Testing Measurement Invariance in Longitudinal Data With Ordered-Categorical Measures

Yu Liu, Roger E. Millsap, Stephen G. West, Jenn-Yun Tein, Rika Tanaka, Kevin J. Grimm

Funded by:

- National Institute of Mental Health

Abstract

A goal of developmental research is to examine individual changes in constructs over time. The accuracy of the models answering such research questions hinges on the assumption of longitudinal measurement invariance: The repeatedly measured variables need to represent the same construct in the same metric over time. Measurement invariance can be studied through factor models examining the relations between the observed indicators and the latent constructs. In longitudinal research ordered-categorical indicators such as self- or observer-report Likert scales are commonly used, and these measures often do not approximate continuous normal distributions. The present didactic paper extends previous work on measurement invariance to the longitudinal case for ordered-categorical indicators. We address a number of problems that commonly arise in testing measurement invariance with longitudinal data including model identification and interpretation, sparse data, missing data, and estimation issues. We also develop a procedure and associated R program for gauging the practical significance of the violations of invariance. We illustrate these issues with an empirical example using a subscale from the Mexican American Cultural Values scale. Finally, we provide comparisons of the current capabilities of three major latent variable programs (lavaan, M*plus*, OpenMx) and computer scripts for addressing longitudinal measurement invariance.

Key words: measurement invariance; ordered-categorical; longitudinal; confirmatory factor analysis; practical significance

Testing Measurement Invariance in Longitudinal Data with Ordered-Categorical Measures

Developmentalists are inherently interested in studying individual and population changes over time (e.g., depression, externalizing behavior, motivation for educational attainment). Models of change (e.g., growth models) hinge on the idea that the repeatedly measured variable carries the same meaning and the same scale over all time points and over all individuals; this is the idea of longitudinal measurement invariance. Often, researchers administer the same scale to participants (e.g., children, other informants) and assume measurement invariance holds; however, in many cases measurement invariance may not hold because the same scale can measure a different construct at different ages (especially given the rapid transitions occurring in adolescence). If measurement invariance does not hold, then the observed changes may reflect changes in what is being measured rather than the level of the construct of interest. Thus, evaluation of longitudinal measurement invariance is critical to drawing valid conclusions about growth and change in the level of latent constructs over time (e.g., Leite, 2007; Wirth, 2008).

Measurement invariance can be evaluated through the use of multivariate measurement models, such as confirmatory factor and item response models. In the published literature, evaluation of measurement invariance has most commonly been applied to cross-sectional data, to study whether the continuous measured indicators reflect the same construct on the same scale in different groups (e.g., gender, ethnic, language groups; see Meredith, 1993; Widaman & Reise, 1997). More recently, methods for establishing longitudinal measurement invariance with continuous measured indicators were developed within a confirmatory analysis (CFA) framework (e.g., Khoo, West, Wu, & Kwok, 2006; Meredith & Horn, 2001) and several demonstrations and applications have appeared in the literature (e.g., Millsap & Cham, 2012;

Widaman, Ferrer, & Conger, 2010). Figure 1 illustrates the basic CFA models used to test measurement invariance in the (a) multiple groups and (b) longitudinal cases for a one-factor model with continuous indicators. Discussion of Figure 1 is deferred to the *Testing Measurement Invariance with Continuous Variables* section.

In many applied situations, key constructs are measured by a collection of ordered-categorical indicators (e.g., self- or observer-report Likert scale items). A comprehensive presentation of issues associated with testing longitudinal measurement invariance with ordered-categorical data has not been presented in the literature. Treating ordered-categorical data as continuous might sometimes be acceptable when there are several (i.e., 5+) response categories. However, this practice can lead to biased parameter estimates when there are fewer categories (e.g., Beauducel & Herzberg, 2006; DiStefano, 2002; Dolan, 1994; Rhemtulla, Brosseau-Liard, & Savalei, 2012) or when the observed indicator distributions are skewed (e.g., asymmetric threshold spacing[1], Rhemtulla et al., 2012). Thus, ordered-categorical CFA will often be the approach of choice when measured indicators are ordered-categorical. Methods of establishing longitudinal measurement invariance for ordered-categorical indicators within the CFA framework build on the foundations of tests of multiple-group measurement invariance for continuous indicators (Meredith, 1993; Widaman & Reise, 1997), tests of longitudinal measurement invariance for continuous indicators (Meredith & Horn, 2001; Khoo et al., 2006), and CFA models for ordered-categorical indicators (Muthén, 1984; Wirth & Edwards, 2007).

