Comparing Surface and Underlying Local Dependence Levels via Polychoric Correlations

Applied Psychological Measurement
2015, Vol. 39(4) 293–302
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621614561492
apm.sagepub.com

Carrie R. Houts¹ and Michael C. Edwards²

Abstract

Item response theory (IRT) is a set of psychometric models used in the social and behavioral sciences. As part of applying these models in practice, a number of assumptions are made. A large literature exists assessing the extent to which these assumptions are satisfied in a given data set. One of these assumptions, local independence, is the focus of the research described here. When the local independence assumption is violated, there is said to be local dependence (LD). Several different models of LD have been proposed, and a number of studies have been conducted examining the performance of different methods at detecting LD. Underlying LD (ULD) and surface LD (SLD) were proposed as two possible mechanisms underlying observed LD in an early exploration of detection procedures. In a number of previous studies, it appears as though ULD is more difficult to detect than SLD. In this article, the authors demonstrate a procedure to examine comparability of induced LD and present results, which suggest a re-interpretation of existing studies on LD detection.

Keywords

IRT, local dependence, surface LD, underlying LD, LD detection

In detailing the mathematical basis of models that are now known as item response theory (IRT), Lord and Novick (1968) discussed two critical assumptions of all IRT models: The chosen mathematical function describes the data well, and the item responses exhibit local independence. Local independence requires that all item responses (e.g., pairs of item, triplets of items) be unrelated to one another (i.e., independent) after conditioning on the latent variable(s). This representation of local independence is often called the strong form of local independence (e.g., McDonald, 1981), because 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 a less stringent definition of local independence, called weak local independence, be assumed. Weak local independence assumes, that for all $i \neq j$,

¹Vector Psychometric Group, LLC, Chapel Hill, NC, USA

²The Ohio State University, Columbus, OH, USA

Corresponding Author:

Carrie R. Houts, Vector Psychometric Group, LLC, 847 Emily Lane, Chapel Hill, NC 27516-0807, USA.

Email: crhouts@vpgcentral.com

$$P(X_i = x_i, X_j = x_j | \theta) = P(X_i = x_i | \theta) \times P(X_j = x_j | \theta). \quad (1)$$

Rather than assuming that every possible higher order set of items is locally independent, only pairs of items (of which there are k -choose-two pairs for a k -item scale) need be locally independent, which makes checking for local dependence (LD; when an item pair fails to be conditionally independent) among items computationally less expensive and the assumption of local independence more likely to hold in practice. Because of the more plausible nature of the assumption and the associated computational savings, weak local independence is the most commonly adopted assumption.

Chen and Thissen (1997) differentiated between LD that is due to highly similar content or placement, what they called surface LD (SLD), and LD that is a result of additional unmodeled latent variables, which they termed underlying LD (ULD). As described in Chen and Thissen, 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, imagine that a childhood autism scale (parent report) 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, parents may not understand these distinctions and might respond to the two items as if they were identical. The degree of similarity in SLD is denoted by π_{LD} , the probability that the second item in the locally dependent pair will have a response identical to the first item. More specifically, the SLD process for an item pair is,

With probability π_{LD} :

$$\text{Response to Item 2} = \begin{cases} 1, & \text{if } X_1 = 1 \\ 0, & \text{if } X_1 = 0 \end{cases}. \quad (2)$$

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 Equation 2, when π_{LD} is "active," the underlying process implied by the IRT model does not determine the response to the second item. For example, if $\pi_{LD} = 0.4$, of 1,000 parents who responded "agree" to the distractibility item of the autism instrument, 400 of them interpreted "difficulty paying attention" 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 answers to the last two items of a depression scale and simply put "strongly disagree" as their response for both items 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 underfactoring. 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 independence 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

Figure 1. Path diagram for the ULD generating model with one pair of locally dependent items.
 Note. ULD = underlying local dependence; LV = latent variable.

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. A diagram of such a model is presented in Figure 1. An example of ULD is a depression scale that includes an item about changes in a respondent's appetite and another item regarding changes in her sleep habits. These two items would not be interpreted as soliciting duplicate information, but given the underlying somatic component, they are likely more similar to each other than either somatic item would be to an additional item regarding an emotional symptom of depression, such as frequent crying.

Because of the negative effect LD can have on parameter estimates (e.g., Reese, 1995; Zenisky, Hambleton, & Sireci, 2002), numerous indices have been proposed to detect the presence of LD. Regardless of the mechanism generating LD, whether from one of the models described above or some other process, the property of excess covariation among the items of a locally dependent pair is a shared feature, and LD indices are typically "marketed" as general purpose, meaning they are not accompanied by suggestions for use with only certain type(s) of suspected LD. By omission, that can be taken to imply that LD measures should be equally good at detecting LD, regardless of the process from which it stems. However, several simulation studies conducted to assess the performance of LD indices have concluded that all of the studied LD indices tend to be better at detecting SLD compared with ULD (e.g., Chen & Thissen, 1997; Houts & Edwards, 2013). Despite this consistent finding, there has been no research as to *why* SLD is more readily detected than ULD.

In the remainder of this article, the authors attempt to accomplish two goals. First, they hope to demonstrate a simple and widely available approach by which researchers studying LD in the future can assess the degree of LD they are inducing using various mechanisms (ULD and SLD are used only as convenient examples). Second, using this approach, the authors intend to demonstrate that previous findings regarding the relative performance of LD indexes at identifying ULD and SLD may require re-interpretation.

The authors hypothesized that the SLD values used in previous studies were inducing a much stronger LD effect than the ULD values. However, because SLD and ULD rely on different mechanisms to create LD, directly comparing the parameters used to control the magnitude of LD induced in each method is not useful. Given that both LD processes induce excess covariation

among the items of a locally dependent pair, polychoric correlations have been employed to investigate the degree of “over-relatedness” in items simulated to exhibit LD. Because polychoric correlations are obtained from raw data, and do not consider the model used to generate the data, they are available when assessing both SLD and ULD and provide a common metric on which to compare the two types of LD. Although more complex, modeling-intensive methods for obtaining a common metric may exist, polychoric correlations were chosen because they are commonly used to assess strength of association between items and widely available in existing software.

Polychoric correlations are a common choice to measure the strength of association between two ordinal variables. By positing the existence of an underlying continuum that gives rise to each observed response, it is possible to use the observed contingency table (along with some assumptions) to estimate the correlations among the underlying variable-level continua. Even if one is not particularly interested in the underlying continua, the mechanics of the process serve to disattenuate estimates as compared with what would be seen using a standard Pearson product-moment correlation. For a more detailed overview of polychoric correlations, see Wirth and Edwards (2007).

Relevant IRT Models

Although there are a variety of IRT models that may be used to fit data, the two-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-type responses, typically with three to seven response options from which a respondent may choose.

Two-Parameter Logistic Model

The 2PLM 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,

$$P(x_j = 1 | \theta) = \frac{1}{1 + \exp[-a_j(\theta - b_j)]}, \quad (3)$$

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 θ is the latent construct (typically assumed to follow a standard Normal distribution).

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 m categories, the probability of responding in the p th category is,

$$P(x_j = p | \theta) = \frac{1}{1 + \exp[-a_j(\theta - b_{jp})]} - \frac{1}{1 + \exp[-a_j(\theta - b_{j(p+1)})]} = P^*(p) - P^*(p+1), \quad (4)$$

where b_{jp} , for item j , is the threshold for category p , which marks the latent construct level where the probability of scoring in p th category or higher reaches 50%, and all other parameters

are as defined before. To facilitate computation, it is specified that $P^*(0) = 1$ and $P^*(m + 1) = 0$. To find the probability of being in the p 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 $m - 1$ threshold parameters.

