULD, for the majority of the indices, the error rates remain at or below a value of 0.06 so most changes across items or degree of ULD are not of any practical significance. The one striking pattern seen in this display is the change seen across increasing degree of LD for the 10 and 20 items cells in the error rates of both the likelihood ratio statistic and MIs. Although the trend is less dramatic in the $G^2$ values compared to the MIs, the similarity seems noteworthy given that both indices are tested against an assumed $\chi^2$ distribution.

4.2.4: Overall comparisons. Within Table 11, it is evident that power tends to be higher for detecting SLD when compared to ULD. Additionally, the performance of LD indices tends to be more variable in ULD cells, as evidenced by the consistently higher
standard deviations. To more rigorously examine such trends, a comprehensive dataset including both SLD and ULD power and type I error rate values was submitted to analysis through ANOVAs. For the power of the indices, the type of LD present was a statistically significant predictor, (all \( p < .001 \)), statistically confirming the lay observation that higher power was observed for detecting SLD. Additionally, the degree of local dependences (re-coded as described in the RMSE overall comparisons section) as well as the interaction of LD type and degree were significant factors in accounting for power. Figure 12 presents the mean power values for the indices, with respect to the type of LD by degree of LD interaction. As shown, the general trend is for power to increase as degree of ULD increases, while in the SLD cells several of the indices (e.g., JSI and \( r \)-to-\( z \) \( Q_1 \)) actually display reduced power at the most extreme level of dependence. The overall analyses for Type I error failed to yield any strikingly prominent pattern across all LD indices, although number of items, as would be expected from the separate SLD and ULD results, and LD type did consistently appear as significant predictors (all \( p < .001 \)) with the expected patterns of decreased Type I error typically associated with larger numbers of items and SLD.
Chapter 5: Graded Response Model Results

All replications for which any program was unable to calculate needed values (e.g., parameter estimates and SEs, modification indices, etc) were excluded and new replications were generated. Across all GRM design cells, ten replications were excluded in total; the majority of excluded replications were in 40 items cells with sample sizes of 250, occurring both in ULD and SLD cells, although replications were also excluded from 10 and 20 item cells. Excluded replications typically resulted from a single problematic item, which through a combination of item parameter and person parameter values, had very few to no observations in at least one response category. One replication was discarded due to Mplus reporting a perfect correlation between the items in a locally dependent pair, which resulted in the program being unable to calculate MIs for the majority of the possible residual correlations.

5.1: Parameter recovery

As was done in the 2PLM results, prior to reporting the performance of the selected LD indices the ability of MULTILOG to recover the generating parameter values in both the null and LD conditions will be addressed. Parameter recovery was assessed using bias and average RMSE, as previously defined in Equations 18 and 19.

5.1.1: No local dependence. To provide a frame of reference with which to later compare the parameter recovery in cells that exhibit LD, average RMSE values for cells exhibiting no local dependence, broken down by number of items and sample size are
presented in Table 14. Individual ANOVAs predicting RMSE for each parameter consistently found sample size to be a statistically significantly predictor (all \( p < .02 \)) while number of items only significantly predicted changes in the discrimination, \( F(2,4) = 31.95, p < .001 \), and second severity parameter (i.e., \( b_2 \); \( F(2,4) = 7.92, p = .04 \)) RMSE values. Parameter recovery for the discrimination and four severity parameters improved with an increase in the number of items and, more substantially, with increased sample size. Additionally, the severity parameters display the anticipated patterns, in that RMSE is smallest for \( b_2 \) and \( b_3 \), where a larger portion of observed responses is expected to occur, while the recovery for the severity parameters at either extreme (i.e., \( b_1 \) and \( b_4 \)) was less accurate. The observed patterns in both the discrimination and severity parameters replicate previous findings regarding the performance of MULTILOG in relation to the GRM (e.g., Reise & Yu, 1990).

5.1.2: Surface local dependence. In discussing the parameter recovery of MULTILOG when SLD was present, results will focus on the RMSE values for only those items exhibiting LD as well as the overall performance, referring to both those items involved in locally dependent pairs and items conforming to the LI assumption. Table 15 presents the average RMSE values for the set of SLD simulation cells displaying the poorest parameter recovery for locally dependent items, in which number of items is 10, the sample size is 250, and \( \pi_{LD} \) took its most extreme value of 0.8; of note is this is the same cell that was most problematic in the 2PLM results. The patterns noted in the null conditions for the severity parameters are again evident, in which the middle severity parameter values are more accurately recovered than the extreme severity
parameters. While the RMSE values for the severity parameters are 1.5 to 2 times greater than those seen in the conditions with no LD, the most noticeable discrepancy between generating and estimated parameters resides in the discrimination parameter.

As seen in the “LD Items only” lines of Table 15, the slope parameter RMSEs for items in an SLD pair were quite high (e.g., 3.7). As reported previously, the distribution from which discrimination parameters were drawn is distributed N(1.7, 0.3), which serves to highlight how discrepant the RMSE values actually are; the RMSE value of 4.7 tabled for the 1 locally dependent pair cell implies the average estimated discrimination parameter would be over 15 standard deviations away from the generating mean value.

To approach these findings in another way, Figure 13 presents the distributions of recovered discrimination parameters for items with no LD and items with LD; a brief visual inspection shows the recovered slope parameters of items constructed to be locally independent approximate the generating distribution (for 1 LD pair, the mean slope value is 1.59 and the standard deviation is 0.38) while the estimated slopes for LD items are extremely varied, ranging from 0.77 to 70.29 with a mean of 3.72 and standard deviation of 4.38. Parameter recovery did improve as more LD pairs were introduced into the simulated data, that is, as the number of pairs of locally dependent items increased from 1 to 2 to 4; however, even in the 4 LD pair cell, the average LD item RMSE value for locally dependent items of 1.74 is still almost 5 times that of the generating distribution’s standard deviation.

While this section has focused on the poor parameter recovery for LD items, it also should be noted that for items generated to meet the LI assumption, the average
RMSE values are also elevated over the corresponding values in the Null cells (e.g., Table 14, 10 items, N = 250). These differences, while less pronounced than those for the LD items, are noteworthy because they imply that the presence of SLD affects the accuracy of parameter values for all items in a scale, not only those items which are involved in the LD.

ANOVA s using the LD-items-only RMSE for all 5 item parameters \((a, b_1, b_2, b_3, b_4)\) as the dependent variables and the simulation design variables and their interactions (up to 3-way interactions) as predictors were conducted. As general conclusions are comparable across item parameters, only the results of the \(a\)-parameter ANOVA, which accounted for 99.4\% of all variance in the RMSE values, will be reported. The most noteworthy predictors were degree of LD, \(F(2,16) = 296.29, p < .001\), number of items, \(F(2,16) = 278.34, p < .001\), and the interaction of degree of LD with number of items, \(F(4, 16) = 203.41, p < .001\). Probing the interaction with Tukey’s HSD post-hoc test, the cell with 10 items and \(\pi_{LD} = 0.8\) (RMSE = 2.31) was found to be statistically higher than all other item and degree combinations (all adjusted \(p\)s < .001); no other cells were significantly different from each other. Figure 14 provides a graphical display of this statistically significant interaction.

5.1.3: Underlying local dependence. The parameter recovery of MULTILOG when ULD was present will, as with SLD, focus on the RMSE values for only those items exhibiting LD as well as the overall performance. Table 16 presents the average RMSE values of the set of ULD simulation cells displaying the poorest parameter
recovery for LD items, in which there are 40 items, the sample size is 250, and the slope ratio took its most extreme value of 1.5.

Figure 15 presents the estimated distributions of the estimated $a$-parameters for items with LD compared to items not involved in a locally dependent pair. Although the distribution of the locally dependent items is less severely comprised than the similar SLD distribution from Figure 13, the discrimination values for these items are clearly biased; specifically, they are underestimated compared to the generating values. For the noted 40 item, $N = 250$, $a$-ratio = 1.5 cell with 2 locally dependent items pairs, while the distribution of the non-LD items from the 1000 replications closely mirrors the $N(1.7, 0.3)$ generating distribution, with a mean of 1.73 and a standard deviation of 0.33, the LD item distribution has a mean $a$-parameter value of only .961 and a standard deviation of .183. Examining the RMSE values in Table 16, it is again seen that the non-locally dependent items are adequately recovered; the tabled values are similar to those reported in the Null condition table. For the locally dependent items only, the RMSE values are less extreme than the comparable entries found in the SLD condition but are still approximately eight times larger than the comparable values from the Null cells, which is problematic. Unlike the SLD conditions, additional pairs of items with LD did not seem to have an effect on parameter recovery under ULD.

ANOVA$s using the RMSE of all 5 item parameters ($a, b_1, b_2, b_3, b_4$) as the dependent variable and the simulation design variables and their interactions (up to 3-way interactions) as predictors were conducted. Unlike SLD results, the general findings varied across the RMSEs for the discrimination parameters and the severity parameters;
results will be reported for the $\alpha$-parameter RMSEs and only one of the severity parameters, as results were common across these four values. The most noteworthy predictors of the slope parameters were degree of LD, $F(2,16) = 71386.90, p < .001$, and number of items, $F(2,16) = 2344.50, p < .001$. Post-hocs of these main effects, controlling for Type I error using Tukey’s HSD, found the RMSE values of 10 items (M = 0.422), 20 items (M = 0.486), and 40 items (M = 0.512) to all be significantly different from each other (all $ps < .001$). Additionally, the RMSE values associated with the levels of $\alpha$-ratio, 0.5 (M = 0.218), 1.0 (M = .476), and 1.5 (M = 0.725), were also all significantly different from each other (all $ps < .001$). For the second severity parameter (i.e., $b_2$), an ANOVA found significant predictors to be the main effects of number of items, $F(2,16) = 8.65, p = .003$, and degree of LD $F(2,16) = 7.05, p = .006$. After controlling for multiple comparisons, there were no significant differences among levels of number of items and, with regard to the magnitude of the slope ratio, it was found that the slope ratio of 1.5 (M = 0.174) was associated with significantly higher RMSE values than both 1.0 (M = 0.122) and 0.5 (M = 0.105), $p < .01$ and .05, respectively.

5.1.4: Overall comparisons. To facilitate broader comparisons, ANOVAs were also conducted on a combined SLD/ULD dataset, using the $\alpha$-parameter and all four severity RMSEs for LD only items as the dependent variable; again, the severity parameter analyses all displayed the same general findings and only one will be reported.