The purpose of this didactic paper is to consider theoretical issues and many of the practical problems that can arise when testing longitudinal measurement invariance with

[1] Asymmetric threshold spacing refers to the situation in which threshold levels of going from one response category to the next (e.g., from "I somewhat believe this" to "I very much believe this") are not distributed symmetrically around the mean of the latent response distribution, such that the peak of the distribution of the observed ordinal indicator falls to the left or right of the center (Rhemtulla et al., 2012).

ordered-categorical indicators. The organization of this manuscript is as follows. First, we briefly review foundational work on methods for establishing measurement invariance with continuous, multivariate-normally distributed indicators across groups and across time. Second, we briefly review foundational work on CFA approaches for ordered-categorical indicators, and explain how this can be extended to test longitudinal measurement invariance for ordered-categorical indicators. Third, we explicate the estimation of latent variable models (e.g., confirmatory factor analysis models; structural equation models) of ordered-categorical variables and provide comparisons of the current capabilities of three major latent variable programs (lavaan, M*plus*, OpenMx) for estimating these models. Fourth, we explicate the model specification and interpretation for each level of longitudinal measurement invariance with ordered-categorical indicators. Fifth, we consider practical problems in the evaluation of longitudinal measurement invariance using ordered-categorical CFA models, including data sparseness, missing data, and gauging the practical significance of *violations* of measurement invariance. Sixth, we illustrate many of these issues with an empirical example using a subscale from the Mexican American Cultural Values scale (Knight et al., 2010), and present computer scripts for addressing longitudinal measurement invariance. Finally, we conclude with a discussion with recommendations, limitations, and future directions. Our goal is to provide a clear didactic presentation of these procedures so applied researchers familiar with measurement invariance can use them in their own research. We limit our presentation to the most common longitudinal data structure in which participants are measured at fixed time points (measurement waves) and there is no group-level clustering (e.g., schools).

Testing Measurement Invariance with Continuous Variables

Multiple Groups

Following a procedure originally developed by Meredith (1993; see also Millsap, 2011; Widaman & Reise, 1997), measurement invariance in multiple groups with continuous indicators is established by sequentially testing a series of nested models in which constraints are added at each step of the hierarchy. Figure 1(a) depicts the basic measurement model used to test a hypothesized one-factor model with J indicators and $G = 2$ groups[2]. (1) The baseline *configural* invariance model tests the hypothesis that the same general pattern of factor loadings holds across groups. The corresponding factor loadings ($\lambda_j^{(1)}, \lambda_j^{(2)}$), intercepts ($\tau_j^{(1)}, \tau_j^{(2)}$), and unique factor variance-covariance matrix ($\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}$) are freely estimated (other than those constrained for model identification[3]) and can take on different values in the G groups. If the configural invariance model fits the data, subsequent models in the hierarchy are tested. (2) In the *weak* invariance model, the corresponding factor loadings are set to be equal ($\lambda_j^{(1)} = \lambda_j^{(2)}$). If the fit of the configural and weak invariance models do not differ, weak invariance is established. Weak invariance implies that the variances and covariances (when there are two or more latent common factors) of the latent common factors can be compared in the G groups (Widaman & Reise, 1997). (3) In the *strong* invariance model, the corresponding factor loadings and intercepts are set to be equal ($\lambda_j^{(1)} = \lambda_j^{(2)}, \tau_j^{(1)} = \tau_j^{(2)}$). If the fit of the weak invariance and strong invariance models do not differ, strong invariance is established. Achieving strong invariance implies that the means, variances, and covariances of the latent common factors can be compared

[2] For ease of presentation, the present work focuses on models with one latent common factor per group (for multiple group models) or per measurement occasion (for longitudinal models), although the results can be easily generalized to cases with more than one latent common factors.

[3] Two approaches can be used to identify the *variance* structure of the latent common factors: The marker variable approach gives the latent common factor a scale that is in the same unit as one of the indicators (the marker variable) by constraining the factor loading of this indicator to 1.0; the second approach sets the variance of the latent common factor to a fixed value, typically 1.0 (Bollen, 1989, p. 239). Two approaches can be used to identify the *mean* structure of the latent common factors: The intercept of the marker variable can be constrained to 0, or the mean of the latent common factor can be constrained to 0. In this present work, we use the marker variable approach to identify the variance structure, and constrain the latent common factor mean in one group (for multiple group models) or at one measurement occasion (for longitudinal models) to 0 to identify the mean structure.

in the *G* groups (Millsap, 2011; Widaman & Reise, 1997). Moreover, group differences in the *means* of the *measured indicators* are solely due to group differences in the latent common factor means (Millsap, 2011). (4) Finally, in the *strict* invariance model, the corresponding factor loadings, intercepts, and unique variances (and any nonzero unique factor covariances) are all set to be equal in the *G* groups ($\lambda_j^{(1)} = \lambda_j^{(2)}, \tau_j^{(1)} = \tau_j^{(2)}, \boldsymbol{\Theta}^{(1)} = \boldsymbol{\Theta}^{(2)}$). If the fit of the strong invariance and strict invariance models do not differ, strict invariance is established, which implies that group differences in the *means, variances,* and *covariances* of the *measured indicators* are solely due to group differences in the latent common factors (Millsap, 2011).