Method

Latent Construct and Item Parameters

Individual θ values were generated from a standard Normal distribution. For conditions including ULD, theta values were generated from a multivariate Normal distribution with means of zero, unit variances, and covariances of zero between dimensions. Item discrimination parameters for the primary factor 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 in peer-reviewed publications. For items generated with two response categories, the b -parameter was generated from a standard Normal distribution, similar to the 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 four needed b -parameter values. Specifically, the first b -parameter was drawn from a Normal distribution with a mean value of -1.5 and a standard deviation of 0.5 and to obtain values for the remaining b -parameters, "shift" values, restricted to positive numbers, were drawn from a $N(1.0, 0.2)$ distribution and added to the previous b -parameter value. For example, if the first b -parameter was drawn as -1.4 and the first shift value was 0.8 , then the second b -parameter would be $(-1.4 + 0.8 = -0.6)$. A second random-draw shift value of 0.4 would result in a third b -parameter of $(-0.6 + 0.4 = -0.2)$. This procedure insures that the b -parameters will maintain the desired ordering. Although not drawn from these distributions directly (with the exception of the first value), the four b -parameters clustered around values of -1.5 , -0.5 , 0.5 , and 1.5 , respectively.

LD Type and Degree of LD

Surface LD. SLD was detailed earlier as a deterministic process in which π_{LD} , the probability that the second item in the locally dependent pair will have a response identical to the first item, denotes the extent of locally dependent present within an item pair. Level of SLD was manipulated through variations of π_{LD} .

Underlying LD. For the purpose of simulating item pairs with ULD, a compensatory multidimensional IRT model was employed. Compensatory models are those in which the I latent traits being estimated are specified to have an additive relationship (Bock, Gibbons, & Muraki, 1988). The multidimensional model was specified so all items loaded onto a general factor and the locally dependent item pair contributed an additional factor, onto which only the two items exhibiting ULD loaded. Specifically, the generating model in the 2PLM case, which is actually a multidimensional 2PLM, was,

$$P(x_j = 1 | \theta) = \frac{1}{1 + \exp[-(a_{j1}\theta_1 + a_{j2}\theta_2 + d_j)]}, \quad (5)$$

where a_{j1} is the slope for item j on the first factor, which in this case would be the primary, or general, factor (θ_1), a_{j2} is the slope for item j on the LD, or specific, factor (θ_2), and d_j is the intercept for item j . Only the items intended to exhibit LD measured the second factor. Because

of identification issues with latent factors that have only two observed variables, the specific factor slopes for locally dependent item pairs were constrained to be equal and the specific latent factors are generated to be uncorrelated with the general factor. This type of latent structure, similar to the testlet response model of Wainer and colleagues (e.g., Wainer, Bradlow, & Wang, 2007), resembles the bi-factor model. The bi-factor model, originally proposed by Holzinger and Swineford (1937), has received a resurgence of interest lately (e.g., Cai, Yang, & Hansen, 2011; Reise, 2012) as potentially useful for modeling psychology constructs. It is also of note that this kind of multidimensional structure is analogous to correlated residuals in a structural equation modeling framework. While other multidimensional structures are possible (e.g., correlated factors), bi-factor-type models are a very reasonable (and recently popular) place to examine this issue. Model mis-specifications could have a major impact on the performance of LD measures, but it seems plausible that, conditional on an otherwise correctly specified model, the ability to detect LD may not depend heavily on the particular structure of that otherwise correctly specified model.

For item pairs with ULD, the degree of LD was defined by the magnitude of the slope of the specific factor in relation to the slope of the general factor. As an example, for a ULD generating value of 0.5, the general slopes were sampled from a $N(1.7, 0.3)$ distribution, as noted earlier, and specific slopes were sampled from a $N(0.85, 0.15)$ 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; a standard deviation of 0.15 was used for all specific slope generating distributions, regardless of the specified mean value.

