Accepted in Multivariate Behavioral Research

Evaluating Structural Equation Models for Categorical Outcomes: A New Test Statistic and a Practical Challenge of Interpretation

Scott Monroe

Li Cai

University of California, Los Angeles

March 18, 2015

The authors thank the Associate Editor and reviewers for their helpful suggestions. Part of this research is supported by an Institute of Education Sciences statistical methodology grant (R305D140046). The views expressed here belong to the authors and do not reflect the views or policies of the funding agency.

Address all correspondence to: Scott Monroe, CRESST, UCLA, Los Angeles, CA, USA 90095-1521. Email: scott.monroe@ucla.edu. Phone: 310.926.9665. Fax: 310.825.3883.

Evaluating Structural Equation Models for Categorical Outcomes: A New Test Statistic and a Practical Challenge of Interpretation

Abstract

This research is concerned with two topics in assessing model fit for categorical data analysis. The first topic involves the application of a limited-information overall test, introduced in the item response theory literature, to Structural Equation Modeling (SEM) of categorical outcome variables. Most popular SEM test statistics assess how well the model reproduces estimated polychoric correlations. In contrast, limited-information test statistics assess how well the underlying categorical data are reproduced. Here, the recently introduced C_2 statistic of Cai and Monroe (2014) is applied. The second topic concerns how the Root Mean Square Error of Approximation (RMSEA) fit index can be affected by the number of categories in the outcome variable. This relationship creates challenges for interpreting RMSEA. While the two topics initially appear unrelated, they may conveniently be studied in tandem since RMSEA is based on an overall test statistic, such as C_2. The results are illustrated with an empirical application to data from a large-scale educational survey.

Keywords: Limited-information testing, structural equation modeling, categorical data analysis, RMSEA

1 Introduction

This research concerns two distinct but related topics in assessing the fit of latent variable models for ordered categorical data. The first topic is the application of the limited-information overall test statistic C_2 (Cai & Monroe, 2014) to Structural Equation Modeling (SEM). The second topic is how the Root Mean Square Error of Approximation index (RMSEA; Steiger & Lind, 1980) is affected by the number of categories in the outcome variable. An important connection between the two topics is that RMSEA is based on non-centrality (population lack of fit) estimated from an overall goodness-of-fit (GOF) test statistic, such as C_2. That is, RMSEA also depends on the choice of underlying overall test statistic, since different test statistics lead to different manifestations of non-centrality.

To appreciate the motivation for the application of C_2, it is helpful to consider the testing of structural models for continuous data. In this case, a sample covariance matrix summarizes the continuous data. Then, following estimation, a test statistic is formed that measures how well the structural model reproduces the sample covariance matrix. Depending on the estimation approach, a moment correction (e.g., Satorra & Bentler, 1994; Asparouhov & Muthén, 2010) can be applied to the test statistic so that it approximately follows a chi-square distribution.

Currently, in many popular SEM software packages, the standard procedure for estimating structural models for ordinal variables is the multistage estimator (e.g., Muthén, 1984). With this estimator, a polychoric correlation matrix is estimated from the categorical data. Then, typically, testing the structural model proceeds as in the continuous case. More specifically, a test statistic is formed that measures how well the structural model reproduces the estimated polychoric correlation matrix. Also, a moment correction is applied to the test statistic.

While the two procedures just described are quite similar, a fundamental distinction exists. As noted in Muthén (1993), unlike the sample covariances of continuous variables, the estimated polychoric correlations of categorical variables are model-based. Specifically, in practice, it is assumed that the observed categorical data arise from discretizing a multivariate normal density. Given this additional stage of estimation, it is arguably necessary to test the structural model directly against the observed categorical data. This can be accomplished using limited-information test statistics, such as C_2.

The C_2 statistic is among a number of limited-information tests that have been developed recently (e.g., Maydeu-Olivares & Joe, 2006; Cai & Hansen, 2013) for models of categorical data. For n observed categorical variables, the data can be organized in an n-way contingency table. While full-information tests, such as Pearson's X^2, depend on the entire n-way table, limited-information tests are "limited" in the sense that they depend on some subset of lower-order marginal tables. For C_2, the subscript denotes the use of marginal tables up to the second-order (i.e., first- and second-order). In comparison to full-information tests, limited-information tests have two main advantages: they are better-calibrated (Maydeu-Olivares & Joe, 2006) and potentially more powerful (Joe & Maydeu-Olivares, 2010). These advantages are more pronounced for sparse contingency tables, which are routinely encountered in applications of SEM to empirical data in the social and behavioral sciences (Bartholomew & Tzamourani, 1999).

While the limited-information testing methodology has been primarily applied to Item Response Theory (IRT) models, the methodology has also been applied to SEM. In an early application of limited-information tests, Maydeu-Olivares (2006) proposed a quadratic form in second-order residuals for this purpose. However, more recent research on the limited-information methodology (e.g., Maydeu-Olivares & Joe, 2006) has yielded tests that are practically and theoretically more appealing. One such test statistic is C_2. As discussed in Cai and Monroe (2014), C_2 is well-calibrated under a variety of conditions, such as second-order marginal table sparseness, and can be computed for models with relatively few outcome variables and relatively many ordinal categories. Further, in comparison to other limited-information test statistics, C_2 can be substantially more powerful in detecting model misspecification (Cai & Monroe, 2014). The first contribution of this research, then, is to apply C_2 to SEM of ordered categorical data, specifically in the context of multistage estimation. This context also provides an opportunity to compare a limited-information test (i.e., C_2) to a moment-corrected test, which, to our knowledge, has not been done before.

As mentioned above, the second contribution of this research concerns the interpretation of RMSEA when the observed variables are categorical. Given a sufficiently large sample size, the presence of any amount of model error (e.g., MacCallum & Tucker, 1991) will lead to a proposed model being rejected by an overall GOF statistic, such as C_2. In the SEM literature, this is commonly referred to as the sample size problem (Cudeck & Henly, 1991). In response to this problem, SEM researchers have, over the years, proposed various fit indices and developed interpretive guidelines for continuous normally-distributed outcomes. For example, with the RMSEA index, a value of less than .05 is indicative of "close-fit" (Browne & Cudeck, 1993).

More recently, researchers have made efforts to adapt these indices and guidelines for use with categorical outcomes. Within the IRT framework, these indices are typically based on the limited-information M_2 statistic (Maydeu-Olivares & Joe, 2006). For example, Maydeu-Olivares (2013) developed a rationale for constructing an M_2-based RMSEA. More recently, Maydeu-Olivares & Joe (2014) expanded on this line of research and proposed some cutoff criteria for approximate fit. Another example is provided by Lee and Cai (2012), which proposed an M_2-based Tucker-Lewis Index (Tucker & Lewis, 1973). Within the SEM framework, these indices have typically been constructed from moment-corrected tests. Notwithstanding the specific framework, the interpretation of these indices has received much less attention for categorical data than for continuous data. To help address this issue, we examine how RMSEA is affected by the number of categories in the outcome variables. This choice is motivated by results reported in Cai and Monroe (2013), which suggest that RMSEA, in a sense, behaves differently depending on the number of categories of the outcome variables.

This RMSEA behavior can conveniently be studied along with C_2 due to the underlying response process formulation of factor analytic measurement models (Thurstone, 1925; Thurstone, 1927; Lord, 1952) assumed under multistage estimation. The underlying response process provides a direct connection between structural models of continuous and categorical data, which can be utilized in the following way.

First, given some form of introducing model error, a population correlation matrix of continuous variables can be created. For a chosen (working) model and discrepancy function (e.g., the maximum likelihood discrepancy function; Browne & Arminger, 1995), minimization of the function for the population correlation matrix yields a population discrepancy function value and derived population RMSEA. Next, underlying response variables can be randomly sampled from this population matrix to create datasets of continuous variables. In accordance with the underlying response variable formulation, these continuous variables may be discretized to generate categorical datasets. All of the datasets contain both model error, because of the nonzero population RMSEA, as well as sampling error. However, for the categorical datasets, the discretization itself does not introduce additional model error, assuming correct distributional specification of the underlying response process variables (e.g., multivariate normal). With a sufficiently large number of Monte Carlo replications, the sampling error may be averaged out. Then, the RMSEA estimates may be directly compared to the uniquely defined population RMSEA. We believe that the simulation results may shed some light on how RMSEA should be practically interpreted for SEM of categorical data.

The rest of the paper is organized as follows. Section 2 presents a motivating example. Section 3 presents a structural model for ordinal data and the multistage estimator. Also, established fit statistics for the multistage estimator are introduced. Then, in Section 4, limited-information testing methodology is presented and the C_2 statistic is introduced. Section 5 presents a simulation study for C_2 and the results. Section 6 explores the behavior of RMSEA, using the results from Section 5. Then, an empirical application of the proposed methods is given in Section 7. Finally, a conclusion and discussion of further research directions are provided in Section 8.

2 A Running Example

The Program for International Student Assessment (PISA; OECD, 2005) administers a student questionnaire containing various schooling and background related variables. One of these topics, surveyed in 2003, is students' perceptions of their own mathematical aptitude. Table 1 presents the 12 items hypothesized to represent three distinct but correlated constructs. These constructs are positive self-concept as a mathematics student (PSC), mathematics anxiety (ANX), and task-specific confidence (TASK). Each of the 12 items has a 4-point response scale. For PSC and ANX, the options are “strongly disagree,” “disagree,” “agree,” and “strongly agree." For TASK, the options are "not at all confident," "not very confident," "confident," and "very confident."

Insert Table 1 about here

One of the reasons PISA administers the student questionnaire is to allow researchers to explore how school and student characteristics relate to achievement outcomes. As an example, consider the full mediation model (see, e.g., Finch, West, & MacKinnon, 1997) shown in Figure 1. While this model is merely illustrative, it is similar to those studied by substantive researchers (see, e.g., Meece, Eccles, & Wigfield, 1990). In the model, ANX is regressed on PSC. Further, TASK is regressed on both ANX and PSC. This ordinal structural model could be estimated by the multistage estimator, at which point a researcher would typically need to examine its fit to data.