Longitudinal Measurement Invariance

The procedure for testing *longitudinal* measurement invariance with continuous indicators closely parallels testing measurement invariance with multiple groups (Meredith & Horn, 2001; Millsap, 2011; Millsap & Cham, 2012). Figure 1(b) depicts the basic measurement model used to test a hypothesized one-factor model with *J* indicators measured at 2 occasions. Three changes from Figure 1(a) are noteworthy: (1) the latent construct at Time 1 is allowed to co-vary with the same construct at Time 2; (2) the corresponding unique factors at Time 1 and Time 2 are allowed to co-vary; and (3) the second subscript now refers to the measurement wave *t*, whereas in Figure 1(a) the superscript (*g*) referred to the group.

Once again, a series of four hierarchical models may be tested. The *configural* invariance model tests the hypothesis that the same general pattern of factor loadings holds across time (Millsap & Cham, 2012). In this model, the corresponding factor loadings ($\lambda_{j1}, \lambda_{j2}$), the intercepts ($\tau_{j1}, \tau_{j2}$), and the unique factor variance-covariance matrix ($\boldsymbol{\Theta}_{11}, \boldsymbol{\Theta}_{22}$) at the two measurement occasions are freely estimated (other than those constrained for model identification). As before, the *weak* invariance model sets the corresponding factor loadings to be

equal over time ($\lambda_{j1} = \lambda_{j2}$), the *strong* invariance model sets the corresponding factor loadings and intercepts to be equal over time ($\lambda_{j1} = \lambda_{j2}$, $\tau_{j1} = \tau_{j2}$), and the *strict* invariance model sets the corresponding factor loadings, intercepts, and unique factor variances (and any non-zero within-wave unique factor covariances) equal over time ($\lambda_{j1} = \lambda_{j2}$, $\tau_{j1} = \tau_{j2}$, $\boldsymbol{\Theta}_{11} = \boldsymbol{\Theta}_{22}$). The interpretation of measurement invariance at each level in the hierarchy parallels that described above for the multiple group invariance case. With more than two measurement waves, additional constraints potentially could be put on the stability of the corresponding lagged covariances of the unique factors across time, which could be considered an additional (higher) level of longitudinal measurement invariance beyond strict invariance.

Ordered-Categorical CFA: Basic and Longitudinal Models

In this section we first review the basic ordered-categorical CFA model, which may not be familiar to applied researchers. We then consider its extension to the longitudinal model used to test measurement invariance over time. We focus our presentation on the case in which one factor is measured over *T* measurement occasions by *J* ordered-categorical indicators that have at least *three* response categories. When the observed indicators are *binary*, special constraints are required to identify the CFA model testing measurement invariance (see Millsap & Tein, 2004).

Basic Model

Let X_{ijt} be the observed ordered-categorical response from the i^{th} person on the j^{th} indicator at time *t*. Following Muthén (1984), we assume that all measured ordered-categorical indicators have score ranges {0, 1,..., *C*}, where $c = 0, 1, \ldots, C$ are the response categories of the observed responses. In the CFA model for ordered-categorical indicators, it is assumed that there are unobserved latent responses X^*_{ijt} that underlie each of the observed ordered-categorical responses X_{ijt}. The unobserved latent responses are assumed to be continuous and multivariate

normally distributed (Muthén, 1984), with a set of threshold parameters ν for each indicator that determine the observed ordered categorical responses, such that

$$X_{ijt} = c, \text{ if } \nu_{jtc} \leq X^*_{ijt} < \nu_{jt(c+1)}, \quad (1)$$

where $c = 0, 1, \ldots, C$ are the response categories of the observed responses, and $\{\nu_{jt0}, \nu_{jt1}, \ldots, \nu_{jt(C+1)}\}$ are the threshold parameters for observed ordered categorical variable *j* at measurement occasion *t* ($\nu_{jt0} = -\infty$, and $\nu_{jt(C+1)} = \infty$). As illustrated in Figure 2, the threshold parameters slice the underlying continuous latent responses into discrete values of the observed ordered-categorical responses. For a given latent response, when the threshold parameters are fixed, the corresponding observed response is completely determined.