Degree of LD generated. Rather than choosing a set number of values to generate ULD and SLD, the authors instead set out to create something akin to a concordance table, linking various ULD and SLD generating values to one another via the polychoric correlation they induce. They manipulated both ULD and SLD generating values as necessary to fill in this table. The values used for SLD ranged from 0.1 to 0.8 and for ULD from 0.5 to 5. The range of these values, both for the SLD and ULD conditions, was much wider than the range of values used in previous simulations studies. For example, Chen and Thissen (1997) and Houts and Edwards (2013) both used SLD values of 0.3, 0.5, and 0.8 in their study of SLD. Houts and Edwards (2013), the only study the authors know of that has induced ULD in a similar manner, used values of 0.5, 1.0, and 1.5.

Data Generation, Parameter Estimation, and LD Index Calculations

Using the item and latent construct parameters estimates detailed above, programs were written in R (version 2.10, R Development Core Team, 2009) to simulate data. Data were generated for a scale length of 20 items and for a sample size of 500.¹ Each data set had only one pair of locally dependent items. For each permutation of model and ULD/SLD value, 1,000 replications were conducted. Generated data were submitted to Mplus 6.1 (Muthén & Muthén, 1998-2010), with all variables specified as categorical, and a lower diagonal matrix of bivariate polychoric correlations was obtained.

After finding comparable LD severity values using the polychoric correlations, LD indices were obtained for a small selection of the simulation cells, to allow for comparison of the detection and Type I error rates. To obtain necessary values for the LD indices, IRT parameter estimation was conducted in MULTILOG (Thissen, 1991). MULTILOG employs full information maximum likelihood estimation, implemented as the Bock–Aitkin expectation-maximization (EM) algorithm. To obtain more precise estimates from MULTILOG, the default number of quadrature points was increased to 91, and the maximum number of EM cycles was increased

Table 1. Severity of LD Conversion Chart Based on Observed Polychoric Correlations.

2PLM			GRM		
Polychoric correlation	ULD	SLD	Polychoric correlation	ULD	SLD
.55	0.50		.55	0.50	0.10
.59		0.10			
.60	0.75		.60	0.75	0.20
.65	1.00		.65	1.00	0.30
.67		0.20			
.70	1.25		.71	1.25	
.75	1.50	0.30	.75	1.50	0.50
.80		0.40	.81		0.60
.82	2.00		.82	2.00	
.86	2.50	0.50	.86	2.50	0.70
.90	3.00		.90	3.00	
.91		0.60	.91		0.80
.92	3.50		.92	3.50	
.94	4.00	0.70	.94	4.00	
.97	5.00	0.80	.96	5.00	

Note. $N = 500$, $k = 20$ items for all cells. ULD refers to the ratio of specific to general slope values. SLD refers to the probability that the second item in the locally dependent pair will have a response identical to the first item (π_{LD}). LD = local dependence; 2PLM = two-parameter logistic model; GRM = graded response model; ULD = underlying local dependence; SLD = surface local dependence.

from 25 (the default) to 200. No prior distributions were specified for the item parameters estimates. MULTILOG item parameter and, when needed, expected a posteriori latent construct person estimates were used in calculating Q_3 (Yen, 1984, 1993), Fisher’s r -to- z transformed Q_3 (Yen, 1984, 1993), and G^2 (Chen & Thissen, 1997). An R program was written for the calculation of G^2 , Q_3 , and Fisher’s r -to- z transformed Q_3 . Jackknife Slope Index (JSI; for example, Houts & Edwards, 2013) values were obtained from flexMIRT (Cai, 2012). Modification indices (MIs) were calculated by MPLUS, which used diagonally weighted least squares.