Insert Figure 1 about here

3 A Structural Equation Model for Ordered Categorical Responses

3.1 The Data and the Model

Let there be $i = 1, \dots, N$ respondents and $j = 1, \dots, n$ variables. Let $\boldsymbol{y}_i^*$ be an $n \times 1$ vector of continuous underlying response variables. It is typically assumed that $\boldsymbol{y}_i^*$ is multivariate normal, that is, $\boldsymbol{y}_i^* \sim \mathcal{N}_n(\mathbf{0}, \mathbf{P})$ where $\mathbf{P}$ is an $n \times n$ correlation matrix. The $d_\rho = n(n-1)/2$ unique correlations are stacked and collected in the $d_\rho \times 1$ vector $\boldsymbol{\rho}$.

It is assumed that a $p \times 1$ vector of latent factors is related to $\boldsymbol{y}^*$ via a factor analytic measurement model. For the ith case, this may be represented as $\boldsymbol{y}_i^* = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i$. Further, the structural relationships among the latent variables is assumed to take the form $\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i$. In the above equations, the unique factors in $\boldsymbol{\epsilon}$ and the disturbance terms in $\boldsymbol{\zeta}$ have zero means. Their covariance matrices are $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$, respectively. Assuming that $\boldsymbol{\epsilon}$ and $\boldsymbol{\zeta}$ are orthogonal, the covariance structure for $\boldsymbol{y}^*$ is

$$cov(\boldsymbol{y}^*) = \boldsymbol{\Lambda}\mathbf{A}\boldsymbol{\Phi}\mathbf{A}'\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \qquad (1)$$

where $\mathbf{A} = (\mathbf{I}_p - \mathbf{B})^{-1}$ is invertible and $\mathbf{I}_p$ is a $p \times p$ identity matrix. To identify the model, it is generally necessary to set diag($\boldsymbol{\Psi}$) = diag($\mathbf{I}_p - \boldsymbol{\Lambda}\mathbf{A}\boldsymbol{\Phi}\mathbf{A}'\boldsymbol{\Lambda}'$). This identification condition implies that $cov(\boldsymbol{y}^*) = \mathbf{P}$ is a correlation matrix.

By the underlying response process formulation, the continuous $\boldsymbol{y}_i^*$ are not observed. Instead, the $n \times 1$ vector of observed categorical variables $\boldsymbol{y}_i$ result from the discretization of $\boldsymbol{y}_i^*$. To facilitate the presentation, we assume that all observed variable have the same number of categories, K. Then, for each variable, there are $K-1$ thresholds, $\tau_1, \dots, \tau_{K-1}$. In all, there are $d_\tau = n(K-1)$ thresholds, which can collected into a $d_\tau \times 1$ vector $\boldsymbol{\tau}$. Finally, y_{ij}^* and y_{ij} are related via the thresholds where

$$y_{ij} = k, \quad \text{if } \tau_{j,k} < y_{ij}^* < \tau_{j,k+1}, \tag{2}$$

with $\tau_{j,0} = -\infty,\ \tau_{j,K} = \infty$.

3.2 Multistage Estimation and Testing

Multistage estimation begins by obtaining an estimate of the (polychoric) correlations in $\boldsymbol{\rho}$. In practice, this is often accomplished in two steps. First, the thresholds are estimated by maximum likelihood, one item at a time, yielding $\hat{\boldsymbol{\tau}}$. Next, treating $\hat{\boldsymbol{\tau}}$ as fixed, the bivariate correlations are estimated by maximum likelihood, one pair of items at a time. This yields a vector of estimated polychoric correlations, $\hat{\boldsymbol{\rho}}$. To facilitate the presentation, we assume that no constraints are imposed on the thresholds. Then, the free structural parameters (e.g., factor loadings and latent regression coefficients) can be estimated by minimizing a weighted least squares (WLS) function of the polychoric correlation residuals. Formally, let the q free parameters be collected in the vector $\boldsymbol{\theta}$, and let $\boldsymbol{\rho}(\boldsymbol{\theta})$ represent the model-implied correlations. Then, the estimator $\hat{\boldsymbol{\theta}}$ is obtained by minimizing

$$F(\boldsymbol{\theta}; \mathbf{W}) = \left(\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}(\boldsymbol{\theta})\right)' \mathbf{W} \left(\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}(\boldsymbol{\theta})\right), \tag{3}$$

where $\mathbf{W}$ is a positive definite weight matrix.

Next, we consider the form of the weight matrix $\mathbf{W}$. Let $\hat{\mathbf{V}}$ be a consistent estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\rho}}$. Further, let $\hat{\mathbf{D}} = \text{diag}(\hat{\mathbf{V}})$ be a diagonal matrix. The most common choices for $\mathbf{W}$ in Equation (3) are as follows. Choosing $\mathbf{W} = \hat{\mathbf{V}}^{-1}$ results in the full weighted least squares estimator (WLS, Muthén, 1978). Choosing $\mathbf{W} = \hat{\mathbf{D}}^{-1}$ results in the diagonally weighted least squares estimator (DWLS, Muthén, du Toit, & Spisic, 1997). Finally, choosing $\mathbf{W} = \mathbf{I}$ results in the unweighted least squares estimator (ULS, Muthén, 1993). While theoretically important, WLS is not often used in practice as it tends to perform poorly unless N is very large. Under correct model specification and standard regularity conditions, the multistage estimator is $\sqrt{N}$-consistent and asymptotically normal (Jöreskog, 1994; Lee, Poon, & Bentler, 1995; Muthén & Satorra, 1995).

In this research, only ULS and DWLS are used to estimate ordinal structural models. Accordingly, let $\hat{\boldsymbol{\theta}}_U$ and $\hat{\boldsymbol{\theta}}_D$ be the vectors of parameter estimates obtained using ULS and DWLS, respectively. Similarly, let $\hat{F}_U$ and $\hat{F}_D$ be the respective minimized discrepancy function values. Such a discrepancy function value, $\hat{F}$, can be used to construct an overall GOF statistic, $T = N \times \hat{F}$. However, for ULS and DWLS, T is not chi-square distributed even under correct model specification (Browne, 1984). But, as suggested by Muthén (1993), moment corrections may be applied to T to construct a test statistic that is approximately chi-square distributed. These moment corrections are analogous to those used in the continuous data case (Satorra & Bentler, 1994). While several adjustments have been proposed, this research utilizes the correction of Asparouhov and Muthén (2010), which is denoted by $\tilde{T}$. An advantage of $\tilde{T}$ is that it scales T so that the resulting statistic is approximately chi-square distributed with the "natural" degrees of freedom (i.e., the difference between the numbers of parameters in the saturated and estimated models). The use of ULS and DWLS to calculate $\tilde{T}$ yields $\tilde{T}_U$ and $\tilde{T}_D$, respectively.

4 Limited-Information Testing Methodology

While test statistics based on quadratic forms in the correlational residuals in $\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}(\boldsymbol{\theta})$ have proven useful in evaluating the fit of ordinal structural models, these statistics were not specifically developed for categorical data and contingency tables. In a certain sense, these statistics may be regarded as afterthoughts, developed as the result of fitting categorical data into a factor-analytic framework largely dominated by continuous outcomes. On the other hand, recent years have seen a number of limited-information statistics specifically developed for latent variable models with categorical outcomes. Generally, these statistics are quadratic forms in linear functions of multinomial cell residuals from the n-way contingency tables formed by the cross-tabulations of the observed responses. Some examples are M_2 (Maydeu-Olivares & Joe, 2006), M_2^* (Cai & Hansen, 2013), and C_2 (Cai & Monroe, 2014). We have chosen to apply and study the C_2 statistic in this research, as it has theoretical and practical advantages over both M_2 and M_2^* (Cai & Monroe, 2014). The presentation here focuses on the application of C_2 to the ordinal structural model with multistage estimation. Readers interested in a more technical account of C_2, or its application to IRT models, are referred to Cai and Monroe (2014).

4.1 Full-Information and Limited-Information Test Statistics

Returning to the structure of the data, recall that K is the number of response categories per item. In total, there are $\kappa = K^n$ possible response patterns, which increases rapidly with K and/or n. For example, for the PISA model introduced in Section 2, $\kappa = 4^{12} > 16$ million. Let the $\kappa \times 1$ vector $\boldsymbol{p}$ collect the κ sample proportions. Similarly, let $\boldsymbol{\pi}(\boldsymbol{\theta})$ collect the κ model-implied response pattern probabilities. Then, let $\boldsymbol{e} = \boldsymbol{p} - \boldsymbol{\pi}(\boldsymbol{\theta})$ be the cell residuals. Assuming the model is correctly specified in the population and given a vector of true parameters $\boldsymbol{\theta}_0$, let the true model-implied probabilities be $\boldsymbol{\pi}_0 = \boldsymbol{\pi}(\boldsymbol{\theta}_0)$. In this case, the observed data may be considered to be a sample of size N from a multinomial with κ categories.

One approach to testing structural models for categorical data is to use a full-information test which directly uses the full set of multinomial residuals. Pearson's X^2 is one such test, and is defined as $X^2 = N \sum_{i=1}^{\kappa} \left[p_i - \pi_i(\widehat{\boldsymbol{\theta}})\right]^2 / \pi_i(\widehat{\boldsymbol{\theta}})$. When a fully-efficient estimator, such as maximum-likelihood, is used to obtain $\widehat{\boldsymbol{\theta}}$, and the model is correctly specified in the population, X^2 is approximately chi-square distributed with $\kappa - q - 1$ degrees of freedom. Despite this asymptotic result, X^2 is not generally useful for testing structural models for categorical data, for several reasons. First, for large values of κ, some model-implied probabilities must necessarily be near-zero. In the literature, this is often referred to as *sparseness* of the contingency table. Under sparseness, the Type I error rates and power of X^2 are both adversely affected (e.g., Bartholomew and Leung, 2002). An accompanying problem is computational. For large K and/or n, κ may be so large that calculating X^2 becomes computationally impractical. Recall that $\kappa > 16$ million for the PISA model, with only 12 variables. Finally, in fitting structural models to categorical data, estimators that are not fully-efficient, such as the multistage estimator, are frequently used. In this case, X^2 will not follow its nominal chi-square distribution with $\kappa - q - 1$ degrees of freedom.