The linear relationships between pairs of the underlying continuous latent responses are represented by polychoric correlations[4]. The factor model for the latent response X^* at time *t* is

$$X^*_{ijt} = \tau_{jt} + \lambda_{jt}\xi_{it} + u_{ijt}, \quad (2)$$

where τ_{jt} is the intercept, λ_{jt} is the factor loading of the continuous latent response *j* on the latent common factor at time *t*, ξ_{it} is the common factor score for person *i* at time *t*, and u_{ijt} is the unique factor score for the person *i* on the j^{th} indicator at time *t*. Typically, all latent intercepts τ_{jt} are fixed to zero to allow for the estimation of the latent threshold parameters.

Extension to Longitudinal Measurement Model

The basic model used to test longitudinal measurement invariance for ordered-categorical indicators is depicted in Figure 3. Consistent with previous work in longitudinal measurement invariance with continuous indicators (e.g., Millsap & Cham, 2012), we allow the common factors to freely covary across time, and have

[4] A polychoric correlation is the theoretical correlation between two bivariate normal, continuous latent responses X^*_{ijt} estimated based on the observed ordered-categorical responses. The best known special case is the tetrachoric correlation which is the estimate of this theoretical correlation based on two binary observed variables.

$$\xi_{it} \sim N(\boldsymbol{\kappa}, \boldsymbol{\Phi}),$$

$$\boldsymbol{\kappa} = [\kappa_1, \kappa_2, \ldots, \kappa_T]', \tag{3}$$

$$\boldsymbol{\Phi} = \begin{bmatrix} \varphi_1 & \varphi_{12} & \cdots & \varphi_{1T} \\ \varphi_{21} & \varphi_2 & \cdots & \varphi_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{T1} & \varphi_{T2} & \cdots & \varphi_T \end{bmatrix},$$

where the diagonal elements of $\boldsymbol{\Phi}$ represent the common factor variances at each occasion and off-diagonal elements representing lagged common factor covariances across time. We allow each unique factor to freely correlate with itself over time but *not* with other unique factors at other measurement occasions, and have

$$u_{ijt} \sim N(0, \boldsymbol{\Theta}), \tag{4}$$

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_{11} & \boldsymbol{\Theta}_{12} & \cdots & \boldsymbol{\Theta}_{1T} \\ \boldsymbol{\Theta}_{21} & \boldsymbol{\Theta}_{22} & \cdots & \boldsymbol{\Theta}_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Theta}_{T1} & \boldsymbol{\Theta}_{T2} & \cdots & \boldsymbol{\Theta}_{TT} \end{bmatrix},$$

where each diagonal element of the supermatrix $\boldsymbol{\Theta}$ is a submatrix $\boldsymbol{\Theta}_{tt}$ representing the unique factor variance-covariance matrix at measurement occasion *t*, and each off-diagonal element of $\boldsymbol{\Theta}$ is a diagonal submatrix $\boldsymbol{\Theta}_{t,t+k}$, with its diagonal elements representing the lagged covariances of each unique factor with itself over time.

Estimation of Latent Variable Models with Ordered-Categorical Variables

In this section we consider estimation issues in some detail. Different statistical packages utilize different estimation methods, model comparison tests, and treatment of missing data; therefore, they can produce different results. Four estimators are commonly used with latent variable models with ordered-categorical variables: Weighted Least Squares (WLS), Diagonally Weighted Least Squares (DWLS), Unweighted Least Squares (ULS), and Maximum Likelihood (ML). WLS, DWLS, and ULS are limited information methods, whereas ML is a full

information method. Assuming multivariate normally distributed latent responses, limited information methods use only the univariate and bivariate information -- the observed univariate marginal frequencies and observed bivariate frequencies of the ordered-categorical indicators (Forero & Maydeu-Olivares, 2009; Maydeu-Olivares & Joe, 2005). In contrast, the full information methods use all of the information in the data -- the observed multivariate frequencies of the ordered categorical indicators. According to asymptotic statistical theory, full information methods should produce parameter estimates with greater efficiency (Joe & Maydeu-Olivares, 2010; Maydeu-Olivares & Joe, 2005). In practice, however, the difference between full information and limited information methods is negligible given sufficient sample size (Forero & Maydeu-Olivares, 2009). Moreover, as the number of observed indicators increases, computing the probabilities of the observed responses based on the multivariate normal distribution of the latent responses becomes difficult.