As anticipated from the SLD and ULD condition ANOVAs, degree of LD, which was re-coded to represent relative magnitude (e.g., both $\pi_{LD} = 0.8$ and $\alpha$-ratio = 0.5 were coded as “low”), was the most prominent predictor, $F(2,64) = 129.96, p < .001$, of
discrimination parameter RMSEs. Average RMSE values for the \( \alpha \)-parameters, by type and magnitude of LD are displayed in the upper half of Figure 16. In contrast to the separated analyses, the overall analysis for the second severity parameter found type of LD, that is surface versus underlying, to be the most prominent predictor of RMSE \( F(1,64) = 224.16, p < .001 \), in which SLD RMSEs (M = 0.253) were significantly higher than those of ULD (M = 0.134). The only other noteworthy predictor for \( b_2 \) was degree of LD, \( F(2,64) = 88.06, p < .001 \). The lower half of Figure 16 displays the means for type and degree of LD and Figure 17 presents the average RMSE for all item parameters within each LD type and demonstrates the “type of LD” difference found in the analysis of the second severity parameter.

5.2: Performance of LD indices

The performance of the selected LD indices will again be assessed using Type I error rate and power, both as detailed previously in the 2PLM results section.

5.2.1: No local dependence. As there are no locally dependent pairs to be identified in the Null cells, only Type I error rates are discussed. Table 17 presents the average Type I error rates for the 5 LD indices appropriate for use with the GRM in the Null condition cells, both overall and decomposed by number of items and sample size cells. The overall Type I error rates for the indices tended to be low, with the unmodified \( Q_3 \) displaying the lowest overall error rate of 0 and the r-to-z transformed \( Q_3 \) returning the highest overall rate, as well as the highest rate in an individual cell (error = .165; Items = 10, \( N =1000 \)). While error rates for most of the indices tend to increase as sample
size increases, across all indices, no broad patterns of note appear within the Null condition error rates.

The indices $G^2$, MIs, and $Q_3$ r-to-z are assumed to follow known distributions, which provide the means of formal statistical testing, and would be expected to display Type I error rates at the nominal level, which was set at 0.05 for all indices. The departure of the indices from the assumed distribution form is theoretically problematic and requires further examination to determine if an alternate distribution is available, with which the associated critical values could bring Type I error to nominal rates, or if distributional tests should be abandoned and a simple cut point should be chosen that both minimizes Type I error and maximizes power in LD detection. Table 18 presents the observed means and standard deviations for the noted indices. The distribution of $G^2$ should now approximate a $\chi^2$ distribution with 16 degrees of freedom, with an expected value of 16 and standard deviation of 5.66. As in the 2PLM case, MIs are expected to approximately follow a 1 degree of freedom $\chi^2$ distribution, and the r-to-z transformed $Q_3$ is expected to approximate a standard Normal distribution. The data in Table 18 show the $G^2$ to have a slightly inflated mean value in all conditions and an extremely inflated standard deviation in the $N = 250$ cells when paired with both 10 and 20 item scales. Overall, the MIs appear to be well-behaved, with the exception of the mean values and standard deviations being noticeably depressed in the 40 item cells. The transformed $Q_3$ descriptives adequately mirror the expected value and standard deviation, although all 10 items cells, but especially those with larger sample sizes, display inflated means.
5.2.2: Surface local dependence. The left portion of Table 19 presents the means and standard deviations for overall power and Type I error rates for the 5 selected indices across all SLD cells. Overall, the power of the indices for detecting SLD was extremely high, with $Q_{3}$ displaying the lowest power of 0.84 and both $G^2$ and MIs exhibiting power very close to 1. As noted previously, high power by itself is not informative of the performance of the indices; considering both power and Type I error together, it is obvious that the MIs are experiencing difficulties and have a statistical testing criterion that is overly liberal, as seen in the conjunction of both high power and an extremely inflated Type I error rate. In terms of overall performance, among the three indices using formal statistical tests, $G^2$ displays the most desirable performance, with high power and a Type I error rate that is closest to the nominal level of 0.05. For the JSI and $Q_{3}$, which both use cut-values as opposed to tests based on assumed distributional forms, the JSI has higher power and an error rate comparable to that of $Q_{3}$. From this overall table, it is also of note that the power of the cut-point indices is much more variable (as seen in the larger standard deviations) than those indices using distribution critical values.

To further examine the differences among the indices, ANOVAs were conducted using both Type I error and power as dependent variables; simulation design variables and their interactions (up to 3-way interactions) were included as predictors. Table 20 provides the ordering of the most prominent F values from these analyses. Of the four design variables included in the ANOVAs, only number of locally dependent pairs fails to appear as the most prominent predictor in Table 20. Although sample size and degree of LD are significant predictors of both MIs and $G^2$, it can be seen in Figure 18 that there
is not much change in power across the levels of either of these variables. The most noticeable patterns in Figure 18, and corresponding to factors in Table 20, is the main effect of degree of LD for the Q₃, in which Q₃ clearly experienced difficulties detecting SLD when \( \pi_{LD} = 0.3 \). Figure 19 displays the interaction of number of items and degree of LD; again, the main effect of degree of LD is apparent in Q₃ and the poor performance of JSI with high LD and few items can also be seen to drive the significant interaction reported in Table 20. Figure 20 displays the number of items by degree of LD interaction, which was found to be a prominent factor in all 5 LD indices. As can be seen, the number of items is the primary force in reducing Type I error rates, resulting in improved identification for all indices as scale length increases. The interaction of degree of LD and number of items is reported in Table 20 as the primary factor in predicting Type I error for Q₃, but the graphical display of the data (i.e., Figure 20) shows the effect to be very small, and not practically meaningful.

**5.2.3: Underlying local dependence.** The right portion of Table 19 presents the means and standard deviations for overall power and Type I error rates for the 5 selected indices across all ULD cells. In general, the power of the indices for detecting ULD was adequate, with Q₃ r-to-z transformed displaying the highest power of .95 and the untransformed Q₃ exhibiting power of 0.72. Considering both power and Type I error, MIs demonstrate the poorest overall performance, with acceptable power but an average Type I error rate well above the nominal 0.05 level. The r-to-z transformed Q₃ displayed the best performance in detecting ULD across all cell conditions with the highest (and
least variable) average power and an observed error rate, that while elevated over 0.05, is not obscenely high.

As was done in the SLD, power and Type I error rates were predicted in ANOVAs by the simulation design variables. The prominent factors are reported in Table 21. Within the ULD cells, the degree of LD was the most prominent factor in predicting power for all 5 LD indices. Figure 21 displays the average power for the indices across the three levels of slope ratios, plainly showing that the main effect is primarily driven by the poor performance when the specific factor slope is only half that of the general factor slope; with equal slopes or a specific factor slope that is larger than the general slope, the ability of all indices to detect the ULD improves to well over 0.90. With regards to the Type I error rates, the number of items appears to be the most consistent predictor across LD indices, although the number of items by number of locally dependent pairs interaction has the largest F value for $Q_3$. To most efficiently depict the necessary factors, Figure 22 displays the interaction of items by pairs from which the main effect of items may also be gleaned for the other indices. Although the interaction is the largest predictor of $Q_3$, this appears to again be a case of very small differences, so small as to be practically not meaningful, being well predicted. However, the main effect of items is clearly shown, with the greatest movement exhibited by the MIs, the Type I error of which dramatically decreases from approximately 0.30 with 10 items to almost zero with 40 items.

5.2.4: Overall comparisons. Within Table 19, it is evident that power tends to be higher for detecting SLD when compared to ULD after fitting the GRM, as was also
found in the 2PLM conditions. Additionally, the performance of LD indices tends to be more variable in ULD cells, as evidenced by the consistently higher standard deviations. To more rigorously examine such trends, a comprehensive dataset including both SLD and ULD power and type I error rate values was submitted to analysis through ANOVAs. For the power of all 5 indices, the type of LD present was a statistically significant predictor, (all $p < .001$), statistically confirming the lay observation that higher power was observed for detecting SLD. Additionally, the magnitude of local dependences (re-coded as described in the RMSE overall comparisons section) was significant factor in accounting for power in all 5 indices. Although there were additional significant differences, after controlling for multiple comparisons within each LD index, the lowest tested level of ULD (slope ratio = 0.5) was always found to exhibit significantly lower power than all other possible LD types (i.e., the other 2 slope ratio values in ULD and all tested $\pi_{LD}$ values). The overall comparisons for Type I error failed to yield any strikingly prominent pattern across all LD indices, although number of items did consistently appear as a significant predictor (all $p < .001$), as would be expected from the individual SLD and ULD results.
Chapter 6: Ancillary Results

Prior to the conclusions and recommendations regarding the overall performance and continued use of the assessed LD indices, there were several findings from the simulation results that prompted additional investigations. Specifically, the topics submitted to more detailed exploration were the differences in detecting ULD versus SLD and the poor performance of the JSI when the number of LD pairs increased.

With respect to the SLD/ULD distinction that has been previously noted (e.g., Chen & Thissen, 1997) and replicated here, the use of a limited information estimator allowed for the comparison of the magnitude of LD, across ULD and SLD, by providing a common metric, the polychoric correlation generated during the fitting of the model in Mplus. As noted previously, when Mplus encountered problems in estimation the flagging of such an issue was generally in the form of a warning regarding highly correlated items; these warnings were encountered more often in SLD cells than ULD cells, which was the prompt for this investigation. Table 22 presents the average polychoric correlation reported between the two items induced to exhibit local dependence in a selection of cells with various scale length and sample size combinations; the average polychoric correlation induced by the most extreme SLD value is consistently higher than the average correlation for the most extreme ULD value included in the simulations, regardless of scale length and sample size manipulations.
Although not a primary focus in the previous reporting of simulation results, conditions in which there was more than one pair of LD items proved rather problematic for the performance of the JSI. Appearing in Tables 12, 13 and 20 as a noteworthy predictor of power, the general trend regarding number of pairs was for power to decrease as the number of pairs increased; while appearing only in Table 13, the number of items by number of pairs interaction also played a non-trivial role. Figure 23 presents, for both the GRM and 2PLM results, the average power of the JSI, parsed by number of items and number of LD pairs present. While the entries in the previous results tables would suggest that the main effect of number of pairs should be considered, Figure 23 decidedly makes the cases that it is the interaction of number of items and number of LD pairs (as indicated in Table 13) which should be further explored. In examining the JSI values within 10 item, 2PLM cells (which, from the graphical display, are the most affected by the issue) more closely (e.g., considering initial parameter standard errors, JSI values prior to summing, etc.) it became apparent that the drop in power was primarily a result of the within-replication standard deviation that increased, in some cells rather dramatically, primarily due to large negative summed JSI values that appeared for item pairs that contained a locally dependent item and an item that was not part of the same locally dependent pair (e.g., a locally independent item or a locally dependent item that was a member of a different pair of items simulated to exhibit LD). As noted previously, the summed JSI values were tested against a value that was set at the mean plus two standard deviations; as the standard deviation increases due to large negative numbers, it
makes flagging the positive values (which, from Formula 12, are those expected to be exhibiting LD) less likely.
Chapter 7: Discussion

7.1: 2-Parameter Logistic Model Discussion

In terms of parameter recovery, there are two main points that warrant attention within the 2PLM results: the sheer number of replications that had to be discarded due to non-convergence in the SLD $\pi_{LD} = 0.8$ cells and the identification of a set of 40 items cells as having the poorest parameter recovery for items exhibiting ULD. With regards to the SLD cells, the failure of MULTILOG to converge to stable estimates for many of the datasets generated to exhibit severe LD may be viewed as a positive aspect of the program. In cells with extreme levels of SLD, especially those with few items and small sample sizes, the failure of MULTILOG to obtain reasonable parameter estimates (e.g., discarded non-converged datasets had “final” discrimination parameter estimates that went as high as 169) served as a method of LD detection without having to resort to post-hoc programs or subroutines to calculate LD indices. While the results reported here discarded such non-converged solutions, it is useful to note that in the most extreme conditions, where LD is having the greatest effect on parameter estimates for locally independent and LD items alike, it can be extremely obvious from just the estimated parameter values that something unexpected is occurring within the data set.