Another, more appealing, approach is provided by limited-information tests. Generally, these tests are quadratic forms that depend on lower-order sample proportions and model-implied probabilities. Different limited-information tests can be distinguished by: 1) which lower-order proportions and probabilities are used; 2) how the proportions and probabilities are combined; and 3) how the distribution of the test is approximated. Here, we focus on first and second-order proportions and probabilities to summarize the categorical data, which is akin to using means and covariances to summarize continuous data.

For a single variable, there are only $K - 1$ independent probabilities as the K cells must sum to 1. Conveniently, a set of independent cells can be obtained by removing any cell with category code $k = 0$. Then, let $\dot{\boldsymbol{p}}$ and $\dot{\boldsymbol{\pi}}(\boldsymbol{\theta})$ be the vectors of length $s_1 = n(K - 1) = d_\tau$, consisting of all linearly independent first-order marginal probabilities and proportions, respectively. Let $\dot{\boldsymbol{e}} = \dot{\boldsymbol{p}} - \dot{\boldsymbol{\pi}}(\boldsymbol{\theta})$ be the vector of linearly independent first-order residual probabilities.

For a pair of variables, there are $(K - 1)^2$ independent second-order marginal proportions or model-implied probabilities upon knowing the first-order margins. Again, an independent set may be obtained by removing any cell in the $K \times K$ two-way table where either category code is 0. Then, let $\ddot{\boldsymbol{p}}$ and $\ddot{\boldsymbol{\pi}}(\boldsymbol{\theta})$ be the vectors of length $s_2 = n(n - 1)/2 \times (K - 1)^2 = d_\rho(K - 1)^2$ of all linearly independent second-order proportions and model-implied probabilities, respectively. And, let $\ddot{\boldsymbol{e}} = \ddot{\boldsymbol{p}} - \ddot{\boldsymbol{\pi}}(\boldsymbol{\theta})$ be the vector of all linearly independent second-order residual probabilities.

With these definitions, we now explain how limited-information tests may be more easily applied than full-information tests. While first and second-order sub-tables can still be affected by sparseness, these tables are necessarily better-filled than the entire n-way contingency table with κ cells. Consequently, limited-information tests are less vulnerable to the sparseness issue that affects the utility of full-information tests. Additionally, limited-information tests are potentially less computationally burdensome than full-information tests. For example, the number of first and second-order probabilities (s_1 and s_2, respectively) may be much smaller than κ. For the PISA model, $s_1 = 36$ and $s_2 = 594$, while $\kappa > 16$ million. Finally, limited-information tests do not require a fully-efficient estimator. Instead, they only require consistency and asymptotic normality (Maydeu-Olivares and Joe, 2006), which are properties enjoyed by numerous estimators for structural models of categorical data, including the multistage, pairwise likelihood (Katsikatsou, Moustaki, Yang-Wallentin, & Jöreskog, 2012), and polychoric instrumental variable (Bollen & Maydeu-Olivares, 2007) estimators.

4.2 Three Limited-Information Test Statistics

The limited-information test of Maydeu-Olivares (2006) is noteworthy due to its application to structural models of categorical data. For convenience, let $\ddot{M}$ denote this statistic. $\ddot{M}$ is an unweighted sum of squares of the second-order residual probabilities in $\ddot{\boldsymbol{e}}$. The distribution of $\ddot{M}$ can be approximated by moment-matching (Satorra & Bentler, 1994).

The M_2 statistic (Maydeu-Olivares & Joe, 2006) is noteworthy here for at least two reasons. First, M_2 and C_2 have analogous structures, which will be presented below. Second, M_2 has been widely-applied in IRT modeling and is available in commercial IRT software (e.g., flexMIRT®, Cai, 2013). Like $\ddot{M}$, M_2 uses the second-order residual probabilities in $\ddot{\boldsymbol{e}}$, but it also incorporates the first-order residual probabilities in $\dot{\boldsymbol{e}}$. Let $\boldsymbol{e}_2 = (\dot{\boldsymbol{e}}', \ddot{\boldsymbol{e}}')'$ be the vector of length $s = s_1 + s_2$ that collects all linearly independent first and second-order residual probabilities. Then, M_2 can be defined as

$$M_2 = N\hat{\boldsymbol{e}}_2{}'\widehat{\boldsymbol{\Omega}}_2\hat{\boldsymbol{e}}_2, \tag{4}$$

where

$$\boldsymbol{\Omega}_2 = \boldsymbol{\Xi}_2^{-1} - \boldsymbol{\Xi}_2^{-1}\boldsymbol{\Delta}_2(\boldsymbol{\Delta}_2'\boldsymbol{\Xi}_2^{-1}\boldsymbol{\Delta}_2)^{-1}\boldsymbol{\Delta}_2'\boldsymbol{\Xi}_2^{-1}, \tag{5}$$

and all matrices are evaluated at $\widehat{\boldsymbol{\theta}}$. In Equation (5), $\boldsymbol{\Xi}_2$ is the asymptotic covariance matrix of the first and second-order sample proportions, and $\boldsymbol{\Delta}_2$ is the matrix of derivatives of the first and second-order model-implied probabilities with respect to the vector of parameter estimates, $\widehat{\boldsymbol{\theta}}$. In words, M_2 is a quadratic form in the first and second-order residual probabilities. The matrix of the quadratic form, $\boldsymbol{\Omega}_2$, weights these residual probabilities so that M_2 is asymptotically chi-square distributed with $s - q$ degrees of freedom (Maydeu-Olivares & Joe, 2006).

While $\ddot{M}$ and M_2 are more robust to sparseness than full-information statistics, they can still be affected by the issue when the number of variable categories is large. As explained by Cai and Hansen (2012), this is because for some pairs of variables, certain response combinations are highly unlikely. For example, with the PISA survey, a student is unlikely to respond "strongly agree" to the item, "I learn mathematics quickly," while also responding "strongly disagree" to the item, "In my mathematics class, I understand even the most difficult work." As shown in Cai and Hansen (2012), this sparseness in the $K \times K$ two-way table can negatively impact the Type I error rates and power of M_2. Additionally, when both K and the number of variables are relatively large (i.e., when s_2 is very large), it can become computationally burdensome to calculate, store, and manipulate all of the second-order residual probabilities in $\ddot{\boldsymbol{e}}$, the derivatives, and the even larger number of elements in the weight matrix.

C_2 addresses these issues by collapsing each $K \times K$ two-way table of residuals into a single residual moment. This is facilitated by using the ordered category codes $k = 0, \ldots, K-1$, as the raw scores. Let $\ddot{e}_{l,m,k_l,k_m}$ be the second-order marginal residual probability for variables l and m in categories k_l and k_m, respectively. The residual moment for variables l and m is given by the weighted sum

$$\ddot{r}_{l,m} = \sum_{k_l=1}^{K-1} \sum_{k_m=1}^{K-1} k_l k_m \ddot{e}_{l,m,k_l,k_m} \,. \tag{6}$$

In words, $\ddot{r}_{l,m}$ sums all of the second-order residual probabilities for variables l and m, weighted by the product of the two corresponding category codes. These second-order marginal residual moments can be collected into a vector $\ddot{\boldsymbol{r}} = (\ddot{r}_{2,1}, \ddot{r}_{3,1}, \ldots, \ddot{r}_{n,n-1})'$ of dimension $s_2^* = n(n-1)/2 = d_\rho$. Then, let the vector $\boldsymbol{r}_2 = (\dot{\boldsymbol{e}}', \ddot{\boldsymbol{r}}')'$, with dimension $d = s_1 + s_2^*$, collect all of the linearly independent first-order marginal residual probabilities as well as the collapsed second-order marginal residual moments.

Then, C_2 is a quadratic form in $\boldsymbol{r}_2$, defined as

$$C_2 = N\hat{\boldsymbol{r}}_2{}'\widehat{\mathbf{U}}_2\hat{\boldsymbol{r}}_2, \tag{7}$$

where

$$\mathbf{U}_2 = \boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_2^{-1}\mathbf{J}_2(\mathbf{J}_2'\boldsymbol{\Sigma}_2^{-1}\mathbf{J}_2)^{-1}\mathbf{J}_2'\boldsymbol{\Sigma}_2^{-1}, \tag{8}$$

and all matrices are evaluated at $\widehat{\boldsymbol{\theta}}$. The construction of C_2 parallels that of M_2, with $\hat{\boldsymbol{r}}_2$ replacing $\hat{\boldsymbol{e}}_2$, and corresponding changes made in the weight matrix $\mathbf{U}_2$. That is, in Equation (8), $\boldsymbol{\Sigma}_2$ is the asymptotic covariance matrix of the first and collapsed second-order sample proportions, and $\mathbf{J}_2$ is the matrix of derivatives of the first-order and collapsed second-order model-implied probabilities with respect to the vector of parameter estimates, $\widehat{\boldsymbol{\theta}}$. The matrix of the quadratic form, $\mathbf{U}_2$, weights the residual probabilities and moments so that C_2 is asymptotically chi-square distributed with $d - q$ degrees of freedom (Cai & Monroe, 2014).

4.3 Technical Details for C_2

A derivation of C_2, and its application to IRT, is given in Cai and Monroe (2014). We refer interested readers to that report. However, the application of C_2 to structural models of categorical data in this research necessitates the presentation of certain technical topics, which are contained in the Appendix.

These topics include: 1) satisfaction of regularity conditions by the multistage estimator; 2) calculation of model-implied probabilities; and 3) calculation of the derivatives of the first and second-order model-implied probabilities with respect to the vector of parameter estimates.