Limited information approach. The general limited information approach for estimating ordered-categorical CFA models involves multiple stages of estimation (Browne & Arminger, 1995; Flora & Curran, 2004; Millsap, 2011; Muthén, 1984). In Stage 1, the observed univariate marginal frequencies of each ordered categorical indicator are used to provide maximum likelihood estimates of the standardized threshold parameters. In Stage 2, conditional maximum likelihood estimates of the polychoric correlations between each pair of indicators are calculated based on estimates of the standardized threshold parameters from Stage 1 and the observed bivariate frequencies of the ordered-categorical indicators. In Stage 3, estimates of the other model parameters (including the unstandardized threshold parameters; Millsap, 2011, p. 134) are obtained based on estimates of the standardized threshold parameters from Stage 1, their estimated large-sample covariance matrix (Muthén, 1984), and estimates of the polychoric

correlations from Stage 2 and their estimated large-sample covariance matrix (Browne & Arminger, 1995; Muthén, 1984; Muthén & Satorra, 1995). Model identification constraints and any other desired constraints are imposed at Stage 3 (Millsap, 2011, p. 134).

In Stage 3 of estimation, in order to obtain estimates of model parameters, one of the estimators from the generalized least squares (GLS) family may be used. The general form of the fit function that these methods attempt to minimize is:

$$F_{GLS} = [\hat{\sigma} - \sigma(\theta)]' W^{-1} [\hat{\sigma} - \sigma(\theta)], \quad (5)$$

where $\hat{\sigma}$ is the vector containing the estimated standardized thresholds from Stage 1 and the estimated polychoric correlations from Stage 2; $\sigma(\theta)$ is the vector containing the model-implied standardized thresholds and polychoric correlations, as functions of the unknown model parameters θ; and W, the weight matrix, is a square matrix. Different weight matrices may be used. The weight matrix W must be inverted and simple forms of W facilitate computation at a cost of asymptotic efficiency of estimates. Simple forms of W are typically preferred when sample size is small, when there are a large number of observed ordered-categorical variables, or when both of these conditions exist. The estimation procedures are presented below in order of decreasing complexity of W and hence decreasing computational burden.

1. Weighted Least Squares (WLS). The weight matrix W contains the large-sample estimates of the variances and covariances of the standardized thresholds and polychoric correlations. This weight matrix contains all of the information, and is the correct weight matrix to use. However, it can become very large quickly as the number of variables in the model increases. Thus when feasible (large sample sizes; small number of indicators; indicators with more response categories), WLS provides optimal estimates. Otherwise, WLS may be computationally infeasible (West, Finch, & Curran, 1995) or prone to non-convergence,

improper solutions, inflated goodness-of-fit test statistics, and biased estimates of factor correlations and factor loadings (Flora & Curran, 2004). This estimator is implemented in M*plus* and the lavaan package, denoted as WLS in both packages.

2. Diagonally Weighted Least Squares (DWLS). DWLS keeps only the diagonal elements of the weight matrix used in WLS estimation, which are the large-sample estimates of the variances of the standardized thresholds and polychoric correlations; all other elements are set to 0. This diagonal weight matrix is easy to invert, thus reducing the computational demands (Millsap, 2011, p. 135). DWLS is *not* asymptotically efficient as its weight matrix contains less information. As a result, default standard errors are no longer accurate, and the goodness-of-fit test statistic does not have a χ^2 distribution. Two different robust adjustment procedures, mean-adjusted weighted least squares and mean- and variance-adjusted weighted least squares have been proposed to overcome these problems (Muthén, du Toit, & Spisic, 1997). These two methods produce identical parameter estimates and standard errors, but their adjustments to the χ^2 goodness-of-fit statistic differ (mean-adjusted versus mean- and variance-adjusted). The mean-adjustment procedure is denoted WLSM and the mean- and variance-adjustment procedure is denoted WLSMV[5] in M*plus*. The lavaan package (Rosseel, 2012) also implements DWLS with the mean- and variance-adjustment procedure, denoted as WLSMV.

3. Unweighted Least Squares (ULS). ULS uses an identity matrix as the weight matrix W (Muthén, 1993). ULS makes no assumption about the distribution of the observed indicators as long as the model is identified (Bollen, 1989, p.112); it is often recommended for categorical indicators at small to medium sample sizes (e.g., Forero, Maydeu-Olivares, & Gallardo-Pujol,

[5] The original version of WLSMV in M*plus* estimated the model degrees of freedom from the sample, such that from sample to sample, "the degrees for freedom may vary within a given model specification" (Flora & Curran, 2004). A new version of the mean- and variance-adjustment procedