Comparing common fit indices across robust estimation methods for multilevel CFA with ordered categorical data

R. Noah Padgett & Grant B. Morgan

Baylor University

Author Note

R. Noah Padgett, Department of Educational Psychology, Baylor University; Grant B. Morgan, Department of Educational Psychology, Baylor University.

This research is part of R. Noah Padgett’s master’s thesis requirements. This research was supported by Baylor University and the Department of Educational Psychology as funding toward the completion of my doctoral degree.

Correspondence concerning this article should be addressed to R. Noah Padgett, Department of Educational Psychology, One Bear Place # 97304, Baylor University, Waco, TX 76798. Contact: [noah\_padgett1@baylor.edu](mailto:noah_padgett1@baylor.edu)

Abstract

GRANT – PLEASE WRITE A 120 ABSTRACT

*Keywords:* ML-CFA, WLSMV, multilevel, categorical, ROC

Comparing common fit indices across robust estimation methods for multilevel CFA with ordered categorical data

Factor analysis models have been utilized in educational and social science investigations for over a century to provide statistical evidence for measure constructs that cannot be directly measured. However, in education and social science, many constructs cannot be investigated fully within the traditional factor analytic framework because students/participants may be influenced by factors that are contextual or beyond the students’ control. The student characteristics are typically the lowest level of analysis, called level-1. While the characteristics that reflect the school are the next level of analysis, level-2. For example, in studying how students perceive the social climate at their school, students’ perception is necessarily influenced by factors that directly affect students (level-1) and factors that influence the environment (level-2) in which they exist. When measuring student-level outcomes both levels of influence need to be accounted for in order to gain a full understanding for how students respond on, say, psychological assessments.

However, the data that arise from this form of assessment in an educational setting are organized hierarchically such that children are grouped into classrooms, and a school itself is organized hierarchically within a district, and so on. Data with a hierarchical structure typically violate independence of observations of most standard statistical methods creating standard errors that are biased (Raudenbush & Bryk, 2002). The effect of the multilevel structure can be estimated by the *intraclass correlation* coefficient (ICC), which is also called the *cluster effect*. When the cluster effect acts on the our example of school climate, one factor that may be of interest is the effect of school membership on perceived climate, which must be investigated through a multilevel measurement model. In this study, we investigated the principles of multilevel measurement models to help guide researchers when multilevel constructs are of immediate importance.

In this study, we build on knowledge that exists for the multilevel measurement models that are developed for categorical data, estimation methods choices, and model fit evaluation.

Limited studies have examined ML-CFA, particularly under conditions when data are ordered categorical. Given that data from educational settings typically arise from hierarchically structured contexts and used ordered categorical response formats, understanding how commonly used fit indices perform warrants investigation.

1. Is there a cut-off for common fit indices that can be used to differentiate a properly specified model from an misspecified model?
2. How does identification of fit vary across potential estimation methods and sample sizes for ML-CFA models?

**Multilevel Confirmatory Factor Analysis**

One of the assumptions underlying traditional single level CFA is that the cases sampled are independent of one another. However, educational and social science data rarely meet this assumption given that data often arise from nested structures. Two general approaches can be taken to account for the dependence among cases. The first approach is a sampling theory perspective. The second approach is a varying parameter modeling perspective. In the first approach, traditional single level analysis are conducted, but the standard errors are corrected for the sampling design, and this approach is sometimes referred to as a “design based” approach (Stapleton, 2013; Stapleton, Yang, & Hancock, 2016). In the second approach, parameters are treated as random components that vary across higher level units. Under this approach, the mechanism by which parameters vary across schools is explicitly modeled. Utilizing this approach accounts for the dependence among cases within unit by modeling and estimating the differences among higher level units. The second approach is known as multilevel confirmatory factor analysis (ML-CFA) and is the focus of this work.

In this study, we are focusing on the use of fit statistics for assessing ML-CFA models. ML-CFA was extended from single-level CFA by decomposing the observed covariance matrix into a pooled-within cluster covariance matrix (i.e. level-1 covariances) and a between cluster (i.e. level-2 covariances) covariance matrix to model the nested structure of data (B. Muthén, 1994). The pooled-within cluster covariance matrix defines how the item responses covary because of the individual contributions of each respondent. For the remainder of this paper, we referred to the pooled-within cluster covariance matrix as the “within covariances” or “level-1 covariances.” In contrast, the between cluster covariance matrix defines how the average response to each item covaries across groups. For the remainder of this paper, we referred to the between cluster covariance matrix as the “between covariances” or “level-2 covariances.” A full description of the technical issues surrounding specification of these two covariance structures is out of scope for the paper, and the interested reader is referred to the excellent article by Stapleton et al. (2016).