5 Simulation Study for C_2

A simulation study was conducted to compare the C_2 statistic with the traditional $\tilde{T}_U$ and $\tilde{T}_D$ statistics in terms of Type I error rates and power. The sample sizes considered were $N = 100$, 200, 500, and 1000. The form of the generating structural model was identical to the theorized mediation model presented in Figure 1. Referring to the notation presented earlier, the latent variables PSC, ANX, and TASK can be considered η_1, η_2, and η_3 respectively. The true structural parameters in **B** were $\beta_{21} = 0.3$, $\beta_{31} = 0.4$, and $\beta_{32} = 0.36$, values used in Finch et al. (1997).

5.1 Design: Data Generation

For the null condition, a population correlation matrix, $\mathbf{P}_0$, was calculated via Equation (1), using the factor loadings and unique variances shown in Table 2. For each of 500 replications, $\boldsymbol{y}_i^* \sim \mathcal{N}_n(\mathbf{0}, \mathbf{P}_0)$ were sampled to form a dataset of continuous underlying variables. Let $\mathbf{Y}^*$ be this dataset. Then, $\mathbf{Y}^*$ was discretized to yield three categorical datasets, $\mathbf{Y}^{(K)}$, for $K = 2$, 4, and 6. For a given replication, the categorical datasets are "nested" in the following sense. First, $\mathbf{Y}^*$ was discretized using 5 thresholds per variable to yield $\mathbf{Y}^{(6)}$**.** Next, a random subset of the thresholds, fixed over replications, was used to create $\mathbf{Y}^{(4)}$. Finally, a further random subset of the thresholds, fixed over replications, was used to create $\mathbf{Y}^{(2)}$. The thresholds and subsets are presented in Table 2.

Insert Table 2 about here

To study the power of C_2, we used the steps just detailed, but introduced model error when generating the population correlation matrices. Specifically, structural model error was introduced using a variation of the Cudeck and Browne (1992) procedure. Given a choice of discrepancy function, the Cudeck and Browne (1992) procedure produces a correlation matrix with a prespecified discrepancy function value. To be consistent with the choice of estimator for the simulated categorical datasets, we chose the ordinary least squares discrepancy function. And, in a slight variation of the original procedure, we specified an exact population RMSEA value instead of the discrepancy function value as the former is more familiar. Let ε_0^* be this value, where the asterisk emphasizes that the definition is at the level of the continuous underlying response variables, $\boldsymbol{y}^*$. The chosen values for ε_0^* were .01, .05, and .10. For continuous normally distributed outcomes, these values are often considered cutoffs for "excellent," "close," and "mediocre" fit, respectively (see, e.g., Browne & Cudeck, 1993), though alternative cutoff values exist (e.g., Hu & Bentler, 1999). An example population correlation matrix for the $\varepsilon_0^* = .10$ model is shown in Table 3.

Insert Table 3 about here

5.2 Design: Estimation and Collected Statistics

For each simulated data set, the mediation model shown in Figure 1 was estimated twice in Mplus (Muthén & Muthén, 2010), once with ULS and once with DWLS. These two model fittings yielded $\tilde{T}_U$ and $\tilde{T}_D$, respectively. The ULS parameter estimates were then used along with the replication's dataset to obtain the C_2 statistic. To the extent that the ULS and DWLS point estimates differ, the resulting C_2 values will also differ. However, we found this difference to be negligible and choose to report only the ULS-based C_2.

Solutions were checked to see if they were proper. Solutions were deemed improper if the estimated error variance was negative for any variable. These replications were discarded and not included in the results. Collected statistics include the proportion of properly converged replications and rejection rates at common alpha levels. For all test statistics, the empirical mean and variance were recorded. Also, for the null condition, two-sided Kolmogorov-Smirnov (K-S) tests were conducted.

After collecting and examining the results, it became clear that the results for DWLS and $\tilde{T}_D$ were very similar to those for ULS and $\tilde{T}_U$. Thus, we only report the latter results.

5.3 Results: Null Condition

Table 4 presents the results for the null condition of the simulation study. As expected, the proportion of valid replications increases with N and K. For instance, whereas the proportion of valid replications for $N = 100$ and $K = 2$ is 0.71, for $N = 1000$ and $K = 4$, all replications converged properly. Generally, the calibration of the statistics also improves with increases in N and K. For the $N = 100$ and $K = 2$ condition, neither statistic is well-calibrated, as measured by the K-S p-values. This conclusion is supported by the Type I error rates, which differ substantially from the nominal values. We can also compare the empirical means and variances of C_2 and $\tilde{T}_U$ to the mean (df) and variance $(2df)$ of the reference chi-square $(df = 51)$. For this condition, the empirical distributions of C_2 and $\tilde{T}_U$ appear stochastically smaller than the reference. On the other hand, for the largest sample size $(N = 1000)$ and $K = 6$ condition, both statistics appear well-calibrated, as evidenced by the Type I error rates and K-S p-values.

Insert Table 4 about here

Examining Table 4 more closely, C_2 appears to be better-calibrated than $\tilde{T}_U$ at smaller sample sizes or with smaller K. At $N = 100$, C_2 appears reasonably well-calibrated for both $K = 4$ and $K = 6$, as evidenced by the non-significant p-values (.099 and .169, respectively) and Type I error rates that approximately track the nominal levels. In contrast, $\tilde{T}_U$ has significant p-values for these conditions (.016 and .003, respectively). Turning to $K = 2$, at $N = 200$, C_2 again appears better-calibrated than $\tilde{T}_U$, as the latter statistic clearly under-rejects the null hypothesis.

In summary, there are conditions, particularly with small N or small K, where C_2 is well-calibrated, while $\tilde{T}_U$ is not. However, there are no conditions where $\tilde{T}_U$ is well-calibrated, while C_2 is not. Thus, C_2 appears to be slightly better calibrated than $\tilde{T}_U$.

5.4 Results: Power

Table 5 presents empirical rejection rates at the $\alpha = .05$ level when model error is introduced via ε_0^*. The cells shaded in gray correspond to conditions under the null where the K-S p-values were significant. Since the significant p-values suggest the statistic may not be well-calibrated, care should be taken in interpreting these rejection rates. If we limit our evaluation to the non-shaded cells, then it is clear that C_2 is generally more powerful than $\tilde{T}_U$. In many cases, the difference in power is quite small. And, at the highest values of ε_0^* and N, both statistics have power at or near 1.0 and cannot be distinguished. However, in other cases, such as $\varepsilon_0^* = .05$, $N = 500$, and $K = 4$, the difference in rejection rates is substantial (.820 and .570, for C_2 and $\tilde{T}_U$, respectively). Also, because C_2 appears generally better-calibrated than $\tilde{T}_U$, there are conditions where the rejection rate for C_2 may be the only meaningful result. Based on Table 5, C_2 has more power than $\tilde{T}_U$ in detecting the model error introduced via the Cudeck and Browne (1992) procedure.

<u>Insert Table 5 about here</u>

As mentioned earlier, in practice, with a sufficiently large sample size and any amount of model error, the proposed model will be rejected by an overall test, such as $\tilde{T}_U$ or C_2. In this event, practitioners routinely examine fit indices, such as RMSEA, to assess the approximate fit of the model. Given our simulation procedure, one RMSEA, which is based on $\tilde{T}_U$, may be obtained using the Mplus output. However, an alternative RMSEA, based on C_2, may also be calculated. In the next section, we compare these two RMSEA estimates, and investigate how they are affected by the number of variable categories.

6 The Relationship Between RMSEA and Number of Categories

This Section uses the simulation results of Section 5 to study RMSEA for structural models of categorical data. To study power in Section 5, structural model error was introduced with a specified population RMSEA value, denoted by ε_0^*. Again, the chosen values for ε_0^* were .01, .05, and .10. Let $\hat{\varepsilon}^{(K)}$ be the sample RMSEA estimate for $\mathbf{Y}^{(K)}$, where $\hat{\varepsilon}^{(K)}$ may be based on either $\tilde{T}_U$ or C_2. Then, the interpretation of RMSEA for categorical data may be studied in two ways. First, for a given simulation condition, the $\hat{\varepsilon}^{(K)}$ may be averaged over the 500 replications for the sampling error to become negligible. Let $\bar{\varepsilon}^{(K)}$ be such an average. Then, $\bar{\varepsilon}^{(K)}$ may be directly compared to ε_0^*, with discrepancies suggesting that the population RMSEA values for the continuous underlying response variables and the discretized categorical variables are not the same. Second, for each $\mathbf{Y}^*$, the $\hat{\varepsilon}^{(K)}$ values for the nested datasets may be compared to one another. Any systematic relationship that holds *across* the Monte Carlo replications would also be of interest.

The RMSEA estimate $\hat{\varepsilon}^{(K)}$ was obtained by

$$\hat{\varepsilon}^{(K)} = \sqrt{\max\left(\frac{T - df}{N \times df}, 0\right)} \tag{9}$$

where T is either C_2 or $\tilde{T}_U$, and df is the corresponding degrees of freedom. For each of the 500 replications, the mean RMSEA values and empirical 5th and 95th percentiles were recorded.

Figure 2 displays the means and empirical 90% confidence intervals for selected simulation conditions. Results corresponding to the $N = 100$ sample size have been omitted, as they are quite similar to the $N = 200$ sample size. A number of trends in Figure 2 are noteworthy. Overall, $\bar{\varepsilon}^{(K)}$ based on C_2 is greater than the corresponding $\bar{\varepsilon}^{(K)}$ based on $\tilde{T}_U$. This is expected, as C_2 is generally the more powerful statistic. Also, as expected, the sampling variability of $\hat{\varepsilon}^{(K)}$ decreases for larger N, as evidenced by the shorter line segments spanning the 90% confidence intervals. Note, however, that for any given ε_0^* and K, the $\bar{\varepsilon}^{(K)}$ values are relatively stable across the various sample sizes.

<u>Insert Figure 2 about here</u>

For the $\varepsilon_0^* = .01$ conditions (the top row of plots in Figure 2), the $\bar{\varepsilon}^{(K)}$ values do not appear to depend on K. Further, all of the $\bar{\varepsilon}^{(K)}$ estimates are near $\varepsilon_0^* = .01$, and for all N and K, the 90% empirical confidence interval of $\hat{\varepsilon}^{(K)}$ spans ε_0^*. For the $\varepsilon_0^* = .05$ conditions (the middle row of plots in Figure 2), the pattern of results is quite different. There is a clear dependence on K, with $\bar{\varepsilon}^{(K)}$ increasing in K. Also, for all N and K, $\bar{\varepsilon}^{(K)} < \varepsilon_0^* = .05$. And, for the largest sample size, the 90% empirical confidence intervals of $\hat{\varepsilon}^{(K)}$ do not span ε_0^*. Finally, the pattern of results for the $\varepsilon_0^* = .10$ conditions (the bottom row of plots in Figure 2) is quite similar to that of the

$\varepsilon_0^* = .05$ conditions. Again, $\bar{\varepsilon}^{(K)}$ clearly increases with K, and is always less than $\varepsilon_0^* = .10$ for the studied conditions.

Figure 3 presents results from another perspective, focusing on the "nested" nature of the datasets for the $N = 1000$ and $\varepsilon_0^* = .10$ condition. That is, Figure 3 gives a more detailed look at the results corresponding to the lower-right plot in Figure 2. For each replication, there is a $\hat{\varepsilon}^{(K)}$ value for $K = 2$, 4, and 6. Further, an RMSEA estimate can be computed upon fitting the structural model to the underlying continuous data for the replication because we have access to them in a simulation. Denote this estimate as $\hat{\varepsilon}^*$. Figure 3 shows the relationship among these various RMSEA estimates (based on C_2 for the categorical data and ordinary least squares for the continuous underlying response data).

Insert Figure 3 about here

For this condition, from Figure 2, we know that $\bar{\varepsilon}^{(K)}$ increases with K. However, Figure 3 makes clear that, for this condition, the RMSEA estimates for "nested" datasets are positively correlated. An implication of Figure 3 is that for a dataset from this condition, any decrease in the number of categories will likely result in a smaller RMSEA estimate. For other conditions, though, the various RMSEA estimates may be more weakly correlated. Factors that influence the strength of the relationships include the magnitudes of N (since a smaller N leads to increased sampling variability) and ε_0^* (since RMSEA is bounded below by 0). Finally, Figure 3 illustrates that with the continuous underlying variables (y-axes for top row of plots), the $\hat{\varepsilon}^*$ values estimate ε_0^* with little bias because the distribution appears to center on the true RMSEA value. In this case, the empirical mean is .099, very close to $\varepsilon_0^* = .10$.

From Figures 2 and 3, it is clear that $\bar{\varepsilon}^{(K)}$ is a poor estimate of ε_0^*. As one extreme example, consider the $K = 2$ and $N = 1{,}000$ condition, when $\varepsilon_0^* = .10$. In this case, $\bar{\varepsilon}^{(2)} = .034$ for C_2 and $.027$ for $\tilde{T}_U$. Based on these large discrepancies, we reason that $\bar{\varepsilon}^{(K)}$ is approximating a *different* population value, due to the discretization process. Let $\varepsilon_0^{(K)}$ be such a value. To the extent that $\bar{\varepsilon}^{(K)}$ is a reasonable estimate for $\varepsilon_0^{(K)}$, it is clear that $\varepsilon_0^{(K)} \neq \varepsilon_0^*$. Also, for relatively large values of ε_0^* (e.g., .05 or .10), $\varepsilon_0^{(K)}$ is always less than ε_0^*. Further, for such conditions, $\varepsilon_0^{(K)}$

appears to converge towards ε_0^* as K increases, though the convergence is slow. Greater values of K (e.g., 10) would be helpful in exploring this apparent convergence. However, such high values are not common in empirical data and were not included in the simulation. In any case, Figures 2 and 3 suggest that the guidelines developed for RMSEA interpretation using continuous data may not be applicable for use with categorical data.

Also, $\bar{\varepsilon}^{(K)}$, and presumably $\varepsilon_0^{(K)}$, clearly depends on the underlying test statistic, C_2 or $\tilde{T}_U$. For the studied conditions, $\bar{\varepsilon}^{(K)}$ based on C_2 is a less biased estimate of ε_0^*. In other words, the C_2-based RMSEA for the categorical datasets is generally a better estimate of the population RMSEA defined at the level of the continuous data. In summary, even when the population RMSEA for the continuous underlying response variables is fixed, the estimated value of RMSEA for categorical variables depends on a number of things, including the discrepancy function, number of categories per variable, and the choice of underlying test statistic.

7 Empirical Application

In this section, we apply C_2 to the PISA example presented in Section 2. We also calculate the RMSEA estimates and discuss their interpretation in light of the simulation study results. Only a random subset ($N = 1000$ complete cases) of the United States school sample is used. For this illustration, we ignore the complex sampling design of the survey, though it would need to be modeled for proper inference. As opposed to the goal of producing valid substantive findings, our goals here are to demonstrate the utility of C_2 in assessing a structural model of real data, and highlight the challenges in interpreting RMSEA for such models.

<u>Insert Table 6 about here</u>

The model was fitted twice in Mplus, once using ULS and once with DWLS. The overall model fit statistics and select fit indices are presented in Table 6. For all of the test statistics (i.e., the ULS-based C_2, $\tilde{T}_U$ and $\tilde{T}_D$), $p < .001$. The large sample size ($N = 1000$) may be an issue in the use of the chi-square test statistics. Turning to the RMSEA estimates, the C_2-based estimate (.036) is less than either the $\tilde{T}_U$ or $\tilde{T}_D$-based estimates (.041 and .054). This is not inconsistent with the simulation study results, where the C_2-based RMSEA estimates were only greater than the $\tilde{T}_U$ or $\tilde{T}_D$-based RMSEA estimates on average and certainly can be smaller on occasion. Also, it is possible that $\tilde{T}_U$ and $\tilde{T}_D$ are more powerful than C_2 against certain types of model error. Applying conventional guidelines for RMSEA interpretation, the observed estimates are all near the .05 cut-off of "close-fit." In particular, the upper-bound of the 90% confidence interval for the C_2-based estimate is .044, which lends further support to the position that the theorized mediation model is close-fitting. However, the results from the simulation study suggest that guidelines developed for use with continuous data may be less applicable for categorical data. More specifically, for models with $K = 4$, the conventional guidelines may be too lenient. Examining Figure 2 again, we may have reason to believe that in the categorical case, with $K = 4$, the RMSEA estimates are smaller by about 20-30% than in the continuous case. Consequently, at least for C_2, perhaps the cut-off between "close" and "not close" should be around .03 as opposed to .05. This, however, is merely a conjecture as opposed to any sort of suggested guideline.

8 Discussion and Conclusion

In this research, limited-information testing principles, heretofore primarily applied in the context of IRT, were applied to SEM of ordinal data. Specifically, the C_2 statistic proposed in Cai and Monroe (2014) was compared to test statistics based on quadratic forms in polychoric correlation residuals. C_2 was shown to perform at least as well as the competing statistics in terms of calibration under the null as well as power. For some conditions, C_2 clearly outperformed the other statistics. This research also took the opportunity presented by the simulation study to examine the behavior of the RMSEA fit index under varying conditions. While guidelines for RMSEA interpretation of continuous variables have been developed over many years, the use of RMSEA for assessing fit of categorical variables is a much more recent phenomenon. The simulation results suggest that the magnitude of RMSEA estimates is surprisingly dependent on the number of variable categories.

While we believe this research has contributed to the area of model fit assessment for categorical SEM, it has also left many questions unanswered. Regarding the C_2 statistic, it is unknown how C_2 will perform under other conditions. Notably, C_2 should be studied with larger models, as the simulation study in this research focused on a relatively small model (with only 12 variables). Also, it would be interesting to study C_2 when the underlying continuous variables are not normal. Presumably, C_2 would have more power to detect this sort of misspecification than statistics that assume multivariate normality of the underlying response variables. Additionally, the statistic itself can be further developed for structural models for categorical data. Under multistage estimation, the sample proportions can be perfectly reproduced by the threshold estimates, leading to all first-order residual probabilities being equal to zero. In this case, perhaps C_2, and other limited-information statistics, can be simplified.

As for the interpretation of RMSEA for categorical data, a number of questions deserve further study. Again, since the simulation study only used one model size, it is unclear to what extent model size will impact the behavior of RMSEA. Additionally, while the Browne and Cudeck (1992) procedure proved convenient in this research as a method of introducing model error, other forms of model misspecification (e.g., omitted cross-loadings) could elicit different behaviors of RMSEA. Also, given how RMSEA appears to depend on the number of categories in the outcome variables, to what extent can corrections or adjustments to RMSEA make the fit index easier to interpret or more useful? Finally, RMSEA is but one fit index. It would stand to reason that other statistics based on chi-square approximations (e.g., TLI) may exhibit interesting behaviors. In any case, both the current research and potential future research topics reinforce the notion that practitioners should exercise caution in interpreting fit index values (see, e.g., Marsh, Hau, & Wen, 2004). In closing, while this research has contributed to the understanding of model fit assessment for categorical data, much work remains.

Appendix

Regularity Conditions for the Multistage Estimator

Maydeu-Olivares and Joe (2006) assumed regularity conditions on the model that must be satisfied for application of the limited-information testing methodology. There must be a matrix $\mathbf{H}$ such that

$$\sqrt{\mathrm{N}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \overset{a}{=} \mathbf{H}\sqrt{\mathrm{N}}(\boldsymbol{p} - \boldsymbol{\pi}), \tag{10}$$

where "$\overset{a}{=}$" denotes asymptotic equivalence. Maydeu-Olivares and Joe (2006) presented $\mathbf{H}$ for the maximum likelihood estimator. Here, $\mathbf{H}$ is presented for the multistage estimator. Essentially, the approach taken here is to piece together results from Maydeu-Olivares (2006), which also considers asymptotic properties of the multistage estimator.

Let $\widetilde{\boldsymbol{\Delta}} = \partial\boldsymbol{\gamma}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$ be a $d \times q$ matrix. Recall that $\mathbf{W}$ is the $d \times d$ matrix used in the third stage of estimation. Then, let $\mathbf{M} = \left(\widetilde{\boldsymbol{\Delta}}'\mathbf{W}\widetilde{\boldsymbol{\Delta}}\right)^{-1}\widetilde{\boldsymbol{\Delta}}'\mathbf{W}$ be a $q \times d$ matrix. The estimates of the structural parameters may be expressed as a linear function of the estimates from the first and second stages,

$$\sqrt{\mathrm{N}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \overset{a}{=} \mathbf{M}\sqrt{\mathrm{N}}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}), \tag{11}$$

which is Equation (18) in Maydeu-Olivares (2006). The $d \times s_2$ matrix $\mathbf{G}$, defined in Equation (14) of Maydeu-Olivares (2006), is used to account for the first and second stages of estimation. Then, the estimates of the structural parameters may be expressed as a linear function of the underlying sample proportions and probabilities,

$$\sqrt{\mathrm{N}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \overset{a}{=} \mathbf{MG\ddot{L}}\sqrt{\mathrm{N}}(\boldsymbol{p} - \boldsymbol{\pi}), \tag{12}$$

where $\ddot{\mathbf{L}}$ is an $s_2 \times \kappa$ operator matrix (see, e.g., Cai and Hansen, 2013) such that $\ddot{\mathbf{e}} = \ddot{\mathbf{L}}\boldsymbol{e}$. Taking $\mathbf{H} = \mathbf{MG\ddot{L}}$ satisfies the requirements for the multistage estimator.

Model-Implied Probabilities

Calculation of $\boldsymbol{r}_2$ requires first and second-order model-implied probabilities. The covariance matrix in Equation (8), $\boldsymbol{\Sigma}_2$, requires first, second, third, and fourth-order model-implied probabilities. Details of the pattern of model-implied probabilities necessary for $\boldsymbol{\Sigma}_2$ can be found in Cai and Hansen (2013). According to the model, we can find the marginal probability of any subset of ν variables as

$$\Pr\left[\bigcap_{j=1}^{\nu} y_j = k_j\right] = \int \underset{\mathbb{P}}{\cdots} \int \phi_\nu(\boldsymbol{y}^*; \mathbf{0}, \mathbf{P}_\nu) d\mathbf{y}^* \tag{13}$$

where $\phi(\cdot)$ denotes a ν-variate normal density and $\mathbb{P}$ is a ν-dimensional parallelepiped region of integration given by $\mathbb{P} = \otimes_{j=1}^{\nu} (\tau_{j,k_j}, \ \tau_{j,k_j+1})$. The correlation matrix $\mathbf{P}_\nu$ is the $\nu \times \nu$ sub-matrix from $\mathbf{P}$. The regions of integration obviously depend on the thresholds $\hat{\boldsymbol{\tau}}$, and the correlations between the underlying variables depend on other free parameters of $\widehat{\boldsymbol{\theta}}$, according to Equation (1). If $\nu = n$, Equation (13) provides the marginal probability of an entire response pattern. And for $\nu < n$, Equation (13) can be used with any subset of the items to find marginal probabilities of any order as needed. For this research, we calculated Equation (13) for up to fourth-order probabilities using the Monte Carlo approach presented in Genz (1992). Though observed proportions could be substituted for the probabilities, these would likely prove unstable, in particular for smaller sample sizes.

Derivatives of the First and Second-Order Model-Implied Probabilities

The weight matrix of C_2 in Equation (8), $\mathbf{U}_2$, depends on $\mathbf{J}_2$. Instead of focusing on the elements of $\mathbf{J}_2$, it is sufficient to focus on the elements of $\boldsymbol{\Delta}_2$, as $\mathbf{J}_2 = \mathbf{T}\boldsymbol{\Delta}_2$ for an appropriate operator matrix $\mathbf{T}$. $\boldsymbol{\Delta}_2$ is the matrix of derivatives of first and second-order model-implied probabilities with respect to $\boldsymbol{\theta}$. Without loss of generality of the method, we make two simplifying assumptions for ease of exposition. Namely, we assume that there are no additional constraints placed on the free parameters, and that the thresholds are saturated, i.e., the model contains as many location parameters as there are thresholds. Following our notational convention, $\boldsymbol{\pi}_2(\widehat{\boldsymbol{\theta}}) = (\dot{\boldsymbol{\pi}}(\widehat{\boldsymbol{\theta}})', \ddot{\boldsymbol{\pi}}(\widehat{\boldsymbol{\theta}})')'$. It is also convenient to partition the components of $\boldsymbol{\theta}$ in the following way. Again, assuming saturated thresholds, let $\boldsymbol{\theta}_{\boldsymbol{\tau}}$ be those parameters that model $\hat{\boldsymbol{\tau}}$, and let $\boldsymbol{\theta}_{\boldsymbol{\rho}}$ be those parameters that model $\widehat{\boldsymbol{\rho}}$ (free parameters in $\boldsymbol{\Lambda}$, $\mathbf{B}$, etc.). Then, $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\boldsymbol{\tau}}', \boldsymbol{\theta}_{\boldsymbol{\rho}}')'$, and $\widehat{\boldsymbol{\Delta}}_2$ may be partitioned as

$$\widehat{\boldsymbol{\Delta}}_2 = \begin{bmatrix} \dfrac{\partial \dot{\boldsymbol{\pi}}(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{\boldsymbol{\tau}}} & \dfrac{\partial \dot{\boldsymbol{\pi}}(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{\boldsymbol{\rho}}} \\ \dfrac{\partial \ddot{\boldsymbol{\pi}}(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{\boldsymbol{\tau}}} & \dfrac{\partial \ddot{\boldsymbol{\pi}}(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{\boldsymbol{\rho}}} \end{bmatrix}. \tag{14}$$

As the first-order moments do not depend on correlations, the upper-right block of $\widehat{\boldsymbol{\Delta}}_2$ is $\mathbf{0}$.

Maydeu-Olivares (2006, Appendix 2) presents results for the upper-left and lower-left blocks of $\widehat{\boldsymbol{\Delta}}_2$. In the same Appendix 2, results are given for $\partial\ddot{\boldsymbol{\pi}}(\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\rho}$. By the chain rule, the lower-right block may be obtained as the product of $\partial\ddot{\boldsymbol{\pi}}(\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\rho}$ and $\partial\widehat{\boldsymbol{\rho}}/\partial\boldsymbol{\theta}_{\boldsymbol{\rho}}$. Thus, the elements of $\partial\widehat{\boldsymbol{\rho}}/\partial\boldsymbol{\theta}_{\boldsymbol{\rho}}$ are needed, which are standard results in the SEM literature (Bock & Bargmann, 1966).

References

Organisation for Economic Co-operation and Development. (OECD). *PISA 2003: Technical report*. Paris, France: OECD Publications; 2005.

Asparouhov, T., & Muthén, B. O. (2010). *Simple second order chi-square correction*. (Unpublished Technical Report). Los Angeles, CA: Muthén & Muthén.

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology*, *55*, 1-15.

Bartholomew, D. J., & Tzamourani, P. (1999). The goodness-of-fit of latent trait models in attitude measurement. *Sociological Methods and Research*, *27*, 525–546.

Bock, R. D., & Bargmann, R. E. (1966). Analysis of covariance structures. *Psychometrika, 31,* 507-534.

Bollen, K. A., & Maydeu-Olivares, A. (2007). A polychoric instrumental variable (PIV) estimator for structural equation models with categorical data. *Psychometrika, 72,* 309-326.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62-83.

Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean- and covariance structure models. In G. Arminger, C. C. Clog, & M. E. Sobel (Eds.), *Handbook of modeling in the social and behavioral sciences* (pp. 185-249). New York, NY: Plenum Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Cai, L. (2013). flexMIRT® version 2: Flexible multilevel item factor analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*, 245-276.

Cai, L., & Monroe, S. (2013). IRT model fit evaluation from theory to practice: Progress and some unanswered questions. *Measurement: Interdisciplinary Research and Perspectives*, *11*, 102-106.

Cai, L., & Monroe, S. (2014). *A new statistic for evaluating item response theory models for ordinal data.* (CRESST Report 839). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika*, *57*, 357-369.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, *109*, 512-519.

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effect of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling*, *4*, 87-107.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics, 1*, 141-149.

Hu, L., -T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75*, 393-419.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*, 381-389.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2001). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics and Data Analysis*, *56*, 4243-4258.

Lee, S. Y., Poon, W. Y., & Bentler, P. M. (1995). A two-stage estimation of structural equation models for continuous and polytomous variables. *British Journal of Mathematical and Psychological Statistics*, *48*, 339-358.

Lee, T., & Cai, L. (2012, July). *A note on a Tucker-Lewis index for item response theory modeling*. Paper presented at the 2012 International Meeting of the Psychometric Society, Lincoln, NE.

Lord, F. M. (1952). *A theory of test scores*. Chicago, IL: Psychometric Monographs # 7.

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common factor model: implications for theory and practice. *Psychological Bulletin*, *109*, 502-511.

Marsh, H. W., Hau, K., -T., & Wen, Z. (2004). In search of golden rules: comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320-341.

Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika, 71*, 57-77.

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

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713–732.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*, 305-328.

Meece, J. L., Eccles, J. S., & Wigfield, A. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology*, *82*, 60-70.

Muthén, B. O. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*, 551-560.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115-132.

Muthén, B. O. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205-243). Newbury Park, CA: Sage Publishing.

Muthén, B. O., du Toit, S. H., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. (Unpublished Technical Report). Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Satorra, A. (1995). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika*, *60*, 489-503.

Satorra, A., & Bentler, P. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications to developmental research* (pp. 399-419). Newbury Park, CA: Sage Publishing.

Steiger, J. H., & Lind, J. C. (1980, June). *Statistically-based tests for the number of common factors*. Paper presented at the 1980 Meeting of the Psychometric Society, Iowa City, IA.

Thurstone, L. L. (1925). A method of scaling educational and psychological tests. *Journal of Educational Psychology*, *16*, 433-449.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 278-286.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1-10.

Table 1

Prompts and Item Wording for the PISA Empirical Example

Construct/ Item	*Stem/* Wording
PSC	*How much do you disagree or agree with the following statements?*
1	I get good <marks> in mathematics.
2	I learn mathematics quickly.
3	I have always believed that mathematics is one of my best subjects.
4	In my mathematics class, I understand even the most difficult work.
ANX	*How much do you disagree or agree with the following statements?*
5	I often worry that it will be difficult for me in mathematics class.
6	I get very tense when I have to do mathematics homework.
7	I get very nervous doing mathematics problems.
8	I feel helpless when doing a mathematics problem.
TASK	*How confident do you feel about having to do the following calculations?*
9	Using a <train timetable>, how long it would take to get from Zedville to Zedtown
10	Calculating how many square metres of tiles you need to cover a floor
11	Finding the actual distance between two places on a map with a 1:10,000 scale
12	Calculating the petrol consumption rate of a car

Note. *PSC* = positive self-concept as a mathematics student. *ANX* = mathematics anxiety. *TASK* = task-specific confidence

Table 2

Simulation Study: True Generating Parameters

Variable (*j*)	$\tau_{j,1}$	$\tau_{j,2}$	$\tau_{j,3}$	$\tau_{j,4}$	$\tau_{j,5}$	$\lambda_{j,1}$	$\lambda_{j,2}$	$\lambda_{j,3}$	$\psi_{j,j}$
1	***−1.27***	−0.69	**−0.28**	0.28	**1.19**	0.70	0	0	0.51
2	−1.11	**−0.71**	−0.07	**0.36**	***0.73***	0.73	0	0	0.47
3	**−0.74**	**−0.39**	−0.03	***0.24***	1.15	0.73	0	0	0.47
4	−1.15	**−0.26**	**0.06**	***0.66***	1.20	0.69	0	0	0.52
5	**−0.64**	***−0.18***	0.21	0.57	**0.94**	0	0.65	0	0.54
6	**−1.17**	−0.54	***−0.23***	0.47	**1.15**	0	0.73	0	0.42
7	***−1.15***	**−0.45**	−0.17	0.18	**0.74**	0	0.73	0	0.42
8	**−1.07**	−0.38	***0.07***	**0.55**	1.09	0	0.67	0	0.51
9	***−0.80***	−0.45	−0.07	**0.22**	**0.52**	0	0	0.62	0.47
10	**−1.02**	−0.26	0.12	**0.46**	***1.06***	0	0	0.68	0.36
11	−1.11	**−0.47**	0.40	***0.76***	**1.19**	0	0	0.76	0.20
12	−1.07	**−0.18**	**0.10**	0.37	***1.10***	0	0	0.61	0.48

Note. For $K = 6$ categories, $\tau_{j,m}$ is the mth ordered threshold for variable j. For $K = 4$, the subset of thresholds is in boldface. For $K = 2$, the further subset of thresholds is also italicized. $\lambda_{j,p}$ is the loading of the jth variable on the pth factor. $\psi_{j,j}$ is unique variance j.

Table 3

Population Correlation Matrices for Correctly Specified Model (Lower Triangle) and Model with $\varepsilon_0^* = .10$ (Upper Triangle)

Item	1	2	3	4	5	6	7	8	9	10	11	12
1	1.000	.532	.459	.455	.219	.099	.116	.121	.266	.254	.303	.231
2	.511	1.000	.569	.479	.164	.164	.117	.184	.255	.147	.265	.185
3	.511	.533	1.000	.549	.094	.162	.147	.206	.141	.213	.290	.218
4	.483	.504	.504	1.000	.185	.064	.219	.130	.226	.200	.249	.213
5	.137	.142	.142	.135	1.000	.473	.515	.510	.292	.270	.204	.157
6	.153	.160	.160	.151	.517	1.000	.645	.541	.220	.298	.262	.294
7	.153	.160	.160	.151	.517	.581	1.000	.469	.252	.268	.232	.337
8	.141	.147	.147	.139	.475	.533	.533	1.000	.284	.220	.295	.184
9	.208	.217	.217	.205	.219	.246	.246	.226	1.000	.568	.694	.436
10	.228	.238	.238	.225	.240	.270	.270	.248	.586	1.000	.721	.628
11	.255	.266	.266	.252	.269	.302	.302	.277	.655	.719	1.000	.646
12	.205	.214	.214	.202	.216	.242	.242	.222	.526	.577	.645	1.000

Table 4

Simulation Results: Null Condition

						Rejection Rates			
K	N	Stat	Reps	Mean	Var	.01	.05	.10	K-S
2	100	C_2	.71	48.6	102.9	.020	.034	.070	< .001
		$\tilde{T}_U$	.71	50.2	60.5	.006	.025	.053	< .001
	200	C_2	.88	50.1	95.0	.009	.034	.066	.371
		$\tilde{T}_U$	.88	50.3	63.2	.007	.016	.052	.001
	500	C_2	.98	51.5	113.6	.012	.059	.128	.348
		$\tilde{T}_U$	.98	51.5	95.0	.012	.053	.108	.490
	1000	C_2	1.00	50.8	116.2	.012	.064	.114	.526
		$\tilde{T}_U$	1.00	51.4	98.6	.006	.062	.100	.885
4	100	C_2	.97	51.9	97.0	.010	.052	.115	.099
		$\tilde{T}_U$	.97	51.2	68.6	.008	.027	.054	.013
	200	C_2	1.00	51.5	99.6	.014	.064	.102	.459
		$\tilde{T}_U$	1.00	50.8	75.7	.002	.028	.074	.090
	500	C_2	1.00	51.4	114.3	.014	.064	.116	.260
		$\tilde{T}_U$	1.00	51.3	96.8	.012	.054	.108	.561
	1000	C_2	1.00	51.4	109.6	.018	.052	.102	.667
		$\tilde{T}_U$	1.00	51.2	95.9	.010	.054	.082	.762
6	100	C_2	.99	51.9	96.3	.010	.062	.123	.169
		$\tilde{T}_U$	.99	51.4	69.1	.006	.036	.073	.002
	200	C_2	1.00	51.6	107.1	.016	.074	.106	.632
		$\tilde{T}_U$	1.00	51.0	86.1	.010	.050	.096	.183
	500	C_2	1.00	51.3	108.7	.008	.064	.112	.516
		$\tilde{T}_U$	1.00	51.2	104.0	.016	.050	.108	.976
	1000	C_2	1.00	51.5	105.1	.014	.056	.108	.699
		$\tilde{T}_U$	1.00	51.2	94.3	.014	.054	.090	.430

Note. K is the number of categories per variable. 'Reps' is the proportion of valid replications. 'K-S' is the two-sided Kolmogorov-Smirnov p-value. The degrees of freedom for the model is 51.

Table 5

Simulation Results: Power at $\alpha = .05$ Level

		$N = 100$			$N = 200$			$N = 500$			$N = 1000$		
ε_0^*	Stat	$K = 2$	$K = 4$	$K = 6$	$K = 2$	$K = 4$	$K = 6$	$K = 2$	$K = 4$	$K = 6$	$K = 2$	$K = 4$	$K = 6$
.01	C_2	.033	.064	.056	.062	.062	.074	.065	.080	.080	.092	.116	.108
	$\tilde{T}_U$	.014	.023	.032	.025	.036	.050	.053	.072	.060	.070	.078	.084
.05	C_2	.057	.176	.219	.118	.364	.450	.185	.820	.920	.476	.996	1.000
	$\tilde{T}_U$	.027	.068	.081	.055	.173	.212	.140	.570	.692	.308	.958	.986
.10	C_2	.085	.708	.884	.241	.982	.998	.794	1.000	1.000	.996	1.000	1.000
	$\tilde{T}_U$	.041	.298	.400	.115	.737	.886	.498	1.000	1.000	.910	1.000	1.000

Note. ε_0^* is population RMSEA. K is the number of categories per variable.

Table 6

PISA Data Example: Test Statistics and Select Fit Indices

Stat	*df*	Value	p-value	TLI	RMSEA	90% CI
C_2	51	116.61	< .001	.997	.036	(.027, .044)
$\tilde{T}_U$	51	138.30	< .001	.989	.041	(.033, .050)
$\tilde{T}_D$	51	199.42	< .001	.992	.054	(.046, .062)

Note. ‘TLI’ = Tucker-Lewis Index. ‘90% CI’ = 90% confidence interval for the RMSEA estimate.

Figure Captions

Figure 1. *Ordinal Structural Model for PISA Example*
Circles represent latent variables. PSC = positive self-concept as a mathematics student. ANX = mathematics anxiety. TASK = task-specific confidence. β = regression weight. ζ = equation disturbances. Squares represent observed variables. ϵ = unique factors.

Figure 2. *Mean and Empirical 90% Confidence Intervals for RMSEA Estimates Based on C_2 and $\tilde{T}_U$*
For each row of plots, the dashed line marks the value of ε_0^*. K is number of categories per variable. N is sample size.

Figure 3. *Bivariate Plots of RMSEA Estimates for "Nested" Datasets when $\varepsilon_0^* = 0.10$ and $N = 1000$*
RMSEA estimates based on C_2. For each plot, each point represents 1 of 500 Monte Carlo replications. The axes labels (K) indicate the number of categories per variable in the dataset. In the top row of plots, y^* indicates continuous data. Dotted lines mark $.05$. Dashed lines mark $\varepsilon_0^* = .10$.

Figure 1

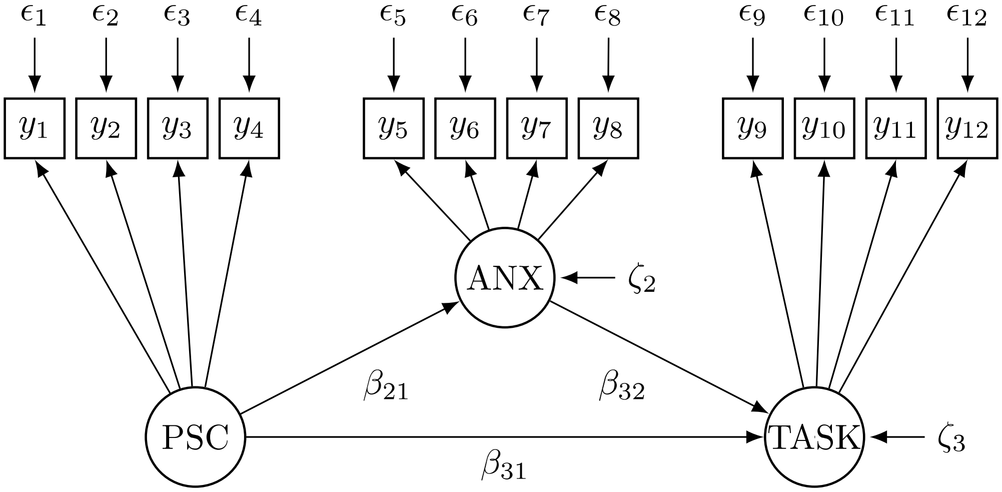
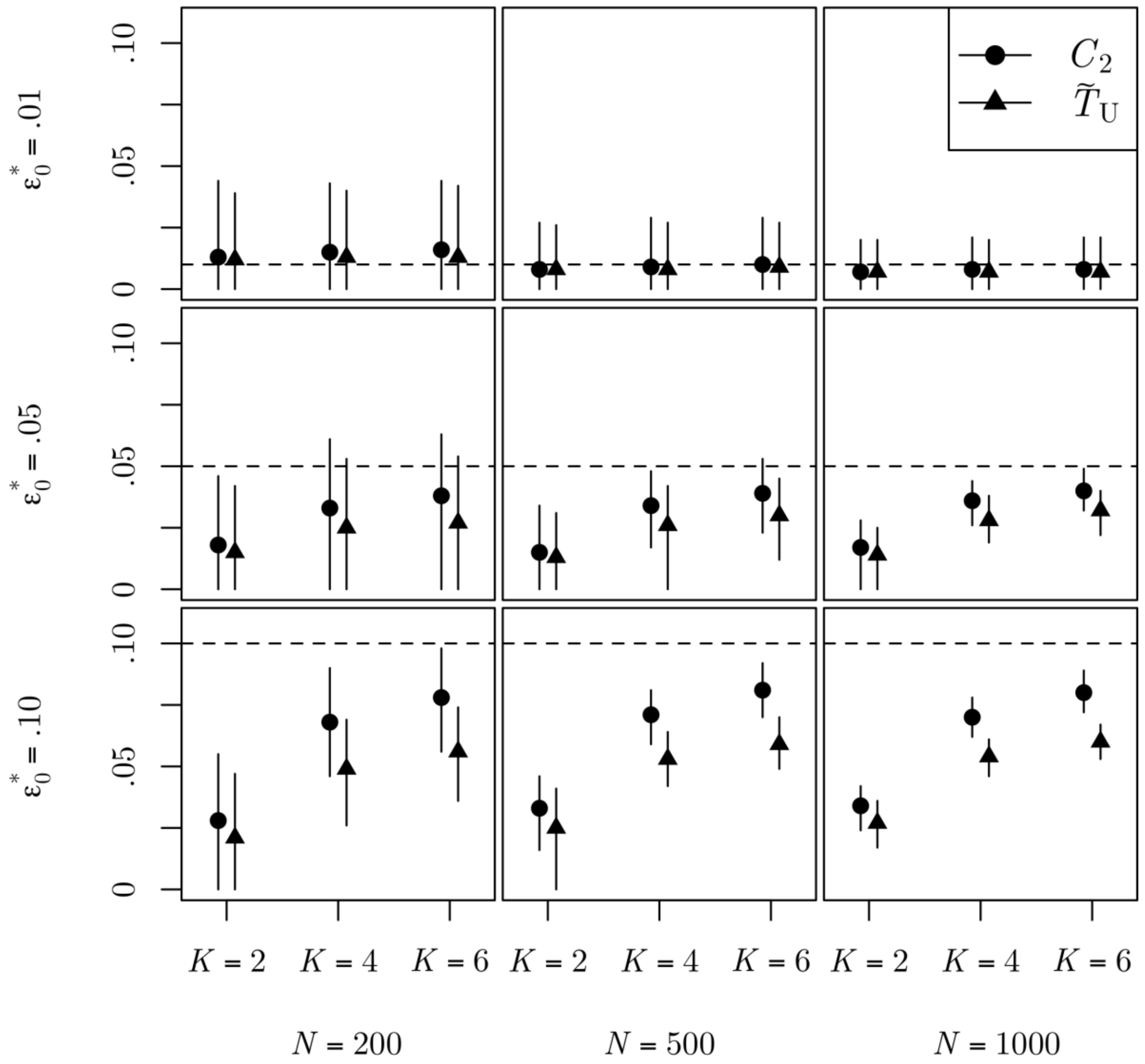
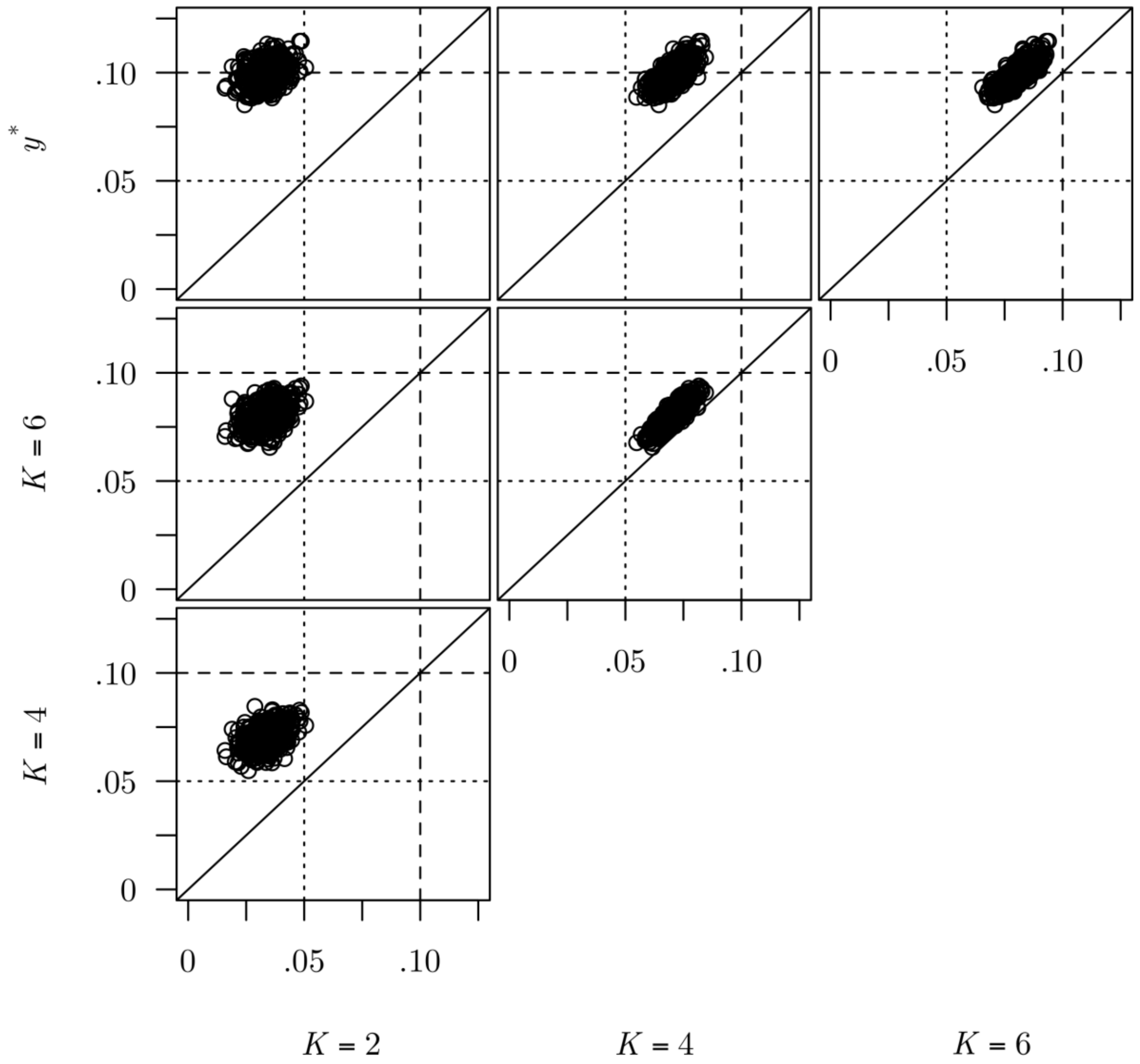

ϵ_1
ϵ_2
ϵ_3
ϵ_4
ϵ_5
ϵ_6
ϵ_7
ϵ_8
ϵ_9
ϵ_{10}
ϵ_{11}
ϵ_{12}
y_1
y_2
y_3
y_4
y_5
y_6
y_7
y_8
y_9
y_{10}
y_{11}
y_{12}
ANX
ζ_2
β_{21}
β_{32}
PSC
TASK
ζ_3
β_{31}

Figure 2



$\varepsilon_0^* = .01$
$\varepsilon_0^* = .05$
$\varepsilon_0^* = .10$
0
.05
.10
C_2
$\tilde{T}_{\mathrm{U}}$
$K = 2$
$K = 4$
$K = 6$
$N = 200$
$N = 500$
$N = 1000$

Figure 3



y^*
K = 6
K = 4
.10
.05
0
0
.05
.10
K = 2
K = 4
K = 6