Article

Educational and Psychological Measurement
2018, Vol. 78(4) 653–678

DOI: 10.1177/0013164417714506
journals.sagepub.com/home/epm
SAGE

On Lagrange Multiplier Tests in Multidimensional Item Response Theory: Information Matrices and Model Misspecification

Carl F. Falk[1] and Scott Monroe[2]

Abstract

Lagrange multiplier (LM) or score tests have seen renewed interest for the purpose of diagnosing misspecification in item response theory (IRT) models. LM tests can also be used to test whether parameters differ from a fixed value. We argue that the utility of LM tests depends on both the method used to compute the test and the degree of misspecification in the initially fitted model. We demonstrate both of these points in the context of a multidimensional IRT framework. Through an extensive Monte Carlo simulation study, we examine the performance of LM tests under varying degrees of model misspecification, model size, and different information matrix approximations. A generalized LM test designed specifically for use under misspecification, which has apparently not been previously studied in an IRT framework, performed the best in our simulations. Finally, we reemphasize caution in using LM tests for model specification searches.

Keywords

multidimensional item response theory, score test, Lagrange multiplier test, modification indices

Introduction

In any initial attempt at fitting a model to data, it is likely that some misspecification will occur. In the context of item response theory (IRT), overall tests of model fit

[1]Michigan State University, East Lansing, MI, USA
[2]University of Massachusetts, Amherst, MA, USA

Corresponding Author:
Carl F. Falk, Measurement and Quantitative Methods, Michigan State University, 458 Erickson Hall, East Lansing, MI 48824, USA.
Email: falkcarl@msu.edu

and targeted fit statistics may help diagnose the source of misfit (Maydeu-Olivares, 2013). For example, in the context of multidimensional IRT (MIRT), the researcher must specify a factor structure and pattern of loadings. If the overall model fit is poor, the researcher may wonder whether the fit may be improved by modifying the pattern of loadings. One way to address this question is through use of the *score test* (Rao, 1948), also called the *Lagrange multiplier* test (LM; Aitchison & Silvey, 1958). In short, LM tests allow a researcher to approximate the change in the log-likelihood if a constrained parameter were to be freely estimated, without actually refitting the model. Alternatively, LM tests may be used to form confidence intervals or perform hypothesis testing for to-be-freed parameters.

In an IRT context, LM tests have been introduced and studied for detecting differential item functioning (Fox & Glas, 2005; Glas, 1998), violations of the functional form of the response functions (Glas, 1999; Glas & Falcón, 2003; Ranger & Kuhn, 2012), and person fit (Glas & Dagohoy, 2007). LM tests have also been studied as a method to detect local dependence (Fox & Glas, 2005; Glas, 1999; Glas & Falcón, 2003; Kim, De Ayala, Ferdous, & Nering, 2011; Liu & Maydeu-Olivares, 2012; Liu & Thissen, 2012, 2014; Obserski, van Kollenburg, & Vermunt, 2013; van der Linden & Glas, 2010). Finally, LM tests have been studied in a general Bayesian framework (Glas, 1999; Fox & Glas, 2005).

This article makes several contributions, the first of which concerns study of the method used to compute LM tests. With full-information estimation (e.g., Bock & Aitkin, 1981; Schilling & Bock, 2005), computational time can sometimes restrict the number of alternative MIRT models that can be computed. As stated above, an appealing feature of LM tests is that they do not require refitting the model to calculate the statistic. However, LM tests do require an approximation of the *information matrix*: The asymptotic covariance matrix of the score vector, including the to-be-freed parameter(s). This matrix is commonly used to obtain standard errors for model parameters in unrestricted models, and various methods may be used to approximate it. These methods, though, may differ in terms of accuracy and computational difficulty (e.g., Paek & Cai, 2014; Tian, Cai, Thissen, & Xin, 2013), as well as appropriateness under misspecification. For example, an estimate could be based on *expected* or *observed* versions of the cross-product of gradients or negative Hessian (Liu & Thissen, 2014; Yuan, Cheng, & Patton, 2014).[1] Among these approaches, the observed cross-product approach is typically the fastest to compute, but may be the least accurate for producing standard errors (e.g., Paek & Cai, 2014). Expected versions require the expectation to be taken over all possible response patterns and are thus computationally intractable with a long test.

Comparisons among the matrices involved in LM test computation are rare. Several studies of LM tests have used only the observed cross-product approach (Glas & Falcón, 2003; Liu & Thissen, 2012), while Glas (1998) and Obserski et al. (2013) used only the observed Hessian approach. In a limited set of simulations, Liu and Maydeu-Olivares (2012) found that LM tests based on expected information yield more accurate results than those based on the observed Cross-product. Glas

(1999) used both the observed cross-product and observed Hessian approaches and found little difference in the performance of the corresponding LM tests. But, recent studies focusing on the information matrix have found that the observed cross-product is generally less accurate than the observed Hessian (Yuan et al., 2014). The computational approach could be important as under some conditions LM tests have been found to have inflated Type I error rates (Glas & Falcón, 2003; Liu & Maydeu-Olivares, 2012; Liu & Thissen, 2012, 2014; Ranger & Kuhn, 2012), and some studies show conflicting results as to whether LM tests perform well with a long test and few respondents (e.g., Glas & Falcón, 2003; Kim et al., 2011; Liu & Thissen, 2012, 2014; Ranger & Kuhn, 2012).

Another possibility is to compute a generalized LM test (Boos, 1992; Engle, 1984; White, 1982). The distinctive feature of the generalized LM test is that it is derived assuming there is some misspecification in the model (Boos, 1992). Since it is analogous to use of the sandwich covariance matrix used to obtain parameter standard errors (Yuan et al., 2014), we later use the term *sandwich* to refer to the generalized LM test. On the other hand, standard forms of the LM test assume the model is exactly correct and may lead to incorrect inference when this assumption is violated. Given that, in practice, models are routinely misspecified (MacCallum, 2003), the generalized LM test is arguably a preferable statistic for applied work. However, to the best of our knowledge, its performance has not yet been evaluated in an IRT context. Thus, this research investigates the performance of the generalized LM test via a simulation study, along with both the observed cross-product and observed Hessian approaches.

The second contribution of the current research concerns the utility of LM tests under varying degrees of misspecification. In any practical research setting, the initially fitted model will most likely be misspecified. The naive researcher may hope that LM tests can definitively distinguish between parameters that are fixed to their true values in the population data-generating model (assuming such a parametric model exists), and parameters that are fixed to incorrect values. In a structural equation modeling context, there are several prominent examples of LM tests being used in model specification searches that typically do not lead to the correctly specified model (Chou & Bentler, 1990; Kaplan, 1988; MacCallum, 1986). While this property of LM tests is inherent in the underlying statistical theory, some research gives the impression that LM tests rejecting parameters fixed to the data-generating values should always count as Type I errors, even under model misspecification (Green, Thompson, & Babyak, 1998). In reality, if the initial model is misspecified, freeing a parameter that is fixed to its data-generating value can often improve model fit and a properly formulated LM test will tend to detect this improvement. Even if one knows the data-generating model, determining whether the null hypothesis is true under an LM test is nontrivial when the baseline and alternative models are misspecified. When the null is true under misspecification, we refer to rejection of the null as a Type I error. When the null is not true under misspecification but the parameter is correctly fixed to its data-generating value, we refer to rejection of the null as a *false*

positive. This point will be further clarified when we provide a concrete example and discuss technical details of LM tests under misspecification. Under misspecification, however, the generalized LM test is theoretically the proper test.

Determining the performance of different LM computational approaches for practical research settings thus requires their study under varying levels of misspecification. Some previous research suggests that LM tests have false positives close to their nominal value or only slightly inflated (Kim et al., 2011; Liu & Maydeu-Olivares, 2012; Liu & Thissen, 2014), yet in some of these studies the overall magnitude of misspecification is arguably small (Liu & Maydeu-Olivares, 2012; Liu & Thissen, 2014). Other studies did not vary the degree of misspecification (Ranger & Kuhn, 2012) or have studied LM tests with few items (e.g., 10 or less; Glas, 1998, 1999). The current investigation is in line with Glas and Falcón (2003), in which the number of items exhibiting local dependence and the degree of local dependence was widely varied. Yet, in most conditions Glas and Falcón (2003) found that LM tests still tend to detect true misspecifications at a greater rate than generating false positives, but did not investigate different computational approaches for LM tests. To our knowledge, in none of these investigations was the misspecified model under the null hypothesis extensively discussed nor were generalized LM tests investigated. For example, Glas (1998) notes that misspecified models may lead to bias in parameter estimates, inflating LM test rejection rates for parameters correctly fixed to their data-generating values, but does not provide an illustrative example nor discuss generalized LM tests.

A final contribution of the current research concerns the use of a MIRT model for the simulation study. Some previous studies on local dependence have used specialized MIRT models (e.g., bifactor, testlet) for the less restrictive model (e.g., Liu & Maydeu-Olivares, 2012; Liu & Thissen, 2012, 2014) in the LM test. However, in these cases, a unidimensional model was used for the more restrictive, or baseline, model. In contrast, the current study uses a two-factor MIRT model for the baseline model. Consequently, the LM test can be used to test for omitted cross-loadings, which has not previously been explored in an IRT framework.

In summary, the main goal of this article is to present an extensive Monte Carlo simulation study that examines the performance of LM tests for omitted cross-loadings in MIRT models: (1) using different methods of score test computation, including the generalized LM test designed specifically for use under model misspecification and (2) under various degrees of model misspecification. In what follows, we first present a motivating example followed by general theory regarding LM tests. Next, we detail the Monte Carlo simulation study. Finally, concluding remarks and limitations are discussed.

Example and Theoretical Background

A Multidimensional Item Response Theory Model

Suppose we have $i = 1, \ldots, N$ independent subjects who complete $j = 1, \ldots, n$ items, with Y_{ij} representing a discrete random variable for subject i's response to the jth item

and y_{ij} its realization. For simplicity, we consider just dichotomous items and model item responses with the multidimensional two-parameter logistic model,

$$P(Y_{ij} = 1 | \mathbf{a}_j, c_j, \boldsymbol{\theta}_i) = \frac{1}{1 + \exp(-(c_j + \mathbf{a}_j' \boldsymbol{\theta}_i))}, \tag{1}$$

where c_j is the item intercept, $\mathbf{a}_j$ is a $D \times 1$ vector of item slopes, and $\boldsymbol{\theta}$ is a vector of the D latent traits. Under the current application, we consider $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = 0$ and the diagonals of $\boldsymbol{\Sigma}$ fixed to 1 for identification of the latent scale. The marginal probability of a response pattern, $\mathbf{y}_i$, is

$$\pi(\mathbf{y}_i | \boldsymbol{\eta}) = \int f(\mathbf{y}_i | \mathbf{c}, \mathbf{a}, \boldsymbol{\theta}) \boldsymbol{\phi}(\boldsymbol{\theta} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{\theta}, \tag{2}$$

where $f(\mathbf{y}_i | \mathbf{c}, \mathbf{a}, \boldsymbol{\theta})$ is the conditional probability of response pattern $\mathbf{y}_i$, $\boldsymbol{\phi}$ is the density function of the multivariate normal distribution, $\mathbf{c}$ and $\mathbf{a}$ collect all intercepts and item slopes, respectively, and $\boldsymbol{\eta}$ in turn collects *all* model parameters. Once item responses are observed, they are treated as fixed and the marginal log-likelihood is defined as the sum of log-likelihood contributions for each respondent:

$$l(\boldsymbol{\eta} | \mathbf{Y}) = \sum_{i=1}^{N} l(\boldsymbol{\eta} | \mathbf{y}_i) = \sum_{i=1}^{N} \log \pi(\mathbf{y}_i | \boldsymbol{\eta}). \tag{3}$$

As a concrete example, consider a correlated two-factor model with simple structure in which each item loads on only a single factor (Figure 1). This model is also known as a between-item multidimensional model. Similar models have been used to investigate the performance of LM tests (or modification indices) in structural equation modeling (Green et al., 1998). In our simulation study, such a model is specified as the baseline model and is fit to all replicated data sets. If the baseline model is the same as the true data-generating model, any significant LM tests constitute Type I errors.

If the baseline model exhibits poor fit, the researcher may explore revising the model. Though the model may be incorrect in numerous ways, we focus on testing for an omitted cross-loading. For example, suppose that the true data-generating model is represented in Figure 2, in which a single cross-loading is present for Item 1. Example item parameters for such a model appear on the top-left of Table 1. In this study, we focus on LM tests for a single cross-loading. In other words, the LM test tells us whether to expect a significant improvement in model fit under a less restrictive model in which a single cross-loading is freed. Continuing with the example, a significant LM test for the omitted cross-loading would (correctly) indicate how the researcher should revise the model.

The situation becomes more complex when both the baseline and the less restrictive models are misspecified. For example, in Figure 3, we consider the case where the baseline model is misspecified and the researcher tests for a cross-loading from

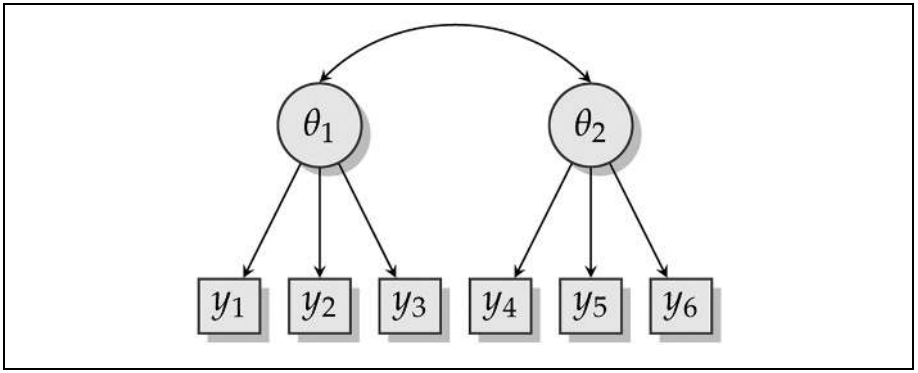

Figure 1. Two-factor model with simple structure.

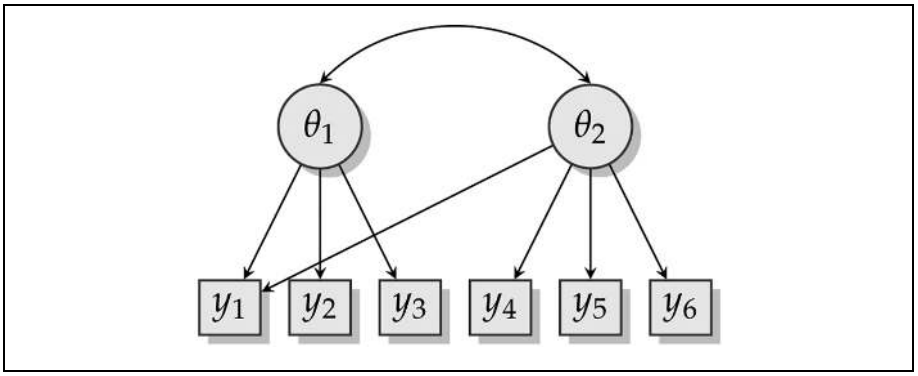

Figure 2. Two-factor model with cross-loading.

θ_2 to Item 3 instead of to Item 1. Under the true model (i.e., Figure 2), the cross-loading from θ_2 to Item 3 is zero. However, under the less restricted model (i.e., the model in Figure 3), this same cross-loading may be nonzero even when fit at the population level. Whether the less restrictive model is misspecified therefore has implications for the truth of the null hypothesis for an LM test. We will return to this topic after presenting LM tests for correctly specified models.

Lagrange Multiplier Tests for Correctly Specified Models

LM tests were proposed as Rao's score test (Rao, 1948) and are asymptotically equivalent to likelihood ratio tests and Wald tests (e.g., Engle, 1984). An equivalent development of the score test was presented in econometrics derived under

Table 1. Hypothetical True Item Parameters and Population Values Under Misspecification.

	True model, $\boldsymbol{\eta}$			Model 2, $\boldsymbol{\eta}^*$			Model 3, $\boldsymbol{\eta}^*$		
Item	c	$\mathbf{a}_1$	$\mathbf{a}_2$	c	$\mathbf{a}_1$	$\mathbf{a}_2$	c	$\mathbf{a}_1$	$\mathbf{a}_2$
1	0.00	1.00	0.70	0.00	1.59		0.00	1.59	
2	−1.00	1.20	0.00	−0.98	1.33	***−0.44***	−0.91	0.89	
3	1.00	1.20	0.00	0.91	0.89		0.98	1.33	***−0.44***
4	0.00	0.00	1.00	0.00		1.00	0.00		1.00
5	−1.00	0.00	1.20	−1.00		1.20	−1.00		1.20
6	1.00	0.00	1.20	1.00		1.20	1.00		1.20
	Model 4, $\boldsymbol{\eta}^*$			Model 5, $\boldsymbol{\eta}^*$			Model 6, $\boldsymbol{\eta}^*$		
Item	c	$\mathbf{a}_1$	$\mathbf{a}_2$	c	$\mathbf{a}_1$	$\mathbf{a}_2$	c	$\mathbf{a}_1$	$\mathbf{a}_2$
1	0.00	1.96		0.00	1.96		0.00	1.96	
2	−0.90	0.86		−0.90	0.86		−0.90	0.86	
3	0.90	0.86		0.90	0.86		0.90	0.86	
4	0.00	***0.00***	1.00	0.00		1.00	0.00		1.00
5	−1.00		1.20	−1.00	***0.00***	1.20	−1.00		1.20
6	1.00		1.20	1.00		1.20	1.00	***0.00***	1.20

Note. Omitted entries are fixed to 0 for the misspecified models. The incorrectly specified cross-loadings appear in bold and italics for misspecified models. c are item intercepts, $\mathbf{a}_1$ are slopes on the first factor, θ_1, $\mathbf{a}_2$ are slopes on the second factor, θ_2. Factor correlations are omitted from the table, but were .30 under the true model, and .60 under Models 2 and 3, and .54 for Models 4 through 6.

constrained estimation using the so-called Lagrange multiplier (Aitchison & Silvey, 1958), hence the name LM tests. In the structural equation modeling literature, LM tests are sometimes called modification indices (e.g., Kaplan, 1988; Sörbom, 1989). In what follows, we present LM tests in their multivariate form, allowing tests of multiple parameters simultaneously.

Let $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \boldsymbol{\eta}_2')'$ be an $r \times 1$ vector of model parameters, where the dimensions of $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are r_1 and r_2, respectively. Let $\boldsymbol{\eta}_0 = (\boldsymbol{\eta}_{10}', \boldsymbol{\eta}_{20}')'$ be the corresponding vector of true parameter values. Suppose we are interested in the following hypothesis test:

$$H_0 : \boldsymbol{\eta}_{20} = \mathbf{d} \text{ vs. } H_1 : \boldsymbol{\eta}_{20} \neq \mathbf{d}, \tag{4}$$

where $\mathbf{d}$ is a vector of fixed constants. While there is more than one way to test this hypothesis, LM tests are useful in the context of constrained estimation, whereby the researcher first obtains the maximum likelihood estimates from a restricted model, $\tilde{\boldsymbol{\eta}} = (\tilde{\boldsymbol{\eta}}_1', \tilde{\boldsymbol{\eta}}_2')' = (\tilde{\boldsymbol{\eta}}_1', \mathbf{d}')'$. Thus, the parameter estimates corresponding to $\boldsymbol{\eta}_{20}$ are fixed to $\mathbf{d}$. This differs from the case whereby the researcher first obtains the unrestricted maximum likelihood estimates, $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2)'$. If this unrestricted model is correctly specified, $\hat{\boldsymbol{\eta}}$ converges to $\boldsymbol{\eta}_0$. If H_0 is also true, then $\tilde{\boldsymbol{\eta}}$ also converges to $\boldsymbol{\eta}_0$. To connect this framework to the current research, let α be the omitted cross-loading that is

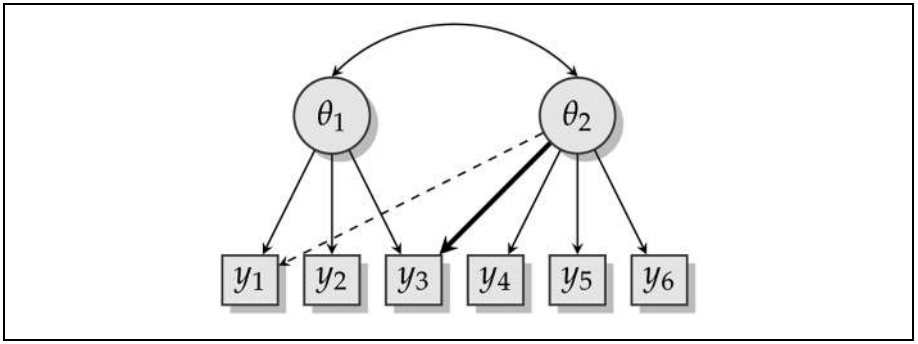

Figure 3. Two-factor model with incorrectly specified cross-loading.
Note. The dashed line indicates a true cross-loading that is omitted from the model. The bold line indicates a cross-loading that is modeled, but not present in the data-generating model.

the focus of the LM test, with population value α_0. Then, $\boldsymbol{\eta}_{20} = \alpha_0$, $r_2 = 1$, and $\mathbf{d} = 0$. That is, the null hypothesis is $H_0 : \alpha_0 = 0$, and the alternative hypothesis is $H_1 : \alpha_0 \neq 0$.

Before presenting the LM test, we define some additional quantities. Denote

$$\dot{l}(\boldsymbol{\eta}|\mathbf{y}_i) = \frac{\partial l(\boldsymbol{\eta}|\mathbf{y}_i)}{\partial \boldsymbol{\eta}} \text{ and } \ddot{l}(\boldsymbol{\eta}|\mathbf{y}_i) = \frac{\partial^2 l(\boldsymbol{\eta}|\mathbf{y}_i)}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'}. \tag{5}$$

as the gradient and Hessian, respectively, of the log-likelihood for a single observation. Then, the Fisher information, or (negative) expected Hessian, for one observation is

$$\mathbf{A}(\boldsymbol{\eta}) = -E\left[\ddot{l}(\boldsymbol{\eta}|\mathbf{y}_i)\right]. \tag{6}$$

Under correct model specification, $\mathbf{A}$ is equivalent to the expected cross-product matrix:

$$\mathbf{B}(\boldsymbol{\eta}) = E\left[\dot{l}(\boldsymbol{\eta}|\mathbf{y}_i)\dot{l}(\boldsymbol{\eta}|\mathbf{y}_i)'\right]. \tag{7}$$

The LM test is defined as

$$LM = \frac{1}{N}\mathbf{s}(\tilde{\boldsymbol{\eta}})'\mathbf{A}(\tilde{\boldsymbol{\eta}})^{-1}\mathbf{s}(\tilde{\boldsymbol{\eta}}), \tag{8}$$

where $\mathbf{s}(\tilde{\boldsymbol{\eta}}) = \sum_{i=1}^{N} \dot{l}(\tilde{\boldsymbol{\eta}}|\mathbf{y}_i)$ is the score vector for the entire sample, and $\mathbf{A}(\tilde{\boldsymbol{\eta}})$ is the Fisher information evaluated at the constrained maximum likelihood estimates. Under H_0 and a correctly specified model, LM is asymptotically distributed as a central chi-square with r_2 degrees of freedom. Since the first r_1 elements in $\mathbf{s}(\tilde{\boldsymbol{\eta}})$ are zero

due to maximum likelihood estimation, Equation (8) is often presented in reduced form (e.g., Glas, 1999):

$$LM = \frac{1}{N}\mathbf{s}_2(\tilde{\boldsymbol{\eta}})'\mathbf{A}(\tilde{\boldsymbol{\eta}})^{(22)}\mathbf{s}_2(\tilde{\boldsymbol{\eta}}), \tag{9}$$

where $\mathbf{s}_2(\tilde{\boldsymbol{\eta}})$ is the subset of $\mathbf{s}(\tilde{\boldsymbol{\eta}})$ corresponding to $\tilde{\boldsymbol{\eta}}_2$, and $\mathbf{A}(\tilde{\boldsymbol{\eta}})^{(22)}$ is the partitioned inverse corresponding to $\boldsymbol{\eta}_2$ (e.g., Schott, 2005). Temporarily suppressing the notation indicating dependence on $\boldsymbol{\eta}$, for the partition

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \tag{10}$$

the partitioned inverse needed for Equation (9) is

$$\mathbf{A}^{(22)} = \left(\mathbf{A}_{11} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}. \tag{11}$$

In practice, the Fisher information in Equations 8 and 9 cannot be calculated as the true parameters, $\boldsymbol{\eta}$, are unknown. Even if substituting $\tilde{\boldsymbol{\eta}}$ for $\boldsymbol{\eta}$, the expectation for $\mathbf{A}(\tilde{\boldsymbol{\eta}})$ requires computation over all possible response patterns, the number of which is exponential in n. Thus, this computation is in general not feasible for long tests. Instead, a different consistent estimate of $\mathbf{A}(\boldsymbol{\eta})$ is used.

A popular approach is to use sample-based counterparts to $\mathbf{A}$ and $\mathbf{B}$, evaluated at the constrained maximum likelihood estimates $\tilde{\boldsymbol{\eta}}$. These consistent estimates are the *observed Hessian*:

$$\hat{\mathbf{A}} = -\frac{1}{N}\sum_{i=1}^{N}\ddot{l}(\tilde{\boldsymbol{\eta}}|\mathbf{y}_i), \tag{12}$$

and *observed cross-product*:

$$\hat{\mathbf{B}} = \frac{1}{N}\sum_{i=1}^{N}\dot{l}(\tilde{\boldsymbol{\eta}}|\mathbf{y}_i)\dot{l}(\tilde{\boldsymbol{\eta}}|\mathbf{y}_i)'. \tag{13}$$

Both of these approximations are studied in our simulations and are computed after parameter estimation. The derivatives involved in these matrices are well known (e.g., see Liu & Thissen, 2012; Yuan et al., 2014). In practice, $\hat{\mathbf{B}}$ is faster to compute than $\hat{\mathbf{A}}$ and is more frequently used to study LM tests in the IRT modeling context.

Lagrange Multiplier Tests for Misspecified Models

When the unrestricted model is correctly specified, the unconstrained maximum likelihood estimates converge to the true population values: $\hat{\boldsymbol{\eta}} \xrightarrow{P} \boldsymbol{\eta}_0$. However, if the model is misspecified, and certain regularity conditions are satisfied (e.g., White, 1982). $\hat{\boldsymbol{\eta}}$ converges to a different stationary point, say $\boldsymbol{\eta}^* = (\boldsymbol{\eta}_1^{*'}, \boldsymbol{\eta}_2^{*'})'$. In general,

$\boldsymbol{\eta}^* \neq \boldsymbol{\eta}_0$, though depending on the form and degree of misspecification, some elements of $\boldsymbol{\eta}^*$ may equal their counterparts in $\boldsymbol{\eta}_0$. In particular, we may wonder whether $\boldsymbol{\eta}_2^* = \boldsymbol{\eta}_{20}$ or even $\boldsymbol{\eta}_{20} = \mathbf{d}$. However, under misspecification, it is only possible to test hypotheses such as:

$$H_0 : \boldsymbol{\eta}_2^* = \mathbf{d} \text{ vs. } H_1 : \boldsymbol{\eta}_2^* \neq \mathbf{d}. \tag{14}$$

If this null hypothesis is true, the restricted estimates, $\tilde{\boldsymbol{\eta}}$, will also converge to $\boldsymbol{\eta}^*$. In general, we cannot carry out inference for $\boldsymbol{\eta}_{20}$ under misspecification.