Finally, the ML-CFA model can be expressed in a path model where each level of analysis gets its own fully specified set of relationships. An example of a simple two level ML-CFA with five indicators is shown in Figure 1. The reader will notice that the model structure is equivalent across both levels. For example, B. Muthén (1994) representation explicitly shows the effect of the level-2 model on the observed indicators (see Figure 1). When the ML-CFA is specified with equivalent structure and factor loadings are constrained equal across levels, the researcher can appropriately calculate an ICC for the latent variable(s). However, a precursor to estimating the ICC in the latent variables is establishing fit of the measurement model across both levels (Stapleton, 2013). Establishing fit of ML-CFA models is an active area of research given the complexities of specifying a model for both levels, and the utility of model statistics is the focus of our investigation.

« FIGURE 1 ABOUT HERE »

**Model Fit.** A limited literature exists on the performance of fit indices in ML-CFA with categorical data (Hsu, 2009; Navruz, 2016). However, a few studies have investigated fit indices in ML-CFA when indicators are multivariate normally distributed (Asparouhov & Muthén, 2007; Hox & Maas, 2001; Hsu, 2009; Hsu, Kwok, Lin, & Acosta, 2015; Ryu & West, 2009). Under data simulated as multivariate normally distributed, the consensus so far is that CFI, TLI, RSMEA, SRMR-W are only sensitive to misspecification of the level-1 model, while the SRMR-B is able to detect misspecification of the level-2 model as well. However, the study that has most thoroughly investigated fit in ML-CFA is limited in generality due to the methods employed (Hsu et al., 2015). Hsu et al. (2015) found that the Hu and Bentler (1999) cut-off criteria for CFI, TLI, and RSMEA performed well for detecting misspecification at level-1 but not when only level-2 is specified. However, when using categorical data with the WLSMV or WLSM estimation method, Navruz (2016) found that none of the fit statistics were able to consistently identify the correct model with the traditional cut-offs.

**Recommended Estimation Methods.** To date, there is no clear recommendations about which estimation method(s) should be used with categorical data in multilevel settings. In Asparouhov and Muthén (2007), the authors investigated ML-CFA with categorical indicators and compared two estimation methods: MLR and WLSM. Asparouhov and Muthén (2007) recommended using WLSM over MLR for model estimation due to WLSM resulted in less biased parameters on average in the one condition simulated. On the other hand, Study 2 of Hsu (2009) only used WLSMV to investigate the effects of dichotomous indicators on ML-CFA parameter recovery. In Navruz (2016), the authors compared WLSM with WLSMV and concluded that under the conditions included, neither estimation method should be used for ordered categorical data in multilevel settings. The choice of estimation method appears to depend on the circumstance at hand without a clear consensus from the methodological literature.

Purpose and Research Questions

**Methods**

To assess fit statistics in ML-CFA, a Monte Carlo simulation was employed in this study. A cursory introduction to the technical details of ML-CFA models and the estimation of these models are given below in the section on ML-CFA Model. Additional details on the estimation and fit statistics are given below. The description of data generation and simulation design are also given below. The study outcomes are explained along with the statistical analyses performed are described below.

**Model Estimation Methods.** All models were estimated using M*plus* (L. Muthén & Muthén, 2017, version 8.2). Researchers have the choice of many potential estimation methods in M*plus*, and each was initially developed to solve a particular estimation limitation of another estimation methods. A full list of the available estimation methods for models in M*plus* can be found in the User’s Guide (L. Muthén & Muthén, 2017). In this project, we focused on ordered categorical responses, for which the maximum likelihood with robust standard errors (MLR) and weighted least squares mean and variance adjusted (WLSMV) estimation methods have been tested for ML-CFA models. However, for single level CFAs with categorical data, there is evidence that unweighted least squares (ULS) performs well (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009). But, the ULS estimation method is not available for multilevel models in M*plus*, so instead, we used the unweighted least squares mean and variance adjusted (ULSMV) method. Therefore, we investigated three estimation methods (MLR, WLSMV, and ULSMV).

**Model Fit indices.** The commonly used measures of fit of interest in this project are innately reported by M*plus* v8.2 when estimating ML-CFA models (e.g., *χ*2 test, CFI,

TLI, RMSEA, SRMRW, and SRMRB).

**Data Generating Process**

Data for this project were generated using M*plus* v.8.2 (L. Muthén & Muthén, 2017) utilizing the MONTECARLO command along with MplusAutomation (Hallquist & Wiley, 2018) package in R.

**Fixed Design Factors.** Data were generated from a factor model with 10 items and two correlated factors at both levels. Across levels, the structure and factor loadings were specified to be invariant. The factor loadings were set to 0.60, which corresponds to the average of reported factor loadings (DiStefano & Hess, 2005). At level-1, factors were generated with a fixed unit variance. Fixing the factor variance to one made specifying the level-1 correlation simple, meaning that all that is needed is further specifying the covariance to the correlation. The factors at each level were specified to have a correlation of 0.3, similar to Hsu et al. (2015). The level-2 factor (co)variances depended on the ICC condition, and more detail is given in the ICC section of manipulated factors. The correlation is constant, but the specific value of (co)variance at level-2 changed across conditions.