It was generally expected, as was observed in the SLD cells, that the poorest parameter recovery for a set of cells within the ULD condition would occur in cells that were limited in both the number of items and the number of simulees, but this was not the
case. The specific set of cells (i.e., 40 items, N = 250, and slope ratio = 1.5, pairs = 1, 2, and 4) identified within the 2PLM ULD cells as displaying the poorest LD item parameter recovery had the longest scale length possible within the simulation design, which was rather surprising. It should be noted, however, that these cells were “worst” by a somewhat narrow margin; cells that had the same sample size, slope ratio, and number of pairs but only 20 items had average slope parameter RMSE values which were smaller by a negligible margin, in practical terms (i.e., 0.04 less). Inspecting the whole of the RMSE values for the ULD cells, it appears that as the non-locally dependent items are being better estimated (with increasing number of items and sample size) the items in locally dependent pairs are unable to exert as much influence on the definition of the latent variable. That is, as the number of locally independent items increases, their shared variance becomes the variance that the model fits itself to and the LD item parameters, which have unaccounted for variance stemming from the specific factor upon which they also load, become less well estimated due to the “noise” in the locally dependent item responses from the unmodeled factor.

Turning to the power analyses, taken across both ULD and SLD cells, the LD index with the most desirable performance (i.e., high power and low/near nominal level Type I error rates) was the intermediate residual z values, obtained from Equation 14 in the process of calculating $\chi^2_{G/D}$. It should be noted that these z values were assessed independent of any conclusion of LD as determined by the final $\chi^2_{G/D}$ value; if a significant approximate chi-square value was a pre-requisite for examining the z values, their observed levels of power would have been significantly reduced. Other indices
which were adept at LD detection and maintaining low Type I error rates were Chen and Thissen’s (1997) $G^2$ and the JSI of Edwards and Cai (in prep). The r-to-z transformed $Q_3$, while displaying extremely high levels of power also had inflated Type I error rates and, as previously reported (e.g., Chen & Thissen, 1997), failed to follow the theoretical distribution in the null conditions. However, the index with the poorest performance, by far, was $\chi^2_{G/D}$, introduced by Gessaroli and DeChamplain (1996).

While the performance of $\chi^2_{G/D}$ had been found to be adequate to good in previous evaluations (e.g., DeChamplain & Gessaroli, 1996; Gessaroli & DeChamplain, 1996) in this study, it faired rather poorly. The power across all simulation conditions was less than 0.40 and in the Null cells the statistic failed to approximate the theoretical null distribution. Statistically, the most prominent predictors of the index’s power were number of items and degree of local dependence, but during a visual inspection across the individual cells, high power was observed in cells where a high degree of LD was paired with both a large sample and a high number of locally dependent pairs. In this restricted subset of simulation cells, observed power was generally 0.90 or above; this finding is more congruent with the previous research.

This disparity between the overall findings and the selected subset may be a result of the factor loading patterns used for the non-unidimensional data simulated in previous studies. For example, in Gessaroli and DeChamplain (1996) the multidimensional data were generated from two factors upon which no items cross-loaded (i.e., items conformed to an independent clusters solution). Using this structure, the authors found overall high power and low Type I error, as well as reporting that power decreased as fewer items...
loaded on the second factor (that is, when the second factor was less well-defined). In this study, a bifactor-type generating MIRT model or the SLD conceptualization was used, both of which also result in factors that are less defined — in the MIRT model, specific factors only have two items loading upon them, and in SLD, there is no “true” second factor. In terms of covariance among items, cells with numerous pairs of LD items with large specific slope values most likely come closest to approximating the covariance observed from a generating independent clusters pattern. If this is the case, it provides an explanation as to why the overall results differ so drastically from previous findings as well as the performance difference between all cells compared to the selected subset of cells, in which $\chi^2_{G/D}$ performs closer to expectations. In the selected subset of cells, when $\chi^2_{G/D}$ was applied to data more similar to the structure it was initially intended to detect, high power and low Type I error were observed; in other cells, especially those with a lesser degree of LD and fewer pairs of locally dependent items, power was diminished, in some cases to essentially zero, which also mirrors the previously detected trend.

7.2: Graded Response Model Discussion

As was the case in the 2PLM results, the ULD parameter recovery data showed that a set of cells containing 40 items was, again, most poorly recovered. As a possible explication for this was detailed in the 2PLM discussion, and the rationale here would be the same, expanding further on this finding would be redundant.

For the four indices with known distributional properties in the Null condition, the theoretical distributions were generally well followed, although $G^2$ displayed inflated
standard deviations with smaller sample sizes, especially when combined with a short scale. The r-to-z transformed $Q_3$, which previous research using dichotomous data (e.g., Chen & Thissen, 1997; Kim, 2007) had found to diverge noticeably from the expected mean and standard deviation, behaved rather well in conjunction with the GRM, especially when paired with longer scales.

With respect to detecting LD when it existed, the best performance was associated with $G^2$ and JSI indices. Over all SLD and ULD cells, both indices had average power values that were over 0.87 and Type I error rates that were low (and for the $G^2$ close to, although somewhat above, the nominal level). The modification indices from the structural equation modeling framework, while displaying high power were also associated with an elevated Type I error rate, reducing their usefulness as an LD detection measure. The r-to-z transformed $Q_3$ had high power across both SLD and ULD, but the elevated average Type I error rate found in the SLD cells lead to questions regarding its continued use.

7.3: Ancillary Discussion

It is possible that the commonly reported deficit in power for detecting ULD is a result of non-congruent degrees of LD being induced across LD types, a hypothesis that was supported by the observed polychoric correlations across the highest levels of induced SLD and ULD in this study. While the current results provide an indication that larger specific-to-general slope ratio values are needed to obtain polychoric correlations equivalent to those seen when $\pi_{LD} = 0.8$, a more comprehensive comparison of SLD/ULD when assessed on the common metric of polychoric correlation is needed. Additional
simulations to find the degree of ULD needed to achieve a comparable polychoric correlation between the items in a pair of locally dependent items and the effect this higher correlation has on the performance of the LD indices with ULD are currently underway.

Further work is also needed in determining the optimum method of employing the calculated JSI values for detecting locally dependent pairs. As noted in the Ancillary Results, the current method of summing JSI values and comparing those to an *ad hoc* cut point displays limitations when more than one pair of locally dependent items is present in the dataset. As seen in Table 23, this problem is somewhat exacerbated as degree of LD increases, which results in the JSI displaying poor power in simulation cells where the LD stands out upon visual inspection of item parameters and should be easily detected. Finding an alternative way of automating the assessment of JSI values that rectifies the current limitation would serve to make this already-promising LD index a stand-alone general purpose LD detection device, recommended without qualifications regarding its performance.

**7.4: General Discussion**

Across both the 2PLM and GRM, it appears that the indices that recommend themselves with consistently high power and low Type I error rates are the JSI and the $G^2$. The residual $z$ values, which displayed the best performance with respect to the 2PLM, are limited by the fact that they are only available for use with dichotomous data. As noted in the Section 7.3, the JSI does have a weakness when there are several pairs of items exhibiting LD in a dataset. Further research may rectify this limitation of the JSI
but a current recommendation is to combine the use of the JSI with a second index that performed well when there was more than one pair of locally dependent items. Speaking generally across all the included simulation conditions and IRT models used here, the best supplementary measure to the JSI is $G^2$. Both measures are available for use with both the GRM and 2PLM, and $G^2$ displayed both high power and reasonable Type I error rates in the cells where the JSI experienced difficulties (i.e., those with more than one pair of locally dependent items). With respect to the 2PLM specifically, $\chi^2_{G/D}$ is recommended as the supplementary index; the set of cells where Gessaroli and DeChamplain’s (1996) approximate chi-square is lacking is exactly where the JSI is strongest and vice versa. Additionally, $\chi^2_{G/D}$ displayed an extremely low Type I error rate (lower than that of $G^2$) meaning the combination of both the JSI and $\chi^2_{G/D}$ should have a Type I error rate near that which was seen when using the JSI alone.

As with any simulation study, the results are generalizable only to the extent that the design variables (and the levels of those variables) mirror conditions in the real world. Although attention was paid to finding and employing generating parameters based on the analysis of existing scales and levels of the simulation variables (e.g., test length, sample size) were chosen to be in line with what is seen in datasets typical of psychological assessment, future research could expand our knowledge of how the selected LD indices perform under additional conditions or using other IRT models to generate and fit the data (e.g., 3PLM, multiple choice model, etc).

More general areas of future research could include detailed comparisons between programs that employ full versus limited information estimation routines, as some
differences were noticed during the simulation such as discrimination parameters estimates being less suppressed in the ULD cells when estimated using a limited information estimator within Mplus. An additional area of interest, which may become more important as IRT becomes more prominent and general purpose statistical programs (e.g., SPSS, LISREL, etc) are updated to provide estimates of IRT parameters, is if existing measures of fit commonly reported in the structural equation modeling framework, such as the root mean square error of approximation (RMSEA) or standardized root mean residual (SRMR) can be used or adapted for detecting LD. If findings are able to support fit measures being employed in such a way, this adaptation could allow practitioners to check for the violation of the LI assumption (a necessary step in any IRT analysis) without having to resort to specialized IRT software, with which they may be unfamiliar.
References


demonstration using items from the pediatric quality of life inventory (PedsQL) 4.0 generic core scales. *Medical Care, 45*, S39-S47.