Under misspecification, it is well known that, in general, $\mathbf{A}(\boldsymbol{\eta}) \neq \mathbf{B}(\boldsymbol{\eta})$, and neither $\hat{\mathbf{A}}^{-1}/N$ nor $\hat{\mathbf{B}}^{-1}/N$ provides a consistent estimate of the covariance matrix of the maximum likelihood estimates (White, 1982). Instead, a consistent estimate can be obtained from the so-called sandwich covariance matrix, $\hat{\mathbf{C}}/N = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N$. Similarly, under misspecification, *LM* as defined in Equation (8) will not be asymptotically distributed as a chi-square with r_2 degrees of freedom. But, an adjustment can be made to the weight matrix of *LM* to form a generalized version of the test (e.g., Boos, 1992; Engle, 1984; White, 1982):

$$LM_s = \frac{1}{N}\mathbf{s}_2(\tilde{\boldsymbol{\eta}})'\left(\hat{\mathbf{A}}^{(22)}\right)^{-1}\hat{\mathbf{C}}_{22}^{-1}\left(\hat{\mathbf{A}}^{(22)}\right)^{-1}\mathbf{s}_2(\tilde{\boldsymbol{\eta}}), \tag{15}$$

where $\hat{\mathbf{A}}^{(22)}$ is defined in Equation (11) and $\hat{\mathbf{C}}_{22}$ is the block of $\hat{\mathbf{C}}$ corresponding to $\boldsymbol{\eta}_2^*$. This *sandwich*-based *LM* test has the desired asymptotic distribution, that is, chi-square with r_2 degrees of freedom when H_0 is true. However, this generalized version of *LM* has apparently never been studied in an IRT context.

To connect this theory to the current research, again consider the true model in Figure 2 with item parameters in Table 1. The single true cross-loading is from θ_2 to Item 1. Suppose the researcher fits the simple structure model in Figure 1 and then uses an LM test for the cross-loading from θ_2 to item 3 as in Figure 3. Since the model in Figure 3 is misspecified, $\hat{\alpha} \xrightarrow{P} \alpha^*$. While $\alpha_0 = 0$, the value of α^* is not obvious.

Given the small size of this model, and the true parameters $\boldsymbol{\eta}_0$, we can find $\boldsymbol{\eta}^*$ (and α^*) in the following way.[2] Using $\boldsymbol{\eta}_0$, the true response pattern probabilities for all $2^6 = 64$ unique response patterns can be calculated by Equation (2). Then, the misspecified model (Figure 3) can be fit directly to these true probabilities by maximum likelihood, yielding $\boldsymbol{\eta}^*$. By this method, we find that $\alpha^* = -0.44$ (see Model 3 in Table 1). Given this particular misspecification, $\alpha_0 = 0$, while $\alpha^* \neq 0$. Thus, for the correctly specified model, the null hypothesis $H_0 : \alpha_0 = 0$ is true, while for the misspecified model, $H_0 : \alpha^* = 0$ is false. In this case, rejection of H_0 under the misspecified model is the correct decision and should not constitute a Type I error. However, since this decision is not consistent with the true data-generating model, we may refer to such rejections as *false positives*.

For a different misspecification, it is possible that the null hypothesis is true for both the correct and misspecified models. For example, consider the model with a

single cross-loading from θ_1 to item 4. The corresponding population parameters for this model are in Table 1, under the ''Model 4'' label. For this misspecification, $\alpha_0 = \alpha^* = 0$, and the LM null hypothesis for the misspecified model is true. This pattern holds when there is a single unmodeled cross-loading: If the misspecified cross-loading involves the same latent dimension as the unmodeled cross-loading, $\alpha^* \neq 0$; otherwise $\alpha^* = 0$. Since this is a small model, we report all other single cross-loading specifications in Models 2 through 6 in Table 1.

This example illustrates the challenge of using LM tests in general, and in particular to study cross-loadings under misspecification. Even when $\alpha_0 = 0$, α^* may or may not equal zero, and the effect of misspecification on hypothesis testing can be difficult to evaluate (see, e.g., Yuan, Marshall, & Bentler, 2003). While not considered here, it is also possible that for a particular model, $\alpha_0 \neq 0$, but some misspecification will result in $\alpha^* = 0$. While the current research focuses on LM tests, the impact of misspecification applies equally to other tests. For example, this phenomenon has been explored in the context of likelihood-ratio tests, in both IRT (Maydeu-Olivares & Cai, 2006) and covariance structure modeling (Yuan & Bentler, 2004) frameworks. And the possible bias of LM tests under misspecification has been known for some time (e.g., Byron, 1972).

Monte Carlo Study

Method

One purpose of the simulation study was to compare the performance of the observed cross-product, observed Hessian, and sandwich approaches to computing LM tests under a MIRT model with possibly omitted cross-loadings. We also sought to examine LM test performance in terms of Type I error under both correct and misspecified models such as those in the previous section, and whether these tests have utility under progressively misspecified models in distinguishing between parameters fixed to their population data-generating values and parameters that are not.

To this end, we employed a 3 (LM method: cross-product, Hessian, sandwich) $\times$ 3 (sample size: N = 200, 500, 1,000) $\times$ 3 (number of items, n = 10, 20, 50) overall design. These conditions were fully crossed with 10 different conditions representing varying levels of misspecification (or lack thereof). The data-generating model and that initially fit to the data were always a correlated two-factor model (with .3 correlation) with an equal number of main items per factor such as that presented earlier in this article. Thus, under the null condition, no cross-loadings were present. The remaining conditions had cross-loadings in a 3 (number of cross-loadings: 1 item, 20% of items, 40% of items) $\times$ 3 (size of cross-loading slope: .5, .75, 1) design. In the one cross-loading condition, the single cross-loading was from the second factor to the first item—analogous to the true model in Table 1. In the 20% and 40% of item conditions, the cross-loadings were equally split between the two latent factors and all cross-loadings had the same value for any given cell of the design. Other item parameters were randomly generated across replications. Specifically, for item slopes,

$a \sim$log-normal$(0, .25^2)$, $c \sim \mathcal{N}(0, .5^2)$ for items on Factor 1 and $c \sim \mathcal{N}(0, 1^2)$ for items on Factor 2.

One hundred data sets were generated using R (R Core Team, 2015) under each of these conditions. Given that each possible cross-loading was tested for every data set, 100 replications were deemed sufficient. To each data set, we fit the correlated two-factor model with flexMIRT® (Cai, 2012) using rectangular quadrature with 101 equally spaced points between −6 and 6 for each latent dimension. The score vector and matrices $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ were computed using custom R code to compute analytical derivatives after reading in flexMIRT® output.[3] The integrals involved in computing $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ were evaluated using rectangular quadrature with 49 equally spaced points between −6 and 6 for each latent dimension. Univariate LM tests were then computed separately for all possible cross-loadings for each replication.

Under the null condition (i.e., $\alpha_0 = 0$ for all cross-loadings), the three LM methods are asymptotically equivalent, and Type I error rates can be compared. Since there is no misspecification, we would expect similar performance across the methods, except possibly with smaller sample sizes and longer tests where high Type I error rates are sometimes observed.

If the true model contains one cross-loading, there are three possible scenarios for the LM test. First, if a cross-loading where $\alpha_0 = 0$ and $\alpha^* = 0$ is tested, then we can compare Type I error rates. For this scenario, the alternative model is misspecified, and we would expect the sandwich method to demonstrate the best performance. Second, if the cross-loading where $\alpha_0 \neq 0$ is tested, then we can study power. Third, if a cross-loading where $\alpha_0 = 0$ and $\alpha^* \neq 0$ is tested, false positive rates can be compared (which technically also constitute power since H_0 is false).

Finally, if the true model contains multiple cross-loadings, then greater misspecification results and for any tested cross-loading in our study, $\alpha^* \neq 0$ (i.e., the null hypothesis is never true). Under these conditions, there are still two possible scenarios. If a cross-loading is tested where $\alpha_0 \neq 0$, power can be studied. And if a cross-loading is tested where $\alpha_0 = 0$, false positive rates can be studied. With multiple cross-loadings, we have no theoretical expectations regarding which LM approach may be most powerful. In addition, we are interested in the relative magnitudes of the false positive and power rates.

Results

Model convergence was assessed by recording replications that reached a maximum iteration limit of 5,000 and replications with any slope greater than 22.35 (arguably Heywood cases or improper solutions).[4] In total, only 35 out of 9,000 fitted models did not converge. Most convergence problems were concentrated in the 10-item and $N = 200$ cells with 1 cross-loading (6 with slope size = .75; 7 with slope size = 1.0) or 20% cross-loadings (3 with slope size = .5; 5 with slope size = .75; 8 with slope size = 1.0). However, no cell had fewer than 92/100 valid replications, and all other cells not explicitly mentioned had 99/100 or more valid replications. Null hypothesis

rejection rates for score tests are reported as a proportion considering only valid replications and using the total number of valid tests.