Figure 2 shows the population structure of the ML-CFA simulated for one of the ICC conditions. The two factors were measured with 6 items and 5 items respectively. Notice that one indicator is cross-loaded on both factors at both levels in the population. Previous studies of model fit measures have typically generated data from a population structure with at least one cross-loading so that the effects of misspecification can be examined (Hsu, 2009; Hsu et al., 2015; Hu & Bentler, 1999; Navruz, 2016).

« FIGURE 2 ABOUT HERE »

The observed indicators were generated to be ordered categorical. One of the difficulties simulating data with categorical indicators in ML-CFA is correctly specifying the variance of the observed indicators. For factor models, there is an interdependence between the factor variance, factor loading(s), indicator residual variance, and total variance of of the latent response continuum. For this study, the indicators were set to be approximately normally distributed which is the best case possible for using ordered categorical data. Varying thresholds for generating the appropriate observed distributed were based on the complex interdependence of the model parameters. The condition-varying thresholds used to categorize the continuum based on the scheme given in Equation 3 are reported in reported in Table A1.

**Manipulated Design Factors.** Factors that varied across simulation conditions are indicator ICC, latent variable ICC, number of groups, and number of level-1 units. The ICC for the observed variables (ICC*O*) had three different levels: 0.1, 0.3, and 0.5. Each of the indicator ICCs were manipulated by changing the residual variance at level-2 of each item. The level-2 item variances were 0.11, 0.43, and 1 for each of the ICC conditions, respectively. Furthermore, the latent variable ICCs (ICC*L*) were manipulated using the level-2 factor variances. Values of 0.1 and 0.5 were investigated corresponding to a low and a high level.

Then, the sample size was varied in two ways. First, by the number of groups (N2), which is known to be one of the most influential components to estimation of ML-CFA. For this investigation, four different group sizes were under investigation, namely 30, 50, 100, and 200. Secondly, by the number of level-1 units within groups. The number of units sampled within each group (N1) was held constant across groups for simplicity with three different levels. The number of units varied among 5, 10, and 30.

Summary of Model Fitted

The major focus of this study was the identification of the correctly specified model amongst misspecified models. For each generated dataset, four different model specifications were fit, namely 1) level-1 and level-2 measurement model are correctly specified (C); 2) level-1 measurement model is misspecified (M1); 3) level-2 measurement model is misspecified (M2); and 4) level-1 and level-2 measurement models are misspecified (M12). The misspecification at each level comes from omitting the cross-loadings at the desired level(s). In additional to model specification, three different estimation methods were tested. This creates an additional layer of models that were estimated within each condition. The MLR, ULSMV, and WLSMV estimation methods were used to estimate the specified models in each condition.

Influence of Design Factors

The influence of manipulated design factors on the distributions of each fit statistic was assessed with an experimental design framework. Using an experimental design framework to summarize the results of Monte Carlo simulation studies provides a rigorous framework for describing the large quantities of data.For each fit statistic, the overall influence of all interactions and main effects were assessed with partial-*ω*2.

ROC Analysis

We conducted receiver-operating-characteristic (ROC) analysis to study the relationship between the sensitivity and specificity of cut-off criteria of each fit statistic. Sensitivity refers to how often the cut-off criteria correctly identifies the correct model, which can be optimized by making the criteria liberal and easy to meet. Specificity refers to how often only the correct model is identified, which can be optimized by making the criteria as conservative as possible and difficult to meet. A trade-off between these two criteria is therefore necessary since the aim is to optimize both. The trade-off can be quantified by calculating the area under the curve (AUC) across a range of cut-off criteria. ROC analyses were conducted using the pROC package in R (Robin et al., 2011).

Simulation Summary

For each condition, 500 replications were used resulting in 36,000 generated datasets. Additionally, for each dataset simulated, four model specifications by three estimation methods were tested. Therefore, 12 models were estimated for each dataset with a total of 432,000 models estimated. The specific outcomes for this study are 1) the distribution of fit statistics (CFI, TLI, RMSEA, SRMRW, and SRMRB) across design factors, 2) influence of manipulated design factors (partial-*ω*2), and 3) performance of fit statistic cut-off criteria tested with ROC analysis across conditions and estimation methods. In addition to these outcomes, convergence rates of fitted models across conditions were reported along with any other estimation issues.

**Results**

Effect of Study Design Factors

The results of the ANOVA for each fit statistic are reported in Table B1. The results are reported as the partial-*ω*2 effect size estimates. The meaning of each value can be interpreted for the N1 (number of level-1 units) effect as follows: for CFI, 8.4% of the variability in observed scores can be attributed to the number of level-1 units (N1) that were sampled per group after controlling for all other design factors (i.e., level-2 sample size, observed and latent ICCs, model specification, and estimation method).