Results

Although the magnitude of the difference depended on the severity of the LD induced, the polychoric correlation between items in the locally dependent pair was generally higher than the correlations among items not involved in locally dependent pairs. For instance, with $\pi_{LD} = 0.5$ (SLD) and a 2PLM model, the average polychoric correlation between the two items exhibiting LD, across the 1,000 replications, was 0.91 whereas the average correlation of the non-locally dependent entries in the lower triangle was 0.48. In the ULD GRM cell with a specific-to-general slope ratio of 1.5, the locally dependent items displayed an average correlation of 0.75 whereas all other correlations averaged to 0.46. These provide validation that data generation proceeded properly (i.e., LD was, in fact, induced) and supports the idea that polychoric correlations are sensitive to LD, regardless of the type of LD induced.

Table 1 presents a rough conversion table for both the 2PLM and GRM conditions, detailing the SLD and ULD values needed to obtain similar observed polychoric correlations. It can be seen that extremely high slope ratios for the ULD conditions are needed to obtain polychoric correlation values similar to those obtained from mid-range and high π_{LD} values from the SLD model.

Table 2. Comparison of Type I Error and Power Rates When SLD and ULD Induce Comparable Polychoric Correlation Values.

LD index	2PLM, power (Type I error)		GRM, power (Type I error)	
	SLD	ULD	SLD	ULD
G^2	.979 (.017)	.997 (.017)	1.000 (.065)	.971 (.063)
JSI	.981 (.019)	.950 (.022)	.997 (.015)	.992 (.018)
MI	.993 (.053)	.999 (.069)	.999 (.040)	1.000 (.033)
Q_3	.583 (.000)	.912 (.000)	.671 (.000)	.944 (.000)
Q_3 r-to-z transformed	.992 (.078)	1.000 (.076)	.999 (.076)	1.000 (.073)

Note. $N = 500$, $k = 20$ items. $\pi_{LD} = 0.3$ for both 2PLM and GRM. ULD slope ratio = 1.5 for 2PLM and 1.0 for GRM. Average polychoric correlation of locally dependent items for 2PLM = 0.75, for GRM = 0.65 (see Table 1). SLD = surface local dependence; ULD = underlying local dependence; LD = local dependence; 2PLM = two-parameter logistic model; GRM = graded response model; JSI = Jackknife Slope Index; MI = modification indices.

Table 2 presents the power and Type I error rates for a selection of LD indices. For the values reported, the SLD and ULD cells were matched on the observed polychoric correlation between the items in the locally dependent pair, rather than matching on a rank-ordering of the LD simulation values, which is how SLD/ULD comparisons have been previously done. It appears that differences in correct detection rates or Type I error across SLD and ULD are generally small. The largest differences seen are associated with the power and Type I error rates for Q_3 .

Discussion

Overall, the authors have demonstrated that typical SLD and ULD generating values from simulation studies are inducing noncomparable levels of LD, as assessed by the common metric of polychoric correlations. Although SLD values can, in theory, range from 0 to 1, values at the higher end of the range (e.g., $\pi_{LD} = 0.7$ -0.8), and even mid-range values (e.g., $\pi_{LD} = 0.4$ -0.6), particularly when paired with a 2PLM, can result in questionably strong LD among the items of the locally dependent pair. If one attempts to mirror these levels of LD when using the ULD model to generate data, large ULD slope ratios must be used (e.g., 5.0), which result in unreasonable slope parameter values (e.g., specific slopes > 8.0).

In addition, when comparisons are made for LD index performance across SLD and ULD after matching on the observed polychoric correlation between a pair of items known to be locally dependent, the power and Type I error rates are very similar across the two methods of LD induction. The most noticeable differences in detection rate across ULD and SLD appear in the performance of the Q_3 , in both the 2PLM and GRM cells, in which ULD was more accurately detected than SLD. The generally poor performance of the untransformed Q_3 in the SLD conditions mirrors previous findings (e.g., Houts & Edwards, 2013) and provides empirical support for Chen and Thissen’s (1997) contention that the static cut-value of 0.2 is likely sub-optimal. The differential performance of the Q_3 also confirms Chen and Thissen’s detailed exploration into the functioning of the Q_3 and how its performance is expected to vary by type of LD, with an advantage seen in the detection of ULD by the Q_3 . Taken as a whole, the similar LD detection rates across type of LD suggests that previous research that found SLD is more easily detected than ULD might be more accurately framed as the idea that more LD is easier to find than less LD.

As with all studies, simulation or otherwise, this project has limitations. Differing item parameter distributions were not examined and it cannot be said with certainty what impact they

would have on the results. However, the simulation process used was designed to mimic the kinds of data that are reported in published articles in psychology journals. The authors also chose to focus on a simple case where there is only one locally dependent pair to be found. There are likely to be many situations where there are multiple pairs and/or more than two items exhibiting LD. Although these conditions were not explicitly studied, the nature of LD and previous research suggest that the conclusions reached here are unlikely to change as a result of adding additional LD item pairs (or items to a group). The authors chose to focus on ULD and SLD as generating models of LD due to their popularity in the extant literature, but there are other possible models that could be used to generate LD.

This study suggests a method to gauge the relative comparability of the LD generated by different LD-generating models, and the results suggest that there is, in fact, a need for such a metric. Rather than attempting to recreate the vast array of LD simulation conditions one can find in the literature, the authors choose to focus on describing an approach to generate more similar amounts of LD using different mechanisms. They feel using information like that contained in Table 1 will allow individual researchers to decide on appropriate steps to take given the specific aspects of their own simulations.

In future simulation studies that examine SLD and LD detection index performance, the authors suggest that, barring evidence of larger values existing in nature, the most appropriate course of action is to restrict π_{LD} values to the lower end of the range of possible values. This should allow comparative studies of LD detection methods to accurately differentiate between sensitivity to amount of LD versus sensitivity to mode of LD generation. In addition, focusing research in the restricted range of simulation values would result in LD indices that were optimized for detecting more realistic (but non-trivial) levels of induced LD, which is where the technology should presumably be aimed and is most useful for applied researchers.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The second author (Edwards) has a financial interest in the company that publishes flexMIRT, the software package used for calculating the Jackknife Slope Index values reported in the article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. As the polychoric correlations are estimated bivariate here, the number of items is not a meaningful variable to manipulate. The authors include items without local dependence (LD) in the simulations to provide a base of comparison to inform the understanding of the impact of the induced LD.

References

- Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.
- Cai, L. (2012). flexMIRT™ version 1.86: A numerical engine for multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group.
- Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*, 221-248.

- Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Education and Behavioral Statistics*, 22, 265-289.
- Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-based statistics for testing unidimensionality. *Applied Psychological Measurement*, 31, 292-307.
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57, 423-436.
- Hill, C. D. (2004). *Precision of parameter estimates for the graded item response model* (Unpublished master's thesis). The University of North Carolina at Chapel Hill.
- Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2, 41-54.
- Houts, C. R., & Edwards, M. C. (2013). The performance of LD measures with psychological data. *Applied Psychological Measurement*, 36, 541-562.
- Kim, D. (2007). *Assessing relative performance of local item dependence (LID) indexes* (Unpublished doctoral dissertation). University of Nebraska-Lincoln.
- Lord, F. M., & Novick, D. A. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 34, 100-117.
- McDonald, R. P. (1994). Testing for approximate dimensionality. In D. Levault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 63-86). Ottawa, Ontario, Canada: University of Ottawa.
- Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.
- R Development Core Team. (2009). R: A language and environment for statistical computing (Version 2.10) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.
- Reese, L. M. (1995). *The impact of local dependencies on some LSAT outcomes* (Statistical report 95-02). Newtown, PA: Law School Admission Council.
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47, 667-696.
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychological Monograph No. 17). Richmond, VA: Psychometric Society.
- Stout, W. F. (2002). Psychometrics: From practice to theory and back. *Psychometrika*, 67, 485-518.
- Thissen, D. (1991). *MULTILOG user's guide*. Chicago, IL: Scientific Software International.
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12, 58-79.
- Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.
- Yen, W. (1993). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213.
- Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, 39, 291-309.