The Hessian approach—and only this approach—sometimes resulted in a negative test statistic. Follow-up analyses on select data sets revealed that the profile log-likelihood near zero for restricted cross-loadings was irregular and that $\hat{\mathbf{A}}$ for these score tests was not positive definite. The corresponding element of $\hat{\mathbf{A}}^{-1}$ (or $\hat{\mathbf{A}}^{(22)}$) was thus negative, resulting in a negative score test. How one should treat such test statistics is open to debate. A similar phenomenon has been found to occur with scaled test statistics for model comparisons in structural equation modeling (Chuang, Savalei, & Falk, 2015; Satorra & Bentler, 2001, 2010). In this literature, the model (restricted, unrestricted, or both) is sometimes considered misspecified and not interpreted—especially if overall tests of fit also reject the model. As we will show later, negative test statistics for the Hessian approach can occur under minimal or no misspecification. Alternatively, such test statistics are sometimes rounded to the nearest valid value, which is zero in this case. In this article, we adopt this latter approach and counted such cases as valid non-rejections of H_0, but also explicitly report when such cases occurred.

As described above, $H_0 : \alpha_0 = 0$ was true only under the conditions with 0 or 1 cross-loading. More specifically, the null was true for all cross-loadings under the no cross-loading conditions, and $n/2$ cross-loadings under the one cross-loading condition. Figure 4 depicts the corresponding Type I error rates, with cells with higher Type I error rates shaded in progressively darker shades of red and cells with lower Type I error rates in darker shades of blue. The horizontal panels of this Figure correspond to each type of LM test method—with the number of items varying across each row of each panel. Similarly, vertical panels correspond to the data-generating condition and columns within each vertical panel indicate different sample sizes.

Comparing the three approaches, the cross-product approach tended to have inflated Type I error rates (as high as .195), and these rates tended to increase with *more* items and at *smaller* sample sizes. Even at the largest sample size ($N = 1,000$), these rates tended to exceed .06 and sometimes .07 with $n = 50$ items. The Hessian approach had much better calibrated Type I error rates, with most rejection rates near the nominal rate. However, some overrejection was observed for the Hessian approach (as high as .088) with *fewer* items and at *smaller* sample sizes (e.g., $N = 200$ and $n = 10$). Thus, the Type I error rates for the Hessian tended to drop as the number of items increased. In contrast, the sandwich approach tended to *underreject* (as low as .006) with *fewer* items and *smaller* sample sizes, but otherwise was close to the nominal rate in most conditions and never exceeded .056. Averaging across all data-generating conditions with each cell contributing an equal amount to rejection rates, the LM test methods had rejection rates of .08 (cross-product), .06 (Hessian), and .04 (sandwich).

Figure 4 also allows us to evaluate the relative performance of the sandwich approach under misspecification, when it is the only theoretically correct approach. For the sandwich approach, the pattern of results for the zero cross-loadings

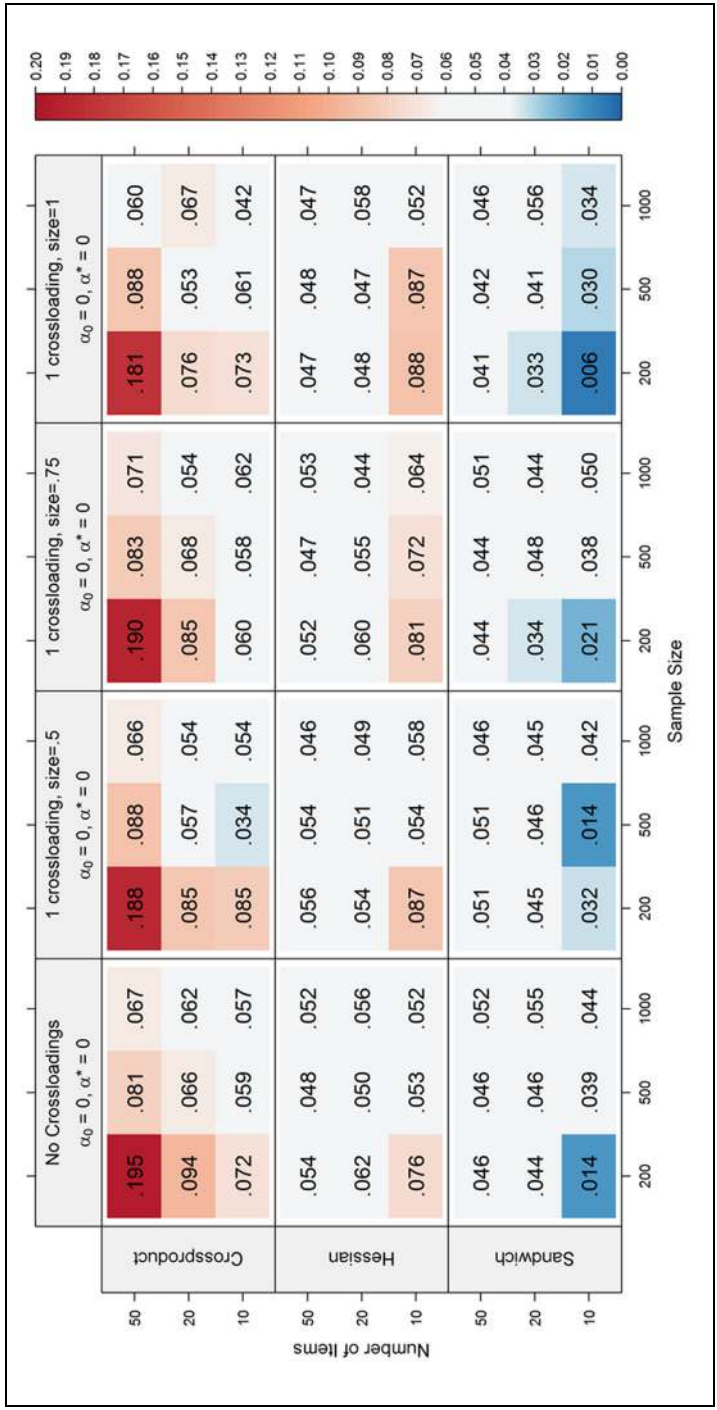

Figure 4. Type I error.
Note. Size = indicates the size of cross-loading condition; $\alpha_0 = 0$ refers to tests for slopes that are zero under the data-generating model; $\alpha^* = 0$ refers to tests for slopes that are zero under the misspecified unrestricted model.

condition (with correct model specification) is similar to the patterns of results for all the one cross-loading conditions (with model misspecification). Thus, the sandwich approach performs as expected and is not adversely affected by the model misspecification. Somewhat surprisingly, the patterns of results for the cross-product and Hessian approaches are also similar when comparing the zero cross-loading to one cross-loading conditions. For the cross-product approach, the same general pattern of liberal Type I error rates continues under misspecification. For the Hessian approach, the Type I error rates with $n = 10$ items appear to increase slightly under misspecification, which is consistent with the theory that the calibration of the Hessian-based LM test may be adversely affected under misspecification.

Figures 5, 6, and 7 depict power for the 1 cross-loading, 20% cross-loadings, and 40% cross-loadings conditions, respectively, with darker shades of red indicating higher power. The layout of these Figures is nearly identical to that of Figure 4, except the vertical panels distinguish between cases where $\alpha_0 = 0$ (i.e., false positives) and where $\alpha_0 \neq 0$. In these Figures, due to misspecification, the null hypothesis $H_0 : \alpha^* = 0$ is never true. On the other hand, the tested cross-loading may or may not actually be zero in the data-generating model. The results for these conditions (i.e., $\alpha_0 = 0$ and $\alpha_0 \neq 0$) are presented side-by-side so that it is easier to discern whether LM tests have greater power to detect cross-loadings where $\alpha_0 \neq 0$. Under misspecification, LM tests will be more useful in revising a model when rejection rates are relatively high when $\alpha_0 \neq 0$ compared to rejection rates when $\alpha_0 = 0$.

Power to detect cross-loadings was usually highest when only one cross-loading was present, when the size of the cross-loading was larger, and with more items and/or greater sample sizes (Figure 5). Typically, power under the 1 cross-loading conditions was high when $\alpha_0 \neq 0$, averaging .90 (cross-product), .76 (Hessian), and .78 (sandwich) across other data-generating conditions (sample size, number of items, size of omitted cross-loadings). Power (or false positives) under 1 cross-loading conditions was much lower when $\alpha_0 = 0$, averaging .16 (cross-product), .12 (Hessian), and .11 (sandwich) across other data-generating conditions. That is, under such mild misspecification, LM tests appear to be able to detect true cross-loadings at a relatively high rate and do not tend to reject the null hypothesis for slopes that are actually zero under the true data-generating model.

One exception to the above pattern was the tendency for false positives to increase under *larger* sample sizes and *fewer* items. Although this pattern was common to all LM test approaches, the false positive rate for the Hessian approach was sometimes higher than power to detect true cross-loadings (e.g., $n = 10$ and a single cross-loading of 1.0). Given the minimal misspecification of one cross-loading, this latter result is somewhat surprising. However, the Hessian results can be partly explained by an increase in the tendency for negative score tests under similar conditions. As shown in Figures 8 and 9, the Hessian approach had up to a .98 proportion of tests that were negative, especially for $a_0 \neq 0$ slopes under large sample sizes and when there was a substantial amount of misspecification (cross-loadings $\geq .75$). Under relatively less misspecification (e.g., $a^* = 0$ or cross-loadings of .5), these negative tests were less

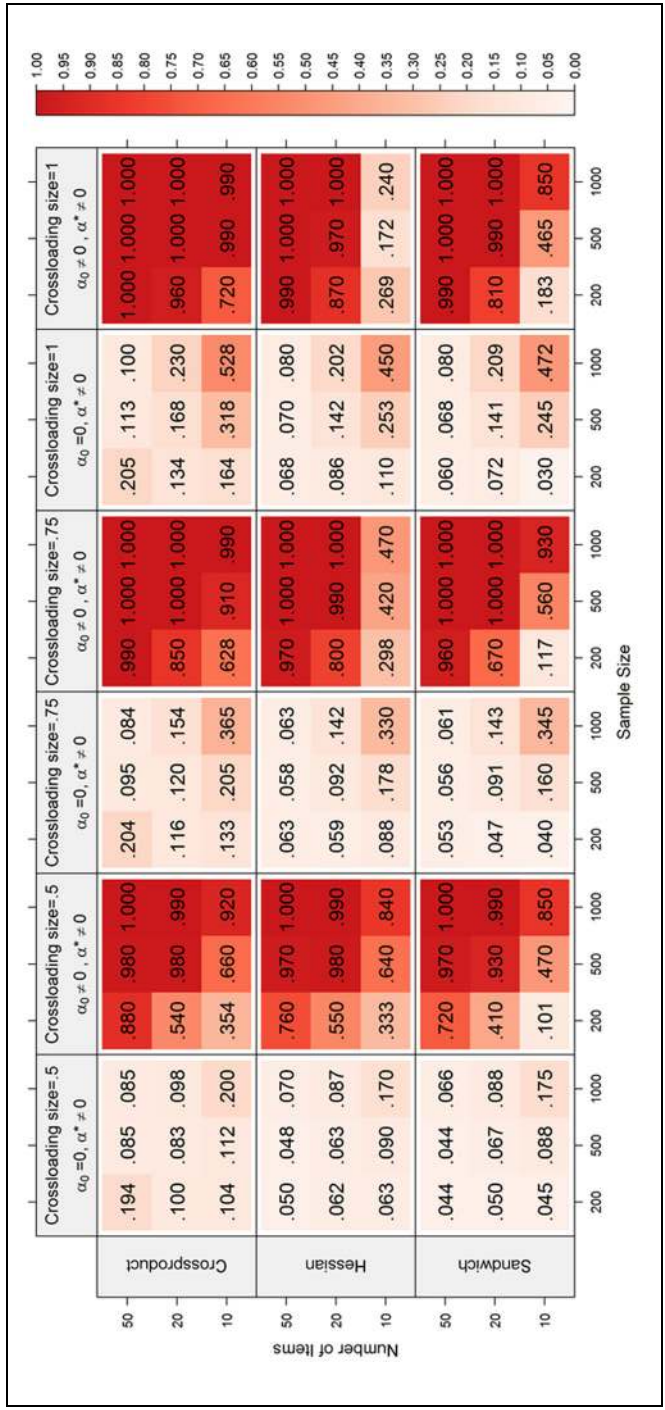

Figure 5. Power for 1 cross-loading conditions.

Note. Size = indicates the size of cross-loading condition; $\alpha_0 = 0$ with $\alpha^* \neq 0$ refers to tests for slopes that are zero under the data-generating model, but are *not* zero under the misspecified unrestricted model (*false positives*); $\alpha_0 \neq 0$ with $\alpha^* \neq 0$ refers to tests for slopes that are nonzero under both the data-generating model and unrestricted models.

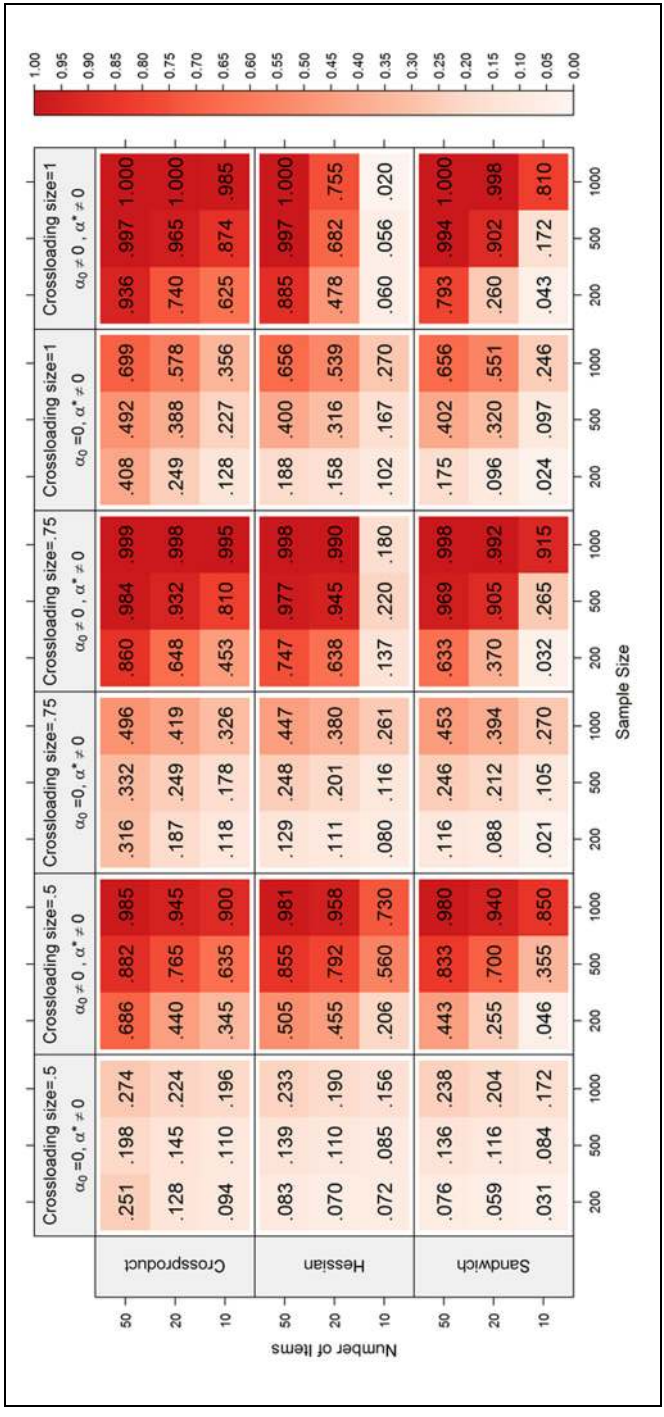

Figure 6. Power for 20% cross-loading conditions.

Note. Size = indicates the size of cross-loading condition; $\alpha_0 = 0$ with $\alpha^* \neq 0$ refers to tests for slopes that are zero under the data-generating model, but are *not* zero under the misspecified unrestricted model (*false positives*); $\alpha_0 \neq 0$ with $\alpha^* \neq 0$ refers to tests for slopes that are nonzero under both the data-generating model and unrestricted models.

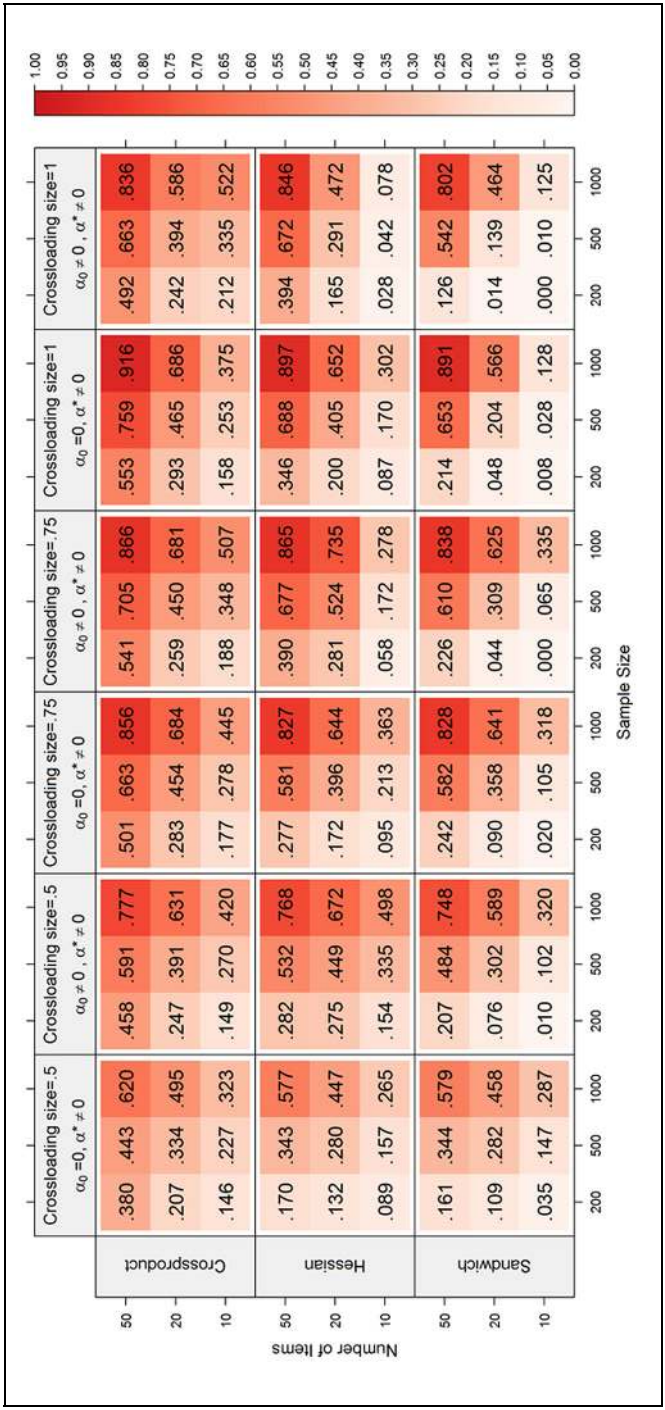

Figure 7. Power for 40% cross-loading conditions.