Overall, the trends across fit statistics are that the number of level-1 units, model specification, estimation method, and the interaction between number of level-2 units and estimation method are the most influential factors on the distribution of fit statistics. Choice of estimation method accounted for between 5-45% of the variability if fit scores across all statistics after controlling for other design effects (see Table B1). For CFI, TLI, RMSEA, and SRMRW, the interaction between level-2 sample size and estimation method accounted for 7-15% of the variability while for SRMRB only 1% was accounted for by this interaction. These trends of sample size effects and estimation method effects are the focus of the remaining analyses.

Distribution of Fit Statistics

The distributions of fit statistics across sample sizes, model specification and estimation methods are shown in Figures 3-6. For simplicity and space restrictions, not all fit statistics and conditions are shown; for the full breakdown, the interested reader should refer to the online material (REMOVED FOR PEER REVIEW).The marginal distributions of all fit statistics across all 72 conditions are reported in Appendix C. Model specification, estimation method, and level-2 sample size were focused on for discussion because these three effects accounted for the most variability across fit statistics.

The distribution of CFI and TLI are nearly identical, so only CFI is discussed. These statistic seems to have two major features that vary across model specifications and model estimation methods (see Figure 3). The means of CFI values differed across model specification. The most obvious differences in means of CFI scores occurs when the level-1 model is misspecified. This trend is seen across all estimation methods. Secondly, the distribution of CFI changes in variability mostly depending on which estimation method was used. For example, the range of all CFI scores under WLSMV is only about .6 to 1 while for MLR the range spans from 0 to 1 (i.e., the full range of potential CFI scores). These results were also found only using one type of misspecification (i.e., the omission of one cross loading). MLR also yielded the lowest value for a correctly specified model

(CFI=.214) and this occurred in condition 1 where sample sizes were smallest (N1 = 5, N2 = 30) and ICCs were low (ICC*L* = *.*1*,*ICC*O* = *.*1). This trend of unequal variability is found primarily between estimation methods and not models.

« FIGURE 3 ABOUT HERE »

The distribution for RMSEA is shown across models and estimation methods in

Figure 4. One aspect of the distribution of RMSEA that immediately stands out is that for ULSMV and WLSMV nearly all estimates fall below the commonly used cut-off of .06. This trend is even more apparent when one looks at the distribution of RMSEA across all conditions irrespective model and estimation method (see Figure C3 in the Appendices). However, despite the astoundingly low estimates across all conditions and estimation methods, the estimates for model C were lower on average. The largest changes in RMSEA appear to due to misspecification of the level-1 factor structure.

« FIGURE 4 ABOUT HERE »

The distribution for SRMRW is shown across models and estimation methods in

Figure 5. Two major features stand out for SRMRW. First, there appears to be a clear discrimination between models that have a correctly specified level-1 model and those models that do not (similar to CFI, TLI, and RMSEA). Secondly, no obvious difference between model C and model M2 can be discerned at first glance. This lack of discrimination between model C and model M2 is expected given that SRMRW is only designed to help detect misspecification of the level-1 model (Hsu et al., 2015).

« FIGURE 5 ABOUT HERE »

The distribution for SRMRB is shown across models and estimation methods in Figure 6. One major feature of the distributions of SRMRB stands out. That one feature is little variation we observed between models and estimation methods. Prior literature (Hsu et al., 2015) suggested that SRMRB would perform best in discriminating between a correct and incorrect level-2 covariance structure specification. However, based on Figure 6 we could not identify a clear pattern to SRMRB across models and estimation methods. There is a slight visible increase in thevalues of SRMRB under estimation with MLR, but this increase is so small that there is practically no difference.

Aside from the small amounts of variability observed among model specifications, we noticed one peculiar attribute of the distribution of SRMRB in our study. Some excessively large values were observed in Conditions 1 and 7. In condition 7 (N2 = 30, N1 = 10, ICC*L* = *.*1*,*ICC*O* = *.*1), the highest value observed was 8.49. In this replication, no obvious signs of an in-admissible solution occurred, and according to CFI, TLI, and RMSEA, this model fits perfectly, and the *χ*2 test of goodness of fit was not rejected. This model was estimated under misspecificated Model M1.

« FIGURE 6 ABOUT HERE »

ROC Analyses Results

One of the defining features of using ROC analysis is the creation of ROC curves. A

ROC curve helps create a visual representation for the quality of classification based on a systematically varying cutoff criteria. In this case, the cutoff criteria was the value of the fit statistic that differentiated between a correctly specified model and an incorrectly specified model. In ML-CFA models, there are three ways that a model can be misspecified (see Methods Summary of Models Fitted). Three ROC analyses were initially performed for detecting 1) any misspecification (Model C versus Model M1-M2-M12); 2) misspecification at level-1 (Model C versus Model M1); and 3) misspecification at level-2 (Model C versus Model M2). The results for ROC analyses across all conditions averaged across all estimation methods are shown in Figure 7 and Table 2. The curves in Figure 7

help describe how strongly each fit index discriminates between the correctly specified model and the misspecified models.