### Appendix A: Tables

**Table 1a**

*Example observed contingency table*

<table>
<thead>
<tr>
<th>Item i</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item j</td>
<td>O(_{11})</td>
<td>O(_{12})</td>
<td>O(_{13})</td>
</tr>
<tr>
<td></td>
<td>O(_{21})</td>
<td>O(_{22})</td>
<td>O(_{23})</td>
</tr>
</tbody>
</table>

**Table 1b**

*Example model-predicted contingency table*

<table>
<thead>
<tr>
<th>Item i</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item j</td>
<td>E(_{11})</td>
<td>E(_{12})</td>
<td>E(_{13})</td>
</tr>
<tr>
<td></td>
<td>E(_{21})</td>
<td>E(_{22})</td>
<td>E(_{23})</td>
</tr>
</tbody>
</table>
Table 2a

*Raw JSI value matrix*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[2,]</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>[3,]</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>[4,]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>[5,]</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Table 2b

*Summed JSI upper triangle matrix*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[2,]</td>
<td></td>
<td>4</td>
<td>0</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>[3,]</td>
<td></td>
<td></td>
<td>0</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>[4,]</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>[5,]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3

*Independent variables manipulated in simulation study*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale length</td>
<td>$k = 10, 20, 40$</td>
</tr>
<tr>
<td>Respondents</td>
<td>$N = 250, 500, 1000$</td>
</tr>
<tr>
<td>Locally dependent pairs</td>
<td>$p = 1, 2, 4$</td>
</tr>
<tr>
<td>Type of local dependence</td>
<td>None</td>
</tr>
<tr>
<td>Degree of local dependence</td>
<td>Surface</td>
</tr>
<tr>
<td></td>
<td>$\pi_{LD} = 0.3, 0.5, 0.8$</td>
</tr>
<tr>
<td>Underlying</td>
<td>Slope ratio $= 0.5, 1.0, 1.5$</td>
</tr>
<tr>
<td>Response categories</td>
<td>$C = 2, 5$</td>
</tr>
</tbody>
</table>
Table 4

*RMSE for Null 2PLM by items and sample size*

<table>
<thead>
<tr>
<th>Average RMSE</th>
<th>10 items</th>
<th>20 items</th>
<th>40 items</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| N = 250 | 0.479 | 0.186 |
| N = 500 | 0.318 | 0.131 |
| N = 1000 | 0.226 | 0.092 |

| N = 250 | 0.441 | 0.181 |
| N = 500 | 0.284 | 0.119 |
| N = 1000 | 0.203 | 0.087 |

| N = 250 | 0.378 | 0.173 |
| N = 500 | 0.262 | 0.121 |
| N = 1000 | 0.193 | 0.088 |

| Overall | 0.309 | 0.131 |
Table 5

*RMSE for cells SLD 2PLM, k = 10, N = 250, \( \pi_{LD} = 0.8, p = 1, 2, \text{ and } 4 \)*

<table>
<thead>
<tr>
<th>Average RMSE</th>
<th>1 LD pair</th>
<th>2 LD pairs</th>
<th>4 LD pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>b₁</td>
<td></td>
</tr>
<tr>
<td>Null Items only</td>
<td>0.462</td>
<td>0.321</td>
<td>0.434</td>
</tr>
<tr>
<td>LD Items only</td>
<td>8.487</td>
<td>0.757</td>
<td>7.804</td>
</tr>
<tr>
<td>Null Items only</td>
<td>0.434</td>
<td>0.271</td>
<td>0.436</td>
</tr>
<tr>
<td>LD Items only</td>
<td>7.804</td>
<td>0.703</td>
<td>5.429</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Average RMSE</th>
<th>1 LD pair</th>
<th>2 LD pairs</th>
<th>4 LD pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>b₁</td>
<td></td>
</tr>
<tr>
<td>Null Items only</td>
<td>0.462</td>
<td>0.321</td>
<td>0.434</td>
</tr>
<tr>
<td>LD Items only</td>
<td>8.487</td>
<td>0.757</td>
<td>7.804</td>
</tr>
<tr>
<td>Null Items only</td>
<td>0.434</td>
<td>0.271</td>
<td>0.436</td>
</tr>
<tr>
<td>LD Items only</td>
<td>7.804</td>
<td>0.703</td>
<td>5.429</td>
</tr>
</tbody>
</table>
Table 6

*Group means for the significant items by degree interaction for 2PLM α-parameter RMSE*

<table>
<thead>
<tr>
<th>Average RMSE</th>
<th>subset 1</th>
<th>subset 2</th>
<th>subset 3</th>
<th>subset 4</th>
<th>subset 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$π_{LD} = 0.3$</td>
<td>0.556</td>
<td></td>
<td>0.988</td>
<td></td>
<td>7.077</td>
</tr>
<tr>
<td>$π_{LD} = 0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$π_{LD} = 0.8$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$π_{LD} = 0.3$</td>
<td>0.409</td>
<td></td>
<td>0.409</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$π_{LD} = 0.5$</td>
<td>0.505</td>
<td></td>
<td>0.505</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$π_{LD} = 0.8$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.269</td>
</tr>
<tr>
<td>40 items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$π_{LD} = 0.3$</td>
<td></td>
<td></td>
<td></td>
<td>0.374</td>
<td></td>
</tr>
<tr>
<td>$π_{LD} = 0.5$</td>
<td></td>
<td></td>
<td>0.417</td>
<td>0.417</td>
<td></td>
</tr>
<tr>
<td>$π_{LD} = 0.8$</td>
<td></td>
<td>0.475</td>
<td>0.475</td>
<td>0.475</td>
<td></td>
</tr>
</tbody>
</table>

Note. Means within subsets are not significantly different. All other differences are statistically significantly different (all adjusted ps < .05)
Table 7

*RMSE for cells ULD 2PLM, k = 40, N = 250, slope ratio = 1.5, p = 1, 2, and 4*

<table>
<thead>
<tr>
<th>Average RMSE</th>
<th>a</th>
<th>b₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 LD pair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null Items only</td>
<td>0.391</td>
<td>0.213</td>
</tr>
<tr>
<td>LD Items only</td>
<td>0.771</td>
<td>0.258</td>
</tr>
<tr>
<td>2 LD pairs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null Items only</td>
<td>0.371</td>
<td>0.195</td>
</tr>
<tr>
<td>LD Items only</td>
<td>0.772</td>
<td>0.278</td>
</tr>
<tr>
<td>4 LD pairs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null Items only</td>
<td>0.386</td>
<td>0.175</td>
</tr>
<tr>
<td>LD Items only</td>
<td>0.770</td>
<td>0.257</td>
</tr>
</tbody>
</table>
Table 8

*Group means for the significant item by degree interaction for 2PLM a-parameter RMSE*

<table>
<thead>
<tr>
<th>Average RMSE</th>
<th>subset 1</th>
<th>subset 2</th>
<th>subset 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SLD</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 items</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 0.5$</td>
<td>0.556</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 1.0$</td>
<td></td>
<td>0.988</td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 1.5$</td>
<td></td>
<td></td>
<td>1.078</td>
</tr>
<tr>
<td>20 items</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 0.5$</td>
<td>0.409</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 1.0$</td>
<td></td>
<td>0.505</td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 1.5$</td>
<td></td>
<td></td>
<td>1.269</td>
</tr>
<tr>
<td>40 items</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 0.5$</td>
<td>0.374</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 1.0$</td>
<td></td>
<td>0.417</td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 1.5$</td>
<td></td>
<td>0.475</td>
<td></td>
</tr>
<tr>
<td><strong>ULD</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 items</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 0.5$</td>
<td>0.312</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 1.0$</td>
<td></td>
<td>0.429</td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 1.5$</td>
<td></td>
<td>0.630</td>
<td>0.630</td>
</tr>
<tr>
<td>20 items</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 0.5$</td>
<td>0.289</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 1.0$</td>
<td></td>
<td>0.486</td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 1.5$</td>
<td></td>
<td>0.724</td>
<td>0.724</td>
</tr>
<tr>
<td>40 items</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 0.5$</td>
<td>0.284</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 1.0$</td>
<td></td>
<td>0.514</td>
<td></td>
</tr>
<tr>
<td>$a_s/a_g = 1.5$</td>
<td></td>
<td>0.769</td>
<td>0.769</td>
</tr>
</tbody>
</table>

Note. Means for subsets are not significantly different. All other differences are statistically significantly different (all adjusted ps < .05)
Table 9

*Type I error rates for Null 2PLM by number of items and sample size*

<table>
<thead>
<tr>
<th>Type I Error</th>
<th>G^2</th>
<th>χ^2_{G/H}</th>
<th>JSI</th>
<th>Mls</th>
<th>Q_3</th>
<th>r-to-z Q_3</th>
<th>residual z</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N = 250</td>
<td>0.009</td>
<td>0.000</td>
<td>0.017</td>
<td>0.033</td>
<td>0.000</td>
<td>0.139</td>
<td>0.037</td>
</tr>
<tr>
<td>N = 500</td>
<td>0.009</td>
<td>0.000</td>
<td>0.018</td>
<td>0.041</td>
<td>0.000</td>
<td>0.194</td>
<td>0.037</td>
</tr>
<tr>
<td>N = 1000</td>
<td>0.010</td>
<td>0.000</td>
<td>0.019</td>
<td>0.042</td>
<td>0.000</td>
<td>0.262</td>
<td>0.036</td>
</tr>
<tr>
<td>20 items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N = 250</td>
<td>0.015</td>
<td>0.000</td>
<td>0.024</td>
<td>0.107</td>
<td>0.615</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td>N = 500</td>
<td>0.015</td>
<td>0.000</td>
<td>0.025</td>
<td>0.024</td>
<td>0.106</td>
<td>0.627</td>
<td>0.044</td>
</tr>
<tr>
<td>N = 1000</td>
<td>0.017</td>
<td>0.000</td>
<td>0.024</td>
<td>0.096</td>
<td>0.702</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td>40 items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N = 250</td>
<td>0.018</td>
<td>0.000</td>
<td>0.027</td>
<td>0.005</td>
<td>0.388</td>
<td>0.715</td>
<td>0.046</td>
</tr>
<tr>
<td>N = 500</td>
<td>0.019</td>
<td>0.000</td>
<td>0.028</td>
<td>0.010</td>
<td>0.392</td>
<td>0.782</td>
<td>0.046</td>
</tr>
<tr>
<td>N = 1000</td>
<td>0.021</td>
<td>0.000</td>
<td>0.028</td>
<td>0.013</td>
<td>0.397</td>
<td>0.817</td>
<td>0.046</td>
</tr>
<tr>
<td>Overall</td>
<td>0.015</td>
<td>0.000</td>
<td>0.023</td>
<td>0.024</td>
<td>0.167</td>
<td>0.528</td>
<td>0.042</td>
</tr>
</tbody>
</table>
Table 10

Means and SDs of LD indices using distributional critical values for testing in 2PLM Null cells

<table>
<thead>
<tr>
<th>M</th>
<th>10</th>
<th>20</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>(SD)</td>
<td>250</td>
<td>500</td>
<td>1000</td>
</tr>
<tr>
<td>$G^2$</td>
<td>0.58</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>(0.82)</td>
<td>(0.85)</td>
<td>(0.85)</td>
</tr>
<tr>
<td>$\chi^2_{G/D}$</td>
<td>17.69</td>
<td>18.00</td>
<td>18.11</td>
</tr>
<tr>
<td>MIs</td>
<td>0.91</td>
<td>1.04</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>(1.63)</td>
<td>(2.09)</td>
<td>(2.16)</td>
</tr>
<tr>
<td>r-to-z</td>
<td>0.41</td>
<td>0.59</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>(1.22)</td>
<td>(1.37)</td>
<td>(1.61)</td>
</tr>
</tbody>
</table>

Note. Expected value (SD) for $G^2$ and MIs are 1 (1.41). Expected value (SD) for 10, 20, and 40 items GD $\chi^2$ are 35 (8.37), 170 (18.44), and 740 (38.47), respectively. Expected value (SD) for r-to-z $Q_3$ are 0 (1.0).
Table 11

*Overall power and Type I error rates for SLD and ULD 2PLM cells*

<table>
<thead>
<tr>
<th>Index</th>
<th>SLD</th>
<th></th>
<th></th>
<th>ULD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power Mean (SD)</td>
<td>Type I error Mean (SD)</td>
<td>Power Mean (SD)</td>
<td>Type I error Mean (SD)</td>
<td></td>
</tr>
<tr>
<td>$G^2$</td>
<td>.934 (.118)</td>
<td>.065 (.103)</td>
<td>.752 (.301)</td>
<td>.025 (.025)</td>
<td></td>
</tr>
<tr>
<td>$X^2_{G/D}$</td>
<td>.455 (.451)</td>
<td>--</td>
<td>.257 (.378)</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>JSI</td>
<td>.857 (.249)</td>
<td>.010 (.009)</td>
<td>.666 (.264)</td>
<td>.018 (.009)</td>
<td></td>
</tr>
<tr>
<td>MIs</td>
<td>.986 (.035)</td>
<td>.196 (.184)</td>
<td>.779 (.279)</td>
<td>.097 (.116)</td>
<td></td>
</tr>
<tr>
<td>$Q_3$</td>
<td>.767 (.259)</td>
<td>.041 (.057)</td>
<td>.473 (.364)</td>
<td>.000 (.000)</td>
<td></td>
</tr>
<tr>
<td>r-to-z $Q_3$</td>
<td>.944 (.122)</td>
<td>.317 (.217)</td>
<td>.869 (.187)</td>
<td>.111 (.071)</td>
<td></td>
</tr>
<tr>
<td>residual z</td>
<td>.973 (.050)</td>
<td>.015 (.014)</td>
<td>.764 (.260)</td>
<td>.025 (.017)</td>
<td></td>
</tr>
</tbody>
</table>

Note: $X^2_{G/D}$ does not have Type I error as it only provides an overall test.
Table 12

*Prominent F values for predictors of power and Type I error, in conjunction with the 2PLM, of LD indices with SLD*

<table>
<thead>
<tr>
<th>Largest F values</th>
<th>G$^2$</th>
<th>$\chi^2_{G/D}$</th>
<th>JSI</th>
<th>MI</th>
<th>Q$_3$</th>
<th>Q$_3z$</th>
<th>res Z</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Power</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1$^{st}$</td>
<td>items</td>
<td>items</td>
<td>items</td>
<td>degree</td>
<td>items</td>
<td>items</td>
<td>items</td>
</tr>
<tr>
<td>2$^{nd}$</td>
<td>items* degree</td>
<td>degree</td>
<td>items* degree</td>
<td>n</td>
<td>degree</td>
<td>items* degree</td>
<td>degree</td>
</tr>
<tr>
<td>3$^{rd}$</td>
<td>degree</td>
<td>n* degree</td>
<td>degree</td>
<td>items* degree</td>
<td>degree</td>
<td>n</td>
<td></td>
</tr>
<tr>
<td>4$^{th}$</td>
<td>pairs</td>
<td>pairs</td>
<td>pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Type I error</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1$^{st}$</td>
<td>items</td>
<td>items</td>
<td>items</td>
<td>items</td>
<td>items</td>
<td>items</td>
<td>items</td>
</tr>
<tr>
<td>2$^{nd}$</td>
<td>degree</td>
<td>degree</td>
<td>degree</td>
<td>n</td>
<td>n</td>
<td>degree</td>
<td></td>
</tr>
<tr>
<td>3$^{rd}$</td>
<td>items* degree</td>
<td>n</td>
<td>items</td>
<td>*n</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. $\chi^2_{G/D}$ provides only an overall test and, therefore, has no Type I error rate in cells with LD. * denotes interaction.
Table 13

Prominent predictors of power and Type I error of LD indices, in conjunction with the 2PLM and ULD

<table>
<thead>
<tr>
<th>Largest F values</th>
<th>$G^2$</th>
<th>$\chi^2_{G/D}$</th>
<th>JSI</th>
<th>MI</th>
<th>$Q_3$</th>
<th>$Q_{3,z}$</th>
<th>res Z</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>degree</td>
<td>degree</td>
<td>degree</td>
<td>degree</td>
<td>degree</td>
<td>degree</td>
<td>degree</td>
</tr>
<tr>
<td>2nd</td>
<td>n</td>
<td>items</td>
<td>n</td>
<td>items</td>
<td>n</td>
<td>n* degree</td>
<td>n</td>
</tr>
<tr>
<td>3rd</td>
<td>items*</td>
<td>degree</td>
<td>n*</td>
<td>degree</td>
<td>n</td>
<td>degree</td>
<td></td>
</tr>
<tr>
<td>4th</td>
<td>items*</td>
<td>pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type I error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>degree</td>
<td>items</td>
<td>items</td>
<td>items</td>
<td>items</td>
<td>degree</td>
<td></td>
</tr>
<tr>
<td>2nd</td>
<td>items</td>
<td>degree</td>
<td>n</td>
<td>n</td>
<td>items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd</td>
<td>items*</td>
<td>degree</td>
<td>items*</td>
<td>n</td>
<td>items*</td>
<td>n</td>
<td></td>
</tr>
</tbody>
</table>

Note. $\chi^2_{G/D}$ provides only an overall test and, therefore, has no Type I error rate in cells with LD. * denotes interaction.
Table 14

*RMSE for Null GRM by items and sample size*

<table>
<thead>
<tr>
<th>Average RMSE</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10 items</td>
<td>0.162</td>
<td>0.152</td>
<td>0.098</td>
<td>0.100</td>
<td>0.235</td>
</tr>
<tr>
<td>N = 250</td>
<td>0.218</td>
<td>0.211</td>
<td>0.134</td>
<td>0.138</td>
<td>0.371</td>
</tr>
<tr>
<td>N = 500</td>
<td>0.156</td>
<td>0.145</td>
<td>0.094</td>
<td>0.096</td>
<td>0.229</td>
</tr>
<tr>
<td>N = 1000</td>
<td>0.111</td>
<td>0.101</td>
<td>0.067</td>
<td>0.067</td>
<td>0.106</td>
</tr>
<tr>
<td>20 items</td>
<td>0.150</td>
<td>0.150</td>
<td>0.098</td>
<td>0.099</td>
<td>0.217</td>
</tr>
<tr>
<td>N = 250</td>
<td>0.203</td>
<td>0.208</td>
<td>0.133</td>
<td>0.134</td>
<td>0.362</td>
</tr>
<tr>
<td>N = 500</td>
<td>0.144</td>
<td>0.142</td>
<td>0.094</td>
<td>0.095</td>
<td>0.172</td>
</tr>
<tr>
<td>N = 1000</td>
<td>0.104</td>
<td>0.099</td>
<td>0.066</td>
<td>0.068</td>
<td>0.117</td>
</tr>
<tr>
<td>40 items</td>
<td>0.142</td>
<td>0.146</td>
<td>0.095</td>
<td>0.097</td>
<td>0.153</td>
</tr>
<tr>
<td>N = 250</td>
<td>0.192</td>
<td>0.200</td>
<td>0.128</td>
<td>0.132</td>
<td>0.211</td>
</tr>
<tr>
<td>N = 500</td>
<td>0.136</td>
<td>0.139</td>
<td>0.091</td>
<td>0.092</td>
<td>0.145</td>
</tr>
<tr>
<td>N = 1000</td>
<td>0.097</td>
<td>0.098</td>
<td>0.065</td>
<td>0.066</td>
<td>0.104</td>
</tr>
<tr>
<td>Overall</td>
<td>0.151</td>
<td>0.149</td>
<td>0.096</td>
<td>0.098</td>
<td>0.201</td>
</tr>
</tbody>
</table>
Table 15

*RMSE for cells SLD GRM, k = 10, N = 250, π_{LD} = 0.8, p = 1, 2, and 4*

<table>
<thead>
<tr>
<th>Average RMSE</th>
<th>$a$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$b_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 LD pair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null Items only</td>
<td>0.271</td>
<td>0.246</td>
<td>0.147</td>
<td>0.142</td>
<td>0.252</td>
</tr>
<tr>
<td>LD Items only</td>
<td>4.698</td>
<td>0.471</td>
<td>0.379</td>
<td>0.376</td>
<td>0.524</td>
</tr>
<tr>
<td>2 LD pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null Items only</td>
<td>0.294</td>
<td>0.313</td>
<td>0.167</td>
<td>0.165</td>
<td>0.513</td>
</tr>
<tr>
<td>LD Items only</td>
<td>3.776</td>
<td>0.482</td>
<td>0.386</td>
<td>0.383</td>
<td>0.466</td>
</tr>
<tr>
<td>4 LD pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null Items only</td>
<td>0.283</td>
<td>0.310</td>
<td>0.162</td>
<td>0.164</td>
<td>0.309</td>
</tr>
<tr>
<td>LD Items only</td>
<td>1.744</td>
<td>0.435</td>
<td>0.367</td>
<td>0.371</td>
<td>0.513</td>
</tr>
</tbody>
</table>
Table 16

*RMSE for cells ULD GRM k = 40, N = 250, slope ratio = 1.5, p = 1, 2 and 4*

<table>
<thead>
<tr>
<th>Average RMSE</th>
<th>$a$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$b_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 LD pair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null Items only</td>
<td>0.195</td>
<td>0.209</td>
<td>0.133</td>
<td>0.135</td>
<td>0.298</td>
</tr>
<tr>
<td>LD Items only</td>
<td>0.779</td>
<td>0.289</td>
<td>0.196</td>
<td>0.205</td>
<td>0.300</td>
</tr>
<tr>
<td>2 LD pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null Items only</td>
<td>0.196</td>
<td>0.209</td>
<td>0.132</td>
<td>0.136</td>
<td>0.316</td>
</tr>
<tr>
<td>LD Items only</td>
<td>0.786</td>
<td>0.290</td>
<td>0.199</td>
<td>0.214</td>
<td>0.321</td>
</tr>
<tr>
<td>4 LD pairs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null Items only</td>
<td>0.196</td>
<td>0.211</td>
<td>0.132</td>
<td>0.133</td>
<td>0.335</td>
</tr>
<tr>
<td>LD Items only</td>
<td>0.778</td>
<td>0.303</td>
<td>0.208</td>
<td>0.202</td>
<td>0.300</td>
</tr>
</tbody>
</table>
### Table 17

**Type I error rates for Null GRM by items and sample size**

<table>
<thead>
<tr>
<th>Type I Error</th>
<th>$G^2$</th>
<th>JSI</th>
<th>Mls</th>
<th>$Q_3$</th>
<th>r-to-z $Q_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N = 250</td>
<td>0.054</td>
<td>0.017</td>
<td>0.032</td>
<td>0.000</td>
<td>0.123</td>
</tr>
<tr>
<td>N = 500</td>
<td>0.046</td>
<td>0.017</td>
<td>0.025</td>
<td>0.000</td>
<td>0.088</td>
</tr>
<tr>
<td>N = 1000</td>
<td>0.064</td>
<td>0.018</td>
<td>0.038</td>
<td>0.000</td>
<td>0.117</td>
</tr>
<tr>
<td>20 items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N = 250</td>
<td>0.062</td>
<td>0.021</td>
<td>0.017</td>
<td>0.000</td>
<td>0.073</td>
</tr>
<tr>
<td>N = 500</td>
<td>0.051</td>
<td>0.022</td>
<td>0.007</td>
<td>0.000</td>
<td>0.062</td>
</tr>
<tr>
<td>N = 1000</td>
<td>0.076</td>
<td>0.021</td>
<td>0.016</td>
<td>0.000</td>
<td>0.071</td>
</tr>
<tr>
<td>40 items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N = 250</td>
<td>0.068</td>
<td>0.022</td>
<td>0.003</td>
<td>0.000</td>
<td>0.059</td>
</tr>
<tr>
<td>N = 500</td>
<td>0.055</td>
<td>0.022</td>
<td>0.000</td>
<td>0.000</td>
<td>0.056</td>
</tr>
<tr>
<td>N = 1000</td>
<td>0.084</td>
<td>0.021</td>
<td>0.001</td>
<td>0.000</td>
<td>0.059</td>
</tr>
<tr>
<td>Overall</td>
<td>0.061</td>
<td>0.020</td>
<td>0.017</td>
<td>0.000</td>
<td>0.085</td>
</tr>
</tbody>
</table>
Table 18

*Means and SDs of LD indices using distributional critical values for testing in GRM Null cell*

<table>
<thead>
<tr>
<th></th>
<th>M (SD)</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>10</th>
<th>20</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>G²</td>
<td></td>
<td>16.55</td>
<td>16.68</td>
<td>17.33</td>
<td>16.82</td>
<td>16.86</td>
<td>17.58</td>
<td>16.66</td>
<td>16.99</td>
<td>17.75</td>
<td>(10.10)</td>
<td>(5.77)</td>
<td>(5.96)</td>
<td>(14.03)</td>
<td>(5.74)</td>
<td>(5.99)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(10.10)</td>
<td>(5.77)</td>
<td>(5.96)</td>
<td>(14.03)</td>
<td>(5.74)</td>
<td>(5.99)</td>
<td>(5.60)</td>
<td>(5.75)</td>
<td>(6.00)</td>
<td>(1.18)</td>
<td>(1.28)</td>
<td>(1.38)</td>
<td>(0.76)</td>
<td>(0.99)</td>
<td>(1.15)</td>
</tr>
<tr>
<td>MIs</td>
<td></td>
<td>0.82</td>
<td>0.89</td>
<td>0.96</td>
<td>0.52</td>
<td>0.69</td>
<td>0.81</td>
<td>0.22</td>
<td>0.36</td>
<td>0.52</td>
<td>(1.18)</td>
<td>(1.28)</td>
<td>(1.38)</td>
<td>(0.76)</td>
<td>(0.99)</td>
<td>(1.15)</td>
</tr>
<tr>
<td>r-to-z</td>
<td>Q₃</td>
<td>0.29</td>
<td>0.41</td>
<td>0.58</td>
<td>0.10</td>
<td>0.15</td>
<td>0.21</td>
<td>0.04</td>
<td>0.06</td>
<td>0.08</td>
<td>(1.10)</td>
<td>(1.16)</td>
<td>(1.27)</td>
<td>(1.04)</td>
<td>(1.06)</td>
<td>(1.09)</td>
</tr>
</tbody>
</table>

Note. Expected value (SD) for G² is 16 (5.66). Expected value (SD) for MIs is 1 (1.41). Expected value (SD) for r-to-z Q₃ is 0 (1).
Table 19

*Overall power and Type I error rates for SLD and ULD GRM cells*

<table>
<thead>
<tr>
<th>Index</th>
<th>SLD Power Mean (SD)</th>
<th>SLD Type I error Mean (SD)</th>
<th>ULD Power Mean (SD)</th>
<th>ULD Type I error Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G^2$</td>
<td>.999 (.003)</td>
<td>.085 (.043)</td>
<td>.803 (.267)</td>
<td>.071 (.013)</td>
</tr>
<tr>
<td>JSI</td>
<td>.911 (.187)</td>
<td>.007 (.007)</td>
<td>.841 (.189)</td>
<td>.015 (.008)</td>
</tr>
<tr>
<td>MIs</td>
<td>.996 (.011)</td>
<td>.235 (.235)</td>
<td>.888 (.184)</td>
<td>.131 (.187)</td>
</tr>
<tr>
<td>$Q_3$</td>
<td>.844 (.214)</td>
<td>.001 (.004)</td>
<td>.718 (.355)</td>
<td>.001 (.003)</td>
</tr>
<tr>
<td>r-to-z $Q_3$</td>
<td>.988 (.033)</td>
<td>.117 (.099)</td>
<td>.948 (.086)</td>
<td>.080 (.028)</td>
</tr>
</tbody>
</table>
Table 20

*Prominent predictors of power and Type 1 error of LD indices with SLD*

<table>
<thead>
<tr>
<th>Largest F values</th>
<th>$G^2$</th>
<th>JSI</th>
<th>MI</th>
<th>$Q_3$</th>
<th>r-to-z $Q_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Power</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>N</td>
<td>items</td>
<td>N</td>
<td>degree</td>
<td>items</td>
</tr>
<tr>
<td>2nd</td>
<td>degree</td>
<td>items*degree</td>
<td>degree</td>
<td>items</td>
<td>items*degree</td>
</tr>
<tr>
<td>3rd</td>
<td>N*degree</td>
<td>degree</td>
<td>N*degree</td>
<td>items*degree</td>
<td>degree</td>
</tr>
<tr>
<td>4th</td>
<td>pairs</td>
<td></td>
<td>pairs</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Type I error</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>N</td>
<td>items</td>
<td>items</td>
<td>items*degree</td>
<td>items</td>
</tr>
<tr>
<td>2nd</td>
<td>degree</td>
<td></td>
<td>items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd</td>
<td>degree</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. * denotes interaction.
Table 21

Prominent predictors of power and Type I error of LD indices with ULD

<table>
<thead>
<tr>
<th>Largest F values</th>
<th>$G^2$</th>
<th>JSI</th>
<th>MIs</th>
<th>$Q_3$</th>
<th>r-to-z $Q_3$</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>degree</td>
<td>degree</td>
<td>degree</td>
<td>degree</td>
<td>degree</td>
<td>degree</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type I error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
</tr>
<tr>
<td>2\textsuperscript{nd}</td>
</tr>
<tr>
<td>3\textsuperscript{rd}</td>
</tr>
</tbody>
</table>

Note. * denotes interaction.
Table 22

*Average polychoric correlations for GRM cells with 1 LD item pair exhibiting the highest level of either type of LD*

<table>
<thead>
<tr>
<th>Average polychoric correlation (SD)</th>
<th>SLD $\pi_{LD} = 0.8$</th>
<th>ULD slope ratio = 1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 items, N = 250</td>
<td>0.912</td>
<td>0.754</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>20 items, N = 500</td>
<td>0.905</td>
<td>0.750</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>40 items, N = 1000</td>
<td>0.904</td>
<td>0.750</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.07)</td>
</tr>
</tbody>
</table>
Table 23

*Average mean and standard deviation (from 50 randomly selected replications) of summed JSI values in selected 2PLM cells*

<table>
<thead>
<tr>
<th>Cell</th>
<th>mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLD, $k = 10$, $N = 250$, $\pi_{LD} = 0.5$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 pair</td>
<td>-0.12</td>
<td>1.04</td>
</tr>
<tr>
<td>2 pairs</td>
<td>-0.15</td>
<td>1.29</td>
</tr>
<tr>
<td>4 pairs</td>
<td>-0.17</td>
<td>1.37</td>
</tr>
<tr>
<td>SLD, $k = 10$, $N = 500$, $\pi_{LD} = 0.3$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 pair</td>
<td>-0.02</td>
<td>0.68</td>
</tr>
<tr>
<td>2 pairs</td>
<td>-0.03</td>
<td>0.78</td>
</tr>
<tr>
<td>4 pairs</td>
<td>-0.05</td>
<td>1.00</td>
</tr>
<tr>
<td>SLD, $k = 10$, $N = 500$, $\pi_{LD} = 0.8$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 pair</td>
<td>-0.85</td>
<td>2.84</td>
</tr>
<tr>
<td>2 pairs</td>
<td>-4.30</td>
<td>15.79</td>
</tr>
<tr>
<td>4 pairs</td>
<td>-6.45</td>
<td>19.08</td>
</tr>
<tr>
<td>SLD, $k = 10$, $N = 1000$, $\pi_{LD} = 0.8$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 pair</td>
<td>-1.05</td>
<td>3.90</td>
</tr>
<tr>
<td>2 pairs</td>
<td>-5.41</td>
<td>18.31</td>
</tr>
<tr>
<td>4 pairs</td>
<td>-7.90</td>
<td>22.70</td>
</tr>
<tr>
<td>ULD, $k = 10$, $N = 250$, slope ratio = 1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 pair</td>
<td>-0.03</td>
<td>0.55</td>
</tr>
<tr>
<td>2 pairs</td>
<td>-0.03</td>
<td>0.62</td>
</tr>
<tr>
<td>4 pairs</td>
<td>-0.04</td>
<td>0.81</td>
</tr>
<tr>
<td>ULD, $k = 10$, $N = 500$, slope ratio = 0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 pair</td>
<td>-0.01</td>
<td>0.48</td>
</tr>
<tr>
<td>2 pairs</td>
<td>-0.02</td>
<td>0.50</td>
</tr>
<tr>
<td>4 pairs</td>
<td>-0.02</td>
<td>0.57</td>
</tr>
<tr>
<td>ULD, $k = 10$, $N = 500$, slope ratio = 1.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 pair</td>
<td>-0.03</td>
<td>0.69</td>
</tr>
<tr>
<td>2 pairs</td>
<td>-0.05</td>
<td>1.01</td>
</tr>
<tr>
<td>4 pairs</td>
<td>-0.22</td>
<td>2.14</td>
</tr>
<tr>
<td>ULD, $k = 10$, $N = 1000$, slope ratio = 1.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 pair</td>
<td>-0.02</td>
<td>0.79</td>
</tr>
<tr>
<td>2 pairs</td>
<td>-0.05</td>
<td>1.24</td>
</tr>
<tr>
<td>4 pairs</td>
<td>-0.15</td>
<td>2.29</td>
</tr>
</tbody>
</table>
Table 24

*2PLM cells with greater than 5% discard rates*

<table>
<thead>
<tr>
<th>Cell specifications</th>
<th>Percent of initial replications discarded</th>
</tr>
</thead>
<tbody>
<tr>
<td>2PLM_SLD_10_250_0.8_4</td>
<td>0.547</td>
</tr>
<tr>
<td>2PLM_SLD_10_250_0.8_2</td>
<td>0.531</td>
</tr>
<tr>
<td>2PLM_SLD_10_500_0.8_2</td>
<td>0.484</td>
</tr>
<tr>
<td>2PLM_SLD_10_250_0.8_1</td>
<td>0.477</td>
</tr>
<tr>
<td>2PLM_SLD_10_500_0.8_1</td>
<td>0.424</td>
</tr>
<tr>
<td>2PLM_SLD_10_1000_0.8_2</td>
<td>0.423</td>
</tr>
<tr>
<td>2PLM_SLD_10_500_0.8_4</td>
<td>0.422</td>
</tr>
<tr>
<td>2PLM_SLD_10_1000_0.8_1</td>
<td>0.340</td>
</tr>
<tr>
<td>2PLM_SLD_10_1000_0.8_4</td>
<td>0.321</td>
</tr>
<tr>
<td>2PLM_SLD_20_250_0.8_4</td>
<td>0.610</td>
</tr>
</tbody>
</table>
Table 25

*Overall power and Type I error for LD indices with 2PLM when problem cells are excluded.*

<table>
<thead>
<tr>
<th>Index</th>
<th>SLD Power</th>
<th>SLD Type I error</th>
<th>ULD Power</th>
<th>ULD Type I error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
</tr>
<tr>
<td>$G^*$</td>
<td>.965 (.063)</td>
<td>.035 (.035)</td>
<td>.752 (.301)</td>
<td>.025 (.025)</td>
</tr>
<tr>
<td>$\chi^2_{G/D}$</td>
<td>.382 (.432)</td>
<td>--</td>
<td>.257 (.378)</td>
<td>--</td>
</tr>
<tr>
<td>JSI</td>
<td>.928 (.108)</td>
<td>.011 (.009)</td>
<td>.666 (.264)</td>
<td>.018 (.009)</td>
</tr>
<tr>
<td>MIs</td>
<td>.984 (.036)</td>
<td>.155 (.147)</td>
<td>.779 (.279)</td>
<td>.097 (.116)</td>
</tr>
<tr>
<td>$Q_3$</td>
<td>.783 (.262)</td>
<td>.046 (.059)</td>
<td>.473 (.364)</td>
<td>.000 (.000)</td>
</tr>
<tr>
<td>r-to-z $Q_3$</td>
<td>.981 (.030)</td>
<td>.313 (.226)</td>
<td>.869 (.187)</td>
<td>.111 (.071)</td>
</tr>
<tr>
<td>residual z</td>
<td>.973 (.053)</td>
<td>.017 (.014)</td>
<td>.764 (.260)</td>
<td>.025 (.017)</td>
</tr>
</tbody>
</table>

Note: $\chi^2_{G/D}$ does not have Type I error as it only provides an overall test. SLD means are now based on 71 cells, while ULD values are based on 81.
Appendix B: Figures

Figure 1. CFA path diagram for MIs explication.
Figure 2. Alternate (and equivalent) CFA path diagram for MIs explication, highlighting IRT-based conceptualization.
Figure 3. Recovered discrimination parameter distributions for LD and Null items in the 2PLM SLD cell with 10 items, \( N = 250 \), and \( \pi_{LD} = 0.8 \). Note. Dark gray = Null, light gray = LD).
Figure 4. Average RMSE values for 2PLM discrimination parameters, varying across degree of SLD and separated by number of items.
Figure 5. Recovered discrimination parameter distributions for LD and Null items in the 2PLM ULD cell with 40 items, N = 250, and slope ratio = 1.5. Note. Dark gray = Null, light gray = LD.)
Figure 6. Average RMSE values for 2PLM discrimination parameters, varying across degree of LD and number of items, separated by type of LD.
Figure 7. Average power (upper portion) and Type I error rate (lower portion) for LD indices, in conjunction with the 2PLM and SLD, by number of items.
Figure 8. Average power and Type I error for LD indices, in conjunction with the 2PLM and SLD, by degree of LD.
Figure 9. Average power for LD indices, in conjunction with the 2PLM and SLD, by number of items and degree of LD.
Figure 10. Average power for LD indices, in conjunction with the 2PLM and ULD, by degree of LD.
Figure 11. Average Type I error rate for LD indices, in conjunction with the 2PLM and ULD, by degree of LD and number of items.
Figure 12. Average power for LD indices from overall analysis, in conjunction with the 2PLM, by type and degree of LD.
Figure 13. Recovered discrimination parameter distributions for LD and Null items in the GRM SLD cell with 10 items, \( N = 250 \), and \( \pi_{LD} = 0.8 \). Note. Dark gray = Null, light gray = LD).
Figure 14. Average RMSE values for GRM discrimination parameters, varying across degree of SLD and separated by number of items.
Figure 15. Recovered discrimination parameter distributions for LD and Null items in the GRM ULD cell with 40 items, N = 250, and slope ratio = 1.5. Note. Dark gray = Null, light gray = LD.)
Figure 16. Average RMSE values for selected GRM item parameters, varying across degree of LD and separated by type of LD.
Figure 17. Average RMSE values for all GRM item parameters, separated by LD type.
Figure 18. Average power for LD indices, in conjunction with the GRM, by sample size and degree of SLD.
Figure 19. Average power for LD indices, in conjunction with the GRM, by number of items and degree of SLD.
Figure 20. Average Type I error for LD indices, in conjunction with the GRM, by number of items and degree of SLD.
Figure 21. Power of LD indices, in conjunction with the GRM, by degree of ULD.
Figure 22. Type I error of LD indices, in conjunction with the GRM and ULD, by number of items and locally dependent pairs.
Figure 23: Power of the JSI with the GRM and 2PLM, separated by number of items and number of LD pairs.
Figure 24. Summarized discard rates for the 171 2PLM simulation cells.
Appendix C: R code for LD Indices Calculation

library(MCMCpack);

########################create holders for LD indices
len = (n*(n-1))/2 + n
G2.summ = rep(0,len); Q3.summ = rep(0,len); Q3z.summ = rep(0,len)
res.z.summ = rep(0,len); JSI.summ = rep(0,len); GD.test=0;GD.summ=0
MI.summ = rep(0,len)

########################generate quadrature nodes and weights
Xk = seq(-4.5,4.5,.1)
qp = 91; sax=0; ax = rep(0,91)
for (i in 1:qp){
   ax[i] = (1/(sqrt(2*pi))*exp(-.5*(Xk[i]^2)))
   sax = sax+ ax[i]
}
for (ii in 1:qp){
   ax[ii] = ax[ii] / sax
}

#####################start loop over replications
for (rep in 1:1000){
   print(rep)
   stringe = paste(cond,"_",rep,".dat", sep="")
   stringb = paste(cond,"_",rep,".PAR", sep="")
   stringc = paste(score,"_",rep,".sco", sep="")
   raw.data = matrix(scan(stringe),ncol=n,byrow=TRUE)
   raw.data = raw.data - 1
   parms = matrix(scan(stringb,nline=n),ncol=CAT,byrow=TRUE)
   aparm = parms[,1]
   if (model == "GRM"){bparm = matrix(parms[,2:(CAT)],nrow=n)
    }else {bparm = matrix(parms[,2])}
   theta = matrix(scan(stringc),ncol=2,byrow=TRUE)
   if(model == "2PL"){
      string.NH = paste("NH",rep.".out", sep="")
      out=readLines(paste(string.NH))
   }
start = grep("RESIDUAL MATRIX", out, value=F)+4
# How many lines we need
k=ceiling(n/9)-1; k=0:k; x=sum((n-9*k)-1)+max(k)*1
res.NH = read.table(paste(string.NH), skip=start, fill=T, nrow=x,
col.names=1:10, as.is=T)
length=(n-1)-(k*9)
endrow=cumsum(length); startrow=endrow-length+1
subset = as.numeric(rownames(res.NH))%in%subset==FALSE
res.NH = res.NH[subset,2:10]
if (n == 10){
  new = matrix(0,9,9)
  for (ww in 1:9){
    for (www in 1:9){
      new[ww,www] = as.numeric(res.NH[ww,ww])
    }
  }
  pad = rep(0,9)
  res.NH = rbind(pad, new)
}
if (n > 10) {
  startcol=1+(k*9)
  new.matrix=matrix(data=NA, nrow=n, ncol=n)
  for(i in 1:(max(k)+1)){
    temp=res.NH[startrow[i]:endrow[i],1:9]
    temp=data.frame(temp[,colSums(!is.na(temp))!=0])
    new.matrix[(nrow(new.matrix)-nrow(temp)+1):nrow(new.matrix),startcol[i]:startcol[i]+length(temp)-1]=as.matrix(temp)
  }
  res.NH = new.matrix
}

######################################################################## read in Mplus mod indices and put into matrix
stringh = paste(paste("Mplus","cond","rep", sep="._",".out", sep=""))
source(paste(folder,"Mplus mod indices.R", sep=""))
######################################################################## read in Edwards & Cai
stringJSI = paste(paste("OLOG","cond",sep="._"),".dbg_",".rep",".txt", sep=""
out=readLines(paste(stringJSI))
skip= grep("Edwards-Cai",out,value=F)
JSI = matrix(scan(paste(stringJSI), skip=skip, nlines=n),n,n,byrow=T)
######################################################################## create holder for LD stats.
G2.table = matrix(0,n,n); Q3.table = matrix(0,n,n); Q3z.table = matrix(0,n,n)
res.z.table = matrix(0,n,n); JSI1 = matrix(0,n,n)

begin item pair looping
for (y in 1:(n-1)){
    # loops over first item in item pairs
    for (yy in 2:n){
        # loops over 2nd item in pair
        if (y >= yy) next
        get observed value cont. table
        obs = table(raw.data[,y],raw.data[,yy])
        if (max(raw.data[,yy]) == 3 | max(raw.data[,y] == 3)) obs =
            table(addNA(raw.data[,y]),addNA(raw.data[,yy]))

        construct expected table
        ps = rep(0,1,CAT-1)
        qs = matrix(0,1,CAT-1)
        p = rep(0,1,CAT)
        q = rep(0,1,CAT)

        exp.prob= matrix(0,CAT,CAT)
        for (l in 1:qp){
            for (k in 1:CAT-1){
                ps[k] = 1/(1+ exp((-aparm[y]*(Xk[l] - bparm[y,k]))))
                qs[k] = 1/(1+ exp((-aparm[yy]*(Xk[l] - bparm[yy,k]))))
            }
            q[1] = 1 - qs[1]

            for (kk in 2:4){
                p[kk] = ps[kk-1]-ps[kk]
                q[kk] = qs[kk-1]-qs[kk]
            }

            p[CAT] = ps[CAT-1]
            q[CAT] = qs[CAT-1]

            for (k in 1:CAT){
                for (m in 1:CAT){
                    exp.prob[k,m] = exp.prob[k,m]+p[k]*q[m]*ax[l]
                }
            }
        }
    }
}

ends loop over quad. pts
exp.n = exp.prob*N

calculate G2 for item pair y, yy
G2 = 0
for (kk in 1:CAT){
    for (mm in 1:CAT){
        if (obs[kk,mm] == 0) {G2 = G2 + 0
            } else {G2 = G2 + (obs[kk,mm] * log((exp.n[kk,mm] /
                obs[kk,mm])))
        }
    }
}
G2 = -2*G2
G2.table[y,yy] = G2
if (G2.table[y,yy] > qchisq(.95,(CAT-1)^2)) {G2.table[yy,y] = 1
} else { G2.table[yy,y] = 0
}

calculate Q3 and Q3z for item pair y,yy

ei1 = rep(0,N); ei2=ei1
for (k in 1:(CAT-1)){
ei1 = ei1 + 1/(1+ exp((-aparm[y]*(theta - bparm[y,k]))))
ei2 = ei2 + 1/(1+ exp((-aparm[yy]*(theta - bparm[yy,k]))))
}
d1 = raw.data[,y] - ei1; d2 = raw.data[,yy] - ei2
Q3 = cor(d1,d2)
z = .5*(log((1+Q3)/(1-Q3)))
eq3 = -1 / (n-1)
ez = .5*(log((1+eq3)/(1-eq3))); Q3z = (z - ez) / (sqrt(1/(N-3)))
Q3.table[y,yy] = Q3
Q3z.table[y,yy] = Q3z
if (Q3.table[y,yy] > .2) {Q3.table[yy,y] = 1
} else { Q3.table[yy,y] = 0
}
if (Q3z.table[y,yy] > qnorm(.95)) {Q3z.table[yy,y] = 1
} else { Q3z.table[yy,y] = 0
}

calc res z for GD chi square

if(model == "2PL"){
pj = sum(obs[2,]) / sum(obs)
pk = sum(obs[,2]) / sum(obs)
rjk = res.NH[yy,y] / sqrt((pj*(1-pj)*pk*(1-pk)))
res.z = .5*(log((1+rjk)/(1-rjk)))
res.z.table[y,yy] = res.z
}

test mod indices

if (mod.indices[y,yy] > qchisq(.95,1)) {mod.indices[yy,y] = 1
} else { mod.indices[yy,y] = 0
}

sum JSI
JSI1[y,yy] = JSI[yy,y] + JSI[y,yy]

write obs, exp, and LD indices to output files
out.final = paste(cond,rep,"LD.out", sep=" ")
if (y ==1 & yy==2){write(paste("Item",y, "Item", yy,"observed", sep=" "),
file=out.final,append=F)
}else {write(paste("Item",y, "Item", yy,"observed", sep=" "),
file=out.final,append=T)
}
write(obs,ncol=CAT+1, file=out.final, append=T)
write(paste("Item",y, "Item", yy,"predicted", sep=" "),file=out.final,append=T)
write(exp.n,ncol=CAT, file=out.final,append=T)
write("G2",file=out.final,append=T); write(G2, file =out.final, append=T)
write("Q3",file=out.final,append=T); write(Q3, file =out.final, append=T)
write("Q3z",file=out.final,append=T); write(Q3z, file =out.final, append=T)
write("MI",file=out.final,append=T); write(mod.indices[yy,y], file=out.final, append=T)
write("JSI1",file=out.final,append=T); write(JSI1[y, yy], file=out.final, append=T)
write(" ", file=out.final, append=T)
if(model == "2PL")
{
write("res.z",file=out.final,append=T); write(res.z,
file=out.final, append=T)
write("MI",file=out.final,append=T); write(mod.indices[yy,y], file=out.final, append=T)
write("JSI1",file=out.final,append=T); write(JSI1[y, yy], file=out.final, append=T)
write(" ", file=out.final, append=T)
}
#ends loop over 2nd item item in pairs
} # ends loop over 1st item in item pairs
######################################Calc and output overall GD value
to file
if(model == "2PL")
{GDchi = (N - 3)* sum(res.z.table[row(res.z.table)<col(res.z.table)])^2)
df = (.5*n^2(n-1))- n
p = 1-pchisq(GDchi,df)
write("GDchi",file=out.final,append=T); write(GDchi, file=out.final, append=T)
write("GDdf",file=out.final,append=T); write(df, file=out.final, append=T)
write("GD p-value",file=out.final,append=T); write(p, file=out.final, append=T)
if (GDchi > qchisq(.95,df)){GD.test = 1
}else {GD.test = 0
}
}

##########################################test JSI and stand res.z
mn=mean(JSI1[row(JSI1)<col(JSI1)])
std=sd(JSI1[row(JSI1)<col(JSI1)])
zmean = mean(res.z.table[row(res.z.table)<col(res.z.table)])
zstd = sd(res.z.table[row(res.z.table)<col(res.z.table)])
zstand = matrix(0,n,n)
for (i in 1:n){
 for (j in 1:n){
  if (j >= i) next
  } #ends loop over second
} #ends loop over first
for (i in 1:n){
  if (j >= i) next
  
  } #ends loop over second
} #ends loop over first

123
if (JSI1[j,i] > (mn + 1.96*std)) { JSI1[i,j] = 1
} else { JSI1[i,j] = 0
}

zstand[row(zstand)<col(zstand)] = (res.z.table[row(res.z.table)<col(res.z.table)] -
zmean) / zstd
if (abs(zstand[j,i]) > 2) { zstand[i,j] = 1
} else { zstand[i,j] = 0
}

# collapses 0/1 LD results (including diagonal) into vector
# xpnd(summ) will put them back into a sym. matrix
G2.summ = G2.summ + (vech(G2.table));
Q3.summ = Q3.summ + (vech(Q3.table))
Q3z.summ = Q3z.summ + (vech(Q3z.table))
res.z.summ = res.z.summ + (vech(res.z.table))
GD.summ = GD.summ + GD.test; JSI.summ = JSI.summ + (vech(JSI1))
MI.summ = MI.summ + (vech(mod.indices))

write(G2.summ, ncol=1, file="G2.out", append=F)
write(Q3.summ, ncol=1, file="Q3.out", append=F)
write(Q3z.summ, ncol=1, file="Q3z.out", append=F)
write(res.z.summ, ncol=1, file="res.z.out", append=F)
write(GD.summ, ncol=1, file="GD.summ.out", append=F)
write(res.z.summ, ncol=1, file="res.z.out", append=F)
write(JSI.summ, ncol=1, file="JSI-std.out", append=F)
Appendix D: 2PLM Supplement to Address Discarded Replications Issue

Figure 24 presents the initial replication discard rates for all 2PLM cells. As can be seen, the majority of the design cells had negligible discard rates, with 30% (n = 50) having no replications discarded and 51% (n = 87) having fewer than 10 replications, out of the initial 1000, discarded. Cells in which the initial replication discard rate was 5% or above are listed in Table 24, along with the corresponding discard rate; out of the 171 2PLM cells, 10 are included in Table 24. An examination of the entries of the table finds that all discarded cells were generated with the highest level of SLD used, $\pi_{LD} = 0.8$. All but one of the cells had a scale length of 10 items, with the exception having a 20-item scale. All values of sample sizes and number of LD pairs included in the study are represented in Table 24. In summary, and as reported previously, it appears that the combination of a short scale with a high level of induced SLD is the main culprit in generating replications that fail to converge.

Analyses of parameter recovery and LD indices performance were re-run on a modified dataset, in which the cells listed in Table 24 were excluded. This resulted in some differences from the full analyses, particularly given that all possible combinations of 10 items and SLD $\pi_{LD} = 0.8$ were removed, but the overall conclusions were generally unchanged. For example, SLD parameter recovery section in the main results which reads as, “The most noteworthy predictors for the discrimination RMSE values were
degree of LD, $F(2, 16) = 4656.17, p < .001$, number of items, $F(2, 16) = 4313.49, p < .001$, and the interaction of degree of LD with number of items, $F(4, 16) = 3117.74, p < .001$."

"The most noteworthy predictors for the discrimination RMSE values were number of items, $F(2, 11) = 2098.79, p < .001$, degree of LD, $F(2, 11) = 1984.49, p < .001$, and the interaction of degree of LD with number of items, $F(3, 11) = 936.57, p < .001$." The specific values of the F-tests and the “importance” ordering of the three main effects has been modified, but the same effects are still reported.

With respect to the power and Type I error of the LD indices, some variation was also present between the full analyses and those with problem cells deleted. Table 25, which is a modification of Table 11 from the results using all cells, presents the power and Type I error rates for the indices. As can been seen, increases in power are seen for all indices but the $\chi^2_{G/D}$, but the general conclusions/ordering of the indices in terms of suitability does not change. The noted increases in power are most likely due the removal of a substantial portion of the 10 item cells, in which indices generally exhibited decreased power, rather than directly related to the discarding of replications.