Note. Size = indicates the size of cross-loading condition; $\alpha_0 = 0$ with $\alpha^* \neq 0$ refers to tests for slopes that are zero under the data-generating model, but are *not* zero under the misspecified unrestricted model (*false positives*); $\alpha_0 \neq 0$ with $\alpha^* \neq 0$ refers to tests for slopes that are nonzero under both the data-generating model and unrestricted models.

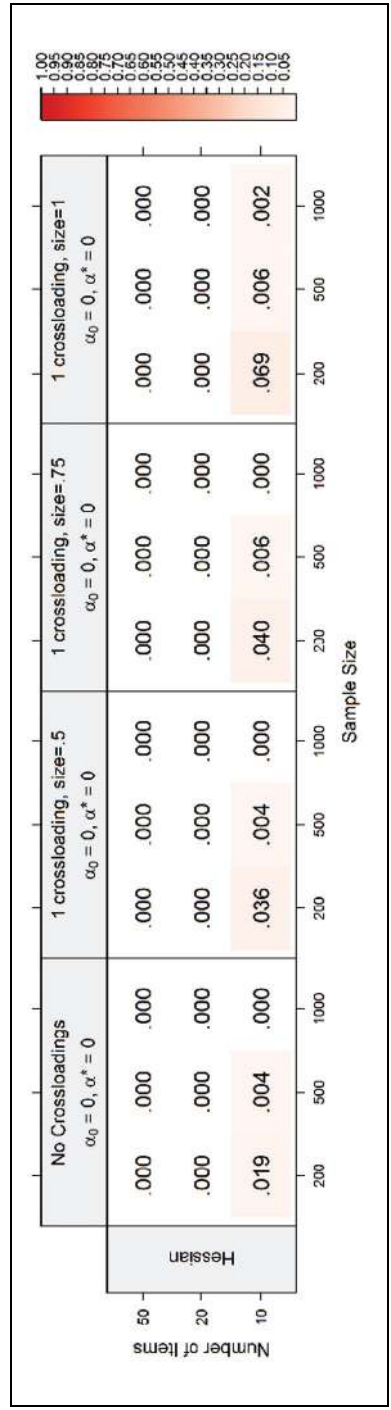

Figure 8. Proportion of negative score tests for Hessian under $\alpha^* = 0$ conditions.
Note. Size = indicates the size of cross-loading condition; $\alpha_0 = 0$ refers to tests for slopes that are zero under the data-generating model; $\alpha^* = 0$ refers to tests for slopes that are zero under the misspecified unrestricted model.

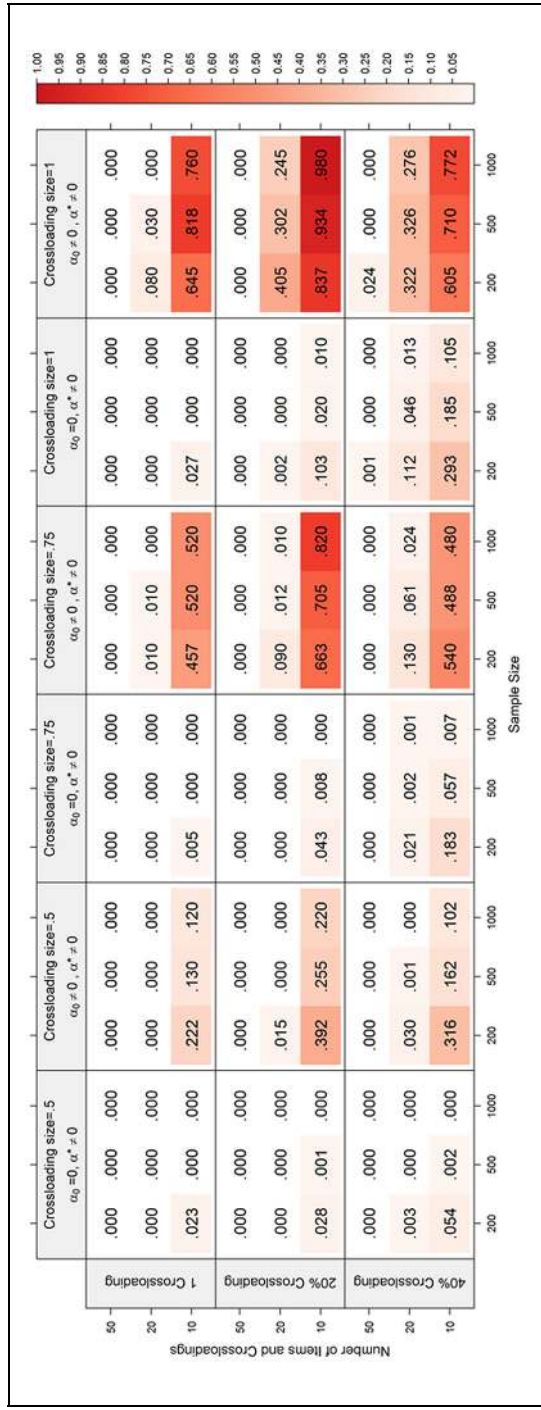

Figure 9. Proportion of negative score tests for Hessian under $\alpha^* \neq 0$ conditions.

Note. Size = indicates the size of cross-loading condition; $\alpha_0 = 0$ with $\alpha^* \neq 0$ refers to tests for slopes that are zero under the data-generating model, but are *not* zero under the misspecified unrestricted model (*false positives*); $\alpha_0 \neq 0$ with $\alpha^* \neq 0$ refers to tests for slopes that are nonzero under both the data-generating model and unrestricted models.

frequent, but tended to occur at $n = 10$ and under smaller sample sizes. Since we have counted negative tests as valid non-rejections of H_0, such results explain the lack of power for the Hessian approach at $n = 10$. This pattern also suggests a tendency for the sandwich approach to better differentiate among zero and nonzero slopes than the Hessian approach.

The results for the 20% (Figure 6) and 40% (Figure 7) conditions were similar to the 1 cross-loading conditions (Figure 5) in that power increased with more items and greater sample sizes. The Hessian approach sometimes displayed the same unexpected pattern of results at $n = 10$, and the sandwich approach tended to have relatively low power with small sample sizes. Across the approaches, the rejection rates for true zero slopes (false positives) tended to increase with the proportion of omitted cross-loadings. At the same time, power to detect true cross-loadings decreased under greater misspecification. For the 40% proportion of cross-loadings conditions, power was very similar for both types of parameters, and there were more cases where false positive rates exceeded power for true cross-loadings even under the same data generation conditions. In the 40% cross-loadings with 1.0 slopes conditions, for example, rejection was greater for true zero slopes than for true cross-loadings in 23/27 cells across all LM test methods. On average under the 20% cross-loading conditions, rejection rates were .29 (cross-product), .22 (Hessian), and .21 (sandwich) for true zero slopes (false positives), and .83 (cross-product), .62 (Hessian), and .64 (sandwich) for true cross-loadings. For 40% cross-loading conditions, these numbers were .44 (cross-product), .36 (Hessian), and .31 (sandwich) for true zero slopes, and .47 (cross-product), .40 (Hessian), and .30 (sandwich) for true cross-loadings.

Figures 5, 6, and 7 make clear that as the degree of misspecification increases, the power of the LM test depends less on the value of α_0 under the true data–generating model. In Figure 5, there is a clear difference in the results depending on whether $\alpha_0 = 0$ or $\alpha_0 \neq 0$. In sharp contrast, in Figure 7, the pattern of results are very similar for the $\alpha_0 = 0$ and $\alpha_0 \neq 0$ conditions. Thus, as misspecification increases, the LM test loses the ability to distinguish between $\alpha_0 = 0$ and $\alpha_0 \neq 0$.

Discussion

Our Monte Carlo simulation study examined the performance of the observed cross-product, Hessian, and sandwich (generalized LM test) approaches to computing LM tests under a variety of levels of model misspecification. Overall, our results indicate a number of differences among LM test approaches that are reconcilable with previously published simulation results. For instance, the tendency of the cross-product approach to have high Type I error rates, especially under longer tests and fewer respondents, is consistent with other research using this approach (Liu & Maydeu-Olivares, 2012; Liu & Thissen, 2012, 2014). In contrast, Type I error rates were in general lower for the Hessian approach and dropped to nominal levels under the same conditions—a pattern of findings consistent with Kim et al. (2011). On the other hand, the sandwich approach never had inflated Type I errors, even under

misspecification. This suggests that the Hessian and sandwich approaches are preferable to the cross-product in terms of Type I error control, and use of these approaches may yield more acceptable Type I error rates for LM tests previously introduced in the IRT literature (e.g., Liu & Maydeu-Olivares, 2012; Liu & Thissen, 2012, 2014).

Any recommendation for the Hessian must be qualified by the observation that it may fail to yield an easily interpretable test statistic when there are few administered items and much misspecification. While we have counted negative tests as nonrejections of H_0, these tended to occur when in fact H_0 was false. It therefore may be prudent in future research to consider such cases actually as possible rejections of H_0, or better yet, to use an alternative such as the sandwich approach.

Turning to power and the ability to differentiate between zero and nonzero data generating slopes under misspecification, a discernible pattern of results across methods was less obvious. The cross-product approach had the highest power, but this was to be expected given the inflated Type I error rates. In addition, we observed poor performance for the Hessian approach for larger samples and fewer items. Thus, given Type I error control and more dependable performance under misspecification, we recommend the sandwich approach and encourage its further study and adoption in IRT software.

The degree of misspecification in the initially fitted and alternative models made the most impactful difference on the ability of LM tests to differentiate among zero and nonzero true slopes. Whereas Glas and Falcón (2003) typically found that LM tests provided some utility under their studied conditions, we have arguably shown that LM tests have questionable utility when the alternative model is highly misspecified. This finding is consistent with results presented in Yuan and Bentler (2004) and Maydeu-Olivares and Cai (2006) on the behavior of the likelihood ratio test under misspecification. Given the asymptotic equivalence between the LM and likelihood ratio tests, this finding is expected. Recent investigations of LM tests in the context of IRT ought to be interpreted with this in mind—the performance of newly developed LM tests is expected to change under misspecification, and it is important to identify various possible forms of misspecification. Thus, one possible future research direction is to study whether a similar pattern of results holds for certain tests of local dependence, functional form of the response function, and so on.

A number of limitations of the present study lead us to suggest other future directions. In the covariance structure modeling literature, Green and colleagues (Green & Thompson, 2010; Green, Thompson, & Poirier, 1999) have attempted a number of methods to control Type I error and any of these approaches could be tried in the IRT context. We would only expect these approaches to have utility in the present context to the extent that rejection rates for true zero slopes (false positives) are reduced at a faster rate than that for true cross-loadings. Although Liu and Maydeu-Olivares (2012) found that the expected Hessian also maintained better Type I error control, a feasible computational approach for longer tests still remains to be seen.

In the current article, we also consider only the case where the misspecification (omitted cross-loading) is of the same type of parameter as that being examined with the LM test. Other unexplored possibilities could include whether LM tests for cross-loadings tend to have inflated Type I error when the functional form assumption is violated, or any other combination of misspecifications. Such a case is likely to be realized in practice. It is also desirable to study the performance of LM tests when some aspect of the data-generating model is nonparametric and/or data follows an unknown distribution.

LM tests are sometimes criticized within the context of model specification searches. If used in such a search, the researcher may focus first on freeing parameter(s) corresponding to the largest LM test. While examining the utility of such a strategy may entail a different set of simulations than we present or different analyses on our simulation results, this strategy may not always work well in practice (MacCallum, 1986). Researchers ought to be cognizant that LM tests are not necessarily designed to find the correctly specified model and may not do so unless there is minimal misspecification. Rather, one needs to be aware that the null hypothesis being tested under misspecification may not be true, even if the parameter of interest is fixed to its true data-generating value. Instead, a final future direction may be to consider other alternatives for specification searches that may be more fruitful than univariate LM tests (Marcoulides, Drezner, & Schumacker, 1998; Marcoulides & Leite, 2014). In sum, significant LM tests may be safely used as an indication that a model is misspecified, but are not guaranteed to identify the source of the misspecification.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

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

Notes

1. In the broader IRT literature, the observed Hessian is sometimes referred to as just the ''observed (Fisher) information matrix,'' the observed cross-product as the ''empirical cross-product,'' and expected Hessian as the ''Fisher information matrix.''
2. The procedure described here is completely analogous to fitting a covariance structure (or structural equation) model to the population covariance matrix.
3. While flexMIRT® can compute $\hat{\mathbf{B}}$ using ''SE = Xpd'' as an option, this matrix will not contain elements corresponding to the to-be-freed parameter(s). Similarly, while flexMIRT® can approximate $\hat{\mathbf{A}}$ by the supplemented EM algorithm (''SE = SEM''; Cai, 2008), again, this matrix will not include the to-be-freed parameter(s). Thus, R code was

written to compute the analytical versions of $\hat{\mathbf{B}}$ and $\hat{\mathbf{A}}$ for the purposes of this research. The accuracy of the programming was checked against flexMIRT® when possible and against numerical derivatives otherwise.

4. This value roughly corresponds to a standardized normal ogive slope of .999 (e.g., see Forero & Maydeu-Olivares, 2009).

References

Aitchison, J., & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to constraints. *Annals of Mathematical Statistics*, *19*, 813-828.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443-459.

Boos, D. (1992). On generalized score tests. *The American Statistician*, *46*, 327-333.

Byron, R. P. (1972). Testing for misspecification in econometric systems using full information. *International Economic Review*, *13*, 745-756.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*, 309-329.

Cai, L. (2012). flexMIRT®: Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group.

Chou, C.-P., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange Multiplier, and Wald tests. *Multivariate Behavioral Research*, *25*, 115-136.

Chuang, J., Savalei, V., & Falk, C. F. (2015). Investigation of Type I error rates of three versions of robust chi-square difference tests. *Structural Equation Modeling*, *22*, 517-530.

Engle, R. (1984). Wald, likelihood ratio, and lagrange multiplier tests in econometrics. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (1st ed., Vol. 2, pp. 775-826). Elsevier.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*, 275-299.

Fox, J. P., & Glas, C. A.W. (2005). Bayesian modification indices for IRT models. *Statistica Neerlandica*, *59*, 95-106.

Glas, C. A.W. (1998). Detection of differential item functioning using lagrange multiplier tests. *Statistica Sinica*, *8*, 647-667.

Glas, C. A.W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, *64*, 273-294.

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, *72*, 159-180.

Glas, C. A. W., & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*, 87-106.

Green, S. B., & Thompson, M. S. (2010). Can specification searches be useful for hypothesis generation? *Journal of Modern Applied Statistical Methods*, *9*, 160-171.

Green, S. B., Thompson, M. S., & Babyak, M. A. (1998). A Monte Carlo investigation of methods for controlling Type I errors with specification searches in structural equation modeling. *Multivariate Behavioral Research*, *33*, 365-383.

Green, S. B., Thompson, M. S., & Poirier, J. (1999). Exploratory analyses to improve model fit: Errors due to misspecification and a strategy to reduce their occurrence. *Structural Equation Modeling*, *6*, 113-126.

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, *23*, 69-86.

Kim, D., De Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2011). The comparative performance of conditional independence indices. *Applied Psychological Measurement*, *35*, 447-471.

Liu, Y., & Maydeu-Olivares, A. (2012). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, *73*, 254-274.

Liu, Y., & Thissen, D. (2012). Indentifying local dependence with a score test statistic based on the bifactor logistic model. *Applied Psychological Measurement*, *36*, 670-688.

Liu, Y., & Thissen, D. (2014). Comparing score tests and other local dependence diagnostics for the graded response model. *British Journal of Mathematical and Statistical Psychology*, *67*, 496-513.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, *100*, 107-120.

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *38*, 113-139.

Marcoulides, G. A., Drezner, Z., & Schumacker, R. E. (1998). Model specification searches in structural equation modeling using Tabu search. *Structural Equation Modeling*, *5*, 365-376.

Marcoulides, G. A., & Leite, W. (2014). Exploratory data mining algorithms for conducting searches in structural equation modeling: A comparison of some fit criteria. In J. J. McAardle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 150-171). New York, NY: Routledge.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*, 71-101.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2(dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, *41*, 55-64.

Obserski, D. L., van Kollenburg, G. H., & Vermunt, J. K. (2013). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, *7*, 267-279.

Paek, I., & Cai, L. (2014). A comparison of item parameter standard error estimation procedures for unidimensional and multidimensional item response theory modeling. *Educational and Psychological Measurement*, *74*, 58-76.

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

Ranger, J., & Kuhn, J.-T. (2012). Assessing fit of item response models using the information matrix test. *Journal of Educational Measurement*, *49*, 247-268.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, *44*, 50-57.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507-514.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*, 243-248.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533-555.

Schott, J. R. (2005). *Matrix analysis for statistics*. Hoboken, NJ: Wiley.

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*, 371-384.

Tian, W., Cai, L., Thissen, D., & Xin, T. (2013). Numerical differentiation methods for computing error covariance matrices in item response theory modeling: An evaluation and a new proposal. *Educational and Psychological Measurement*, *73*, 412-439.

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

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1-25.

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and z-tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, *64*, 737-757.

Yuan, K.-H., Cheng, Y., & Patton, J. (2014). Information matrices and standard errors for MLEs of item parameters in IRT. *Psychometrika*, *79*, 232-254.

Yuan, K.-H., Marshall, L. L., & Bentler, P. M. (2003). Assessing the effect of model misspecifications on parameter estimates in structural equation models. *Sociological Methodology*, *33*, 241-265.