« FIGURE 7 ABOUT HERE »

Across conditions and estimation methods, CFI, TLI, and RMSEA appear to perform approximately equally. The first (left) panel of Figure 7 shows how the predictive value for detecting the correct model varies across fit indices. SRMRW and SRMRB have a poor discriminate curve. The discrimination of SRMRW and SRMRB can also be seen as inadequate or at least under performing compared to CFI, TLI, and RMSEA based on the AUC value for each index reported in Table 2. SRMRB has a poor AUC of .598, though the performance of SRMRW appears only slightly behind that of CFI, TLI, and RMSEA

at.742.

« TABLE 2 ABOUT HERE »

SRMRW had a lower AUC than CFI, TLI, and RMSEA for detecting any model

misspecification detection. However, SRMRW provided the sharpest transition for detecting a level-1 model misspecification (see middle panel of Figure 7). SRMRB has essentially zero predictive power of level-1 misspecification, because the curve is a straight line along the diagonal. In the right most panel, SRMRB performs as well as CFI, TLI, and RMSEA at predicting level-2 misspecification, although this predictive performance is barely greater than random chance for all indices.

Despite the apparent poor performance of these fit statistics at detecting level-2 misspecification, there is at least potential for identifying whether there is any misspecification in a model. The overall optimal threshold for CFI irrespective of estimation method is .977, with an AUC of .816. These results mean that, first, a cutoff value of .977 optimizes the specificity and sensitivity of classifying a model as correctly specified over any type of misspecification. Secondly, an AUC of .816 means that there is approximately a .816 probability that a model with a CFI value above .977 is correctly

specified.

The results for ROC analyses across conditions for each estimation method are shown in Figure 8. The overall trend across estimation methods is that a similar shape occurred between the three types of ROC analyses performed, and these three plots have a similar trend across estimation methods with only minor differences. That is, the estimation method effect is difficult to detect from a coarse view of the ROC analyses. However, in Tables 3-E2, subtle differences can be parsed a out across estimation methods and level-2 sample sizes. Note, we did not conduct any formal statistical tests to compare AUCs, though these comparisons are possible for future investigations.

« FIGURE 8 ABOUT HERE »

One of these estimation methods differences was found in the identification of correctly specified models (left most panels of Figure 8). The WLSMV estimation method yielded the sharpest curve for CFI, TLI, and RMSEA compared to MLR and ULSMV (Figures 8a-8b). Under model estimation with WLSMV, these statistics appear to performed best at identifying the completely correct model specification compared to any form of misspecification. Another subtle difference among these three estimation methods is the shape of the ROC curves under correctly specified level-1 models.

Similar to the overall ROC analyses, the SRMRB performed poorly at identifying the correctly specified models versus any misspecification across estimation methods when level-2 sample size was lower than 100 (see Table 3). In general, the discrimination of these common statistics appears low, but a decision still needs to be made of the adequacy of model fit. To help with this judgement, many researchers use Hu and Bentler (1999) cut-off criteria for CFI, TLI, RMSEA, and SRMRW/B. However, these values were not recovered in our simulation based on these ROC analyses for completely correct model selection.

« TABLE 3 ABOUT HERE »

Beyond complete model selection with these fit indices, there is potential that these statistics can help identify where part of a model is correctly specified. The ROC analyses for classifying models as having a correct level-1 model specification are reported in Table

E1. The ROC analyses for classifying models as having a correct level-2 model specification are reported in Table E2. The statistic that performed “best” was SRMRB, but, the AUCs were still low. These statistics do not appear to help discriminate among incorrect and correct level-2 models. This is contrary to some previous simulation studies that found

SRMRB to be useful in identifying level-2 misfit (Hsu et al., 2015).

Discussion

Multilevel measurement models allow researchers to test hypotheses of complex phenomena using data from complex sampling methods. The complex organizational structure of educational and psychological data are particularly well suited for these types of multilevel measurement models; however, finding evidence for the validity of these measurement model specification is not well established. The use of common fit indices (e.g., CFI, TLI, RMSEA, SRMRW, SRMRB) should be interpreted with caution because the sensitivity of commonly used fit statistics to misspecification may depend on the estimation method used. The influence of robust estimation methods on fit indices in complex settings is still considerably unknown (Asparouhov & Muthén, 2007; DiStefano & Morgan, 2014; Hsu, 2009; Navruz, 2016; Nestler, 2013). For example, DiStefano and

Morgan (2014) found that the CFI and RMSEA were generally not sensitive to variation in number of response categories, nonnormality, and sample size, but slightly poorer estimates of fit were observed when number of categories was only two and the response distribution was skewed. In their study, the performance of CFI and RMSEA may depend on the fact that models were specified correctly. The major difference between the current study and DiStefano and Morgan (2014) is that the categorical CFA model is extended into the multilevel space.

In multilevel CFA with categorical data, CFI and TLI were generally only influenced by misspecification of the level-1 model while only some evidence suggests these statistics are sensitive to sample size and estimation method used. The similarities in performance of CFI and TLI were not observed in a previous simulation studies of ML-CFA (Navruz, 2016). These common fit indices were not equally sensitive to misspecification of the level-1 model versus the level-2 model. The differences in sensitivity to model misspecification at different levels leads to varying degrees of utility of these fit statistics. The difficulty in using these fit statistics for ML-CFA is that misspecification could be due to the within- or between-group model. This consideration is one of the reason the SRMR statistics are so conceptually useful; that is, they guide researchers on which part of the model may be misspecified. SRMRB was found to be variable across conditions and/or models to consistently discriminate between a correctly and incorrectly specified between groups model. The SRMRB statistic also returned some very large values for smaller number of groups. For example, in a condition with a very small number of groups (30) returned an estimated SRMRB for a correctly specified model of 2.54. Researchers may be able to get fine grained information of the fit of each level’s model by using the level specific indices proposed by Ryu and West (2009). However, only one additional study was found that described these level specific statistics (Ryu, 2014). These level-specific indices should be evaluated in future studies for applicability to ML-CFA and multilevel SEM more generally.

Recommendations for ML-CFA Fit Statistics

The availability of fit indices is intended to aid researchers in diagnosing misfit and ultimately select the correct model for interpretation. Based on the mixed findings of the current study, the utility of available fit indices for interpreting ML-CFA models is may be limited. Researchers are strongly encouraged to seek more ways of evaluating model fit beyond the fit indices commonly reported by M*plus*. The fit statistics described in Ryu and West (2009) appear promising to this end. If researchers are use the commonly reported indices, they should be interpreted with caution, especially the number of groups is low (below 100). The recommended cut-off criteria for CFI and TLI match those of those commonly used when the ML-CFA model treats the ordered categorical data as continuous and estimated with MLR. When a robust estimation method is used different criteria were found to be needed. Across all estimation methods, we found that SRMRB performed *best* at detecting level-2 misfit, but this statistic still does not discriminate between correct and incorrect level-2 models well. A summary of the recommended criteria for the commonly used fit statistics is outlined in Table 4.

« TABLE 4 ABOUT HERE »

Limitations & Delimitations

As with any simulation study, the results of only generalize to the limited conditions examined. That said, the conditions chosen were selected to mirror conditions of applied researchers as close as possible while still maintaining parsimony. Estimation with MLR resulted in the most usable cases per cell on average, but this may have occurred because data were treated as continuous. Additionally, the cells of this design ended up with unequal sample sizes due to convergence issues and the number usable replications. In some cells, the number of usable cases was zero. Proper solutions were checked by looking for negative variances; however, another type of improper solution is impossible estimates of factor loadings.

References

Asparouhov, T., & Muthén, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. In *Joint statistical meeting: Section on statistics in epidemiology.* Retrieved from <https://www.statmodel.com/download/JSM2007000746.pdf>

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling*, *21*(1), 102–116. doi: 10.1080/10705511.2014.859510

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

Bollen, K. A., Bauer, D. J., Christ, S. L., & Edwards, M. C. (2010). Overview of structural equation models and recent extensions. In *Statistics in the social sciences: Current methodological developments* (pp. 37–79). doi: 10.1002/9780470583333.ch2

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*(3), 225–241. doi: 10.1177/073428290502300303

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling*, *21*(3), 425–438. doi: 10.1080/10705511.2014.915373

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Charlotte, NC: Information Age Publishing.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, *9*(4), 466–491. doi: 10.1037/1082-989X.9.4.466

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS sstimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 625–641. doi: 10.1080/10705510903203573

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. doi: 10.1080/10705511.2017.1402334

Hox, J. J., & Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, *8*(2), 157–174. doi: 10.1207/S15328007SEM0802\_1

Hsu, H. Y. (2009). *Testing the effectiveness of various commonly used fit indices for detecting misspecifications in multilevel structure equation models* (Unpublished doctoral dissertation). Texas A&M University.

Hsu, H. Y., Kwok, O., Lin, H., & Acosta, S. (2015). Detecting isspecified multilevel structural equation models with common fit indices: A monte carlo study. *Multivariate Behavioral Research*, *50*(2), 197–215. doi: 10.1080/00273171.2014.977429

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. doi: 10.1080/10705519909540118

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. doi: 10.1007/BF02294210

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376–398. doi: 10.1177/0049124194022003006

Muthén, L., & Muthén, B. (2017). *Mplus User’s Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Navruz, B. (2016). *The behaviors of robust weighted least squares estimation techniques for categorical/ordinal data in multilevel CFA models* (Unpublished doctoral dissertation). Texas A&M University.

Nestler, S. (2013). A Monte Carlo study comparing PIV, ULS and DWLS in the estimation of dichotomous confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 127–143. doi: 10.1111/j.2044-8317.2012.02044.x

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Second ed., Vol. 1). SAGE Publications, Inc.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Ryu, E. (2014). Model fit evaluation in multilevel structural equation models. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.00081

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 583–601. doi: 10.1080/10705510903203466

Stapleton, L. M. (2013). Multilevel structural equation modeling with complex sample data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 521–562). Charlotte, NC: Information Age Publishing.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*(5), 481–520. doi: 10.3102/1076998616646200

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling*, *17*(3), 392–423. doi: 10.1080/10705511.2010.489003

Table 2

*Overall ROC analysis AUCs and optimized threshold*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | AUC1 | Thresh[[1]](#footnote-1) | Spec3 | Sens4 |
| CFI | 0.816 | 0.977 | 0.702 | 0.855 |
| TLI | 0.815 | 0.972 | 0.702 | 0.855 |
| RMSEA | 0.803 | 0.015 | 0.685 | 0.829 |
| SRMRW | 0.742 | 0.038 | 0.728 | 0.723 |
| SRMRB | 0.598 | 0.067 | 0.804 | 0.352 |

*Note.* 1 AUC = Area Under the Curve;

Table 3

*Completely correct specification ROC analysis AUCs and optimized threshold by estimation method and level-2 sample size*

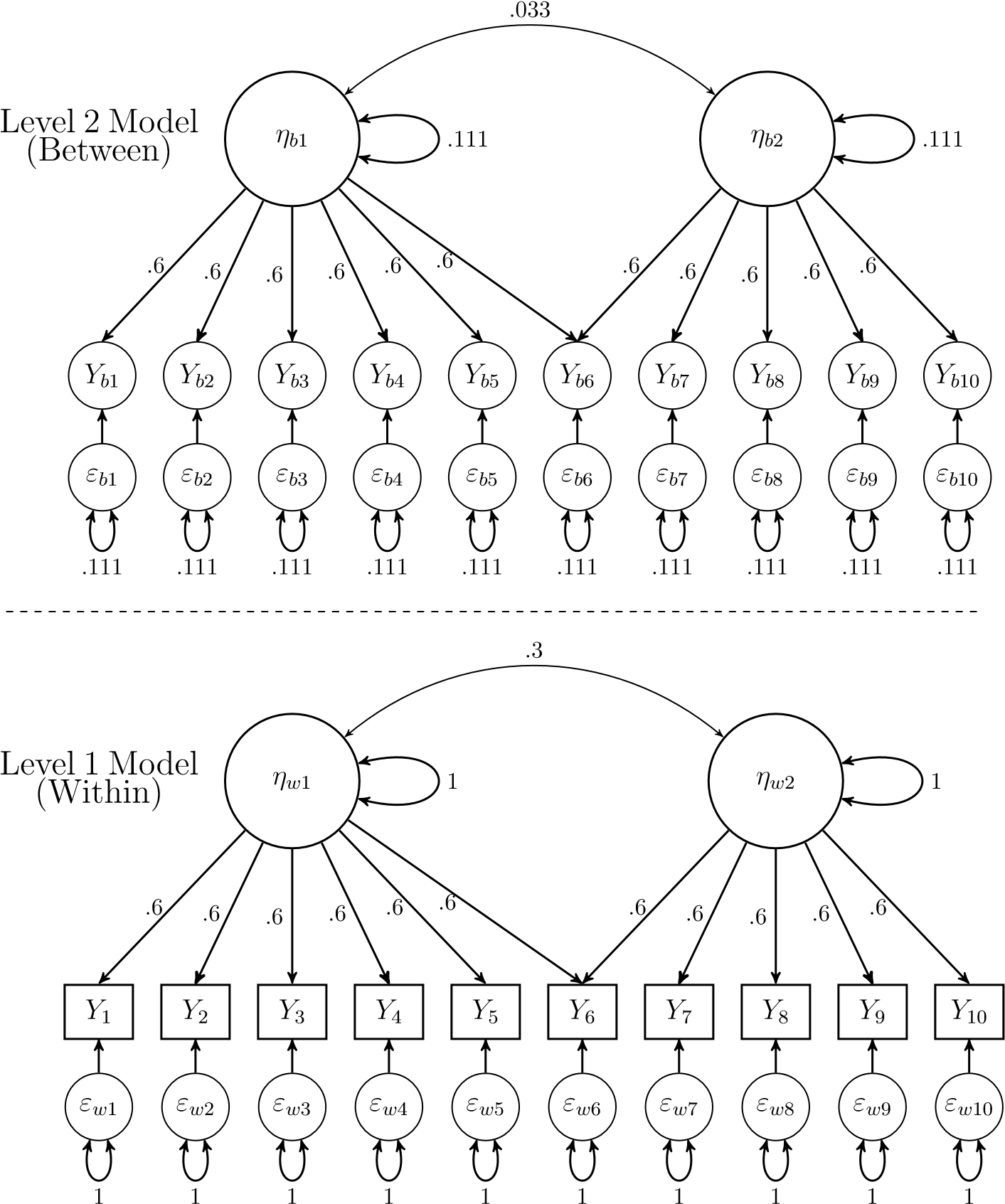
MLR

ULSMV

WLSMV

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Index | N[[2]](#footnote-2) | AUC | Threshold | AUC | Threshold | AUC | Threshold |
| CFI | 30 | 0.747 | 0.940 | 0.628 | 0.982 | 0.705 | 0.983 |
|  | 50 | 0.828 | 0.956 | 0.712 | 0.973 | 0.802 | 0.980 |
|  | 100 | 0.893 | 0.974 | 0.827 | 0.971 | 0.876 | 0.979 |
|  | 200 | 0.917 | 0.981 | 0.911 | 0.975 | 0.910 | 0.986 |
| TLI | 30 | 0.746 | 0.928 | 0.628 | 0.979 | 0.705 | 0.979 |
|  | 50 | 0.827 | 0.947 | 0.712 | 0.968 | 0.801 | 0.976 |
|  | 100 | 0.893 | 0.968 | 0.827 | 0.965 | 0.876 | 0.975 |
|  | 200 | 0.917 | 0.977 | 0.911 | 0.970 | 0.910 | 0.984 |
| RMSEA | 30 | 0.725 | 0.029 | 0.628 | 0.008 | 0.700 | 0.013 |
|  | 50 | 0.814 | 0.026 | 0.706 | 0.009 | 0.790 | 0.011 |
|  | 100 | 0.889 | 0.020 | 0.813 | 0.012 | 0.873 | 0.015 |
|  | 200 | 0.915 | 0.017 | 0.893 | 0.013 | 0.910 | 0.014 |
| SRMRW | 30 | 0.674 | 0.042 | 0.642 | 0.052 | 0.682 | 0.050 |
|  | 50 | 0.737 | 0.037 | 0.709 | 0.048 | 0.733 | 0.047 |
|  | 100 | 0.816 | 0.037 | 0.803 | 0.046 | 0.797 | 0.045 |
|  | 200 | 0.832 | 0.032 | 0.855 | 0.042 | 0.813 | 0.038 |
| SRMRB | 30 | 0.565 | 0.143 | 0.567 | 0.106 | 0.576 | 0.117 |
|  | 50 | 0.610 | 0.119 | 0.596 | 0.081 | 0.613 | 0.096 |
|  | 100 | 0.673 | 0.093 | 0.649 | 0.060 | 0.670 | 0.071 |
|  | 200 | 0.732 | 0.069 | 0.701 | 0.047 | 0.723 | 0.054 |

*Note.* 1 AUC = Area Under the Curve; and



*Figure 2*. Example of simulation data generating model specification.

*Note.* The observed variance of each indicator is fixed to one across all items with an observed and latent ICC of 0.1.

Any Mis.

Level−1 Mis.

Level−2 Mis.

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

Specificity

Sensitivity

Fit Statistics CFI TLI RMSEA SRMRW SRMRB

*Figure 7*. ROC analysis curves for identifying correctly specified models

*Note.* The “Any Mis.” panel reflects the ROC analyses for identifying the data generating model

(i.e., the completely correct specification) over the three types of misspecification. The “Level-1 Mis.” ROC analyses reflect the comparison between models with a correctly specified level-1 model versus models that are misspecified at level-1 (i.e., Models C vs. M1). The “Level-2 Mis.” ROC analyses reflect the comparison between models with a correctly specified level-2 model versus models that are misspecified at level-2 (i.e., Models C vs. M2).

Any Mis.

Level−1 Mis.

Level−2 Mis.

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

Specificity

Sensitivity

Fit Statistics CFI TLI RMSEA SRMRW SRMRB

MLR Estimation Method

Any Mis.

Level−1 Mis.

Level−2 Mis.

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

Specificity

Sensitivity

Fit Statistics CFI TLI RMSEA SRMRW SRMRB

ULSMV Estimation Method

Any Mis.

Level−1 Mis.

Level−2 Mis.

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

Specificity

Sensitivity

Fit Statistics CFI TLI RMSEA SRMRW SRMRB

WLSMV Estimation Method

*Figure 8*. ROC analysis curves across estimation methods

1. Thresh = Threshold determined by best

   Specificity and Sensitivity; 3 Spec =

   Specificity; and 4 Sens = Sensitivity. [↑](#footnote-ref-1)
2. Threshold determined by best specificity and sensitivity. [↑](#footnote-ref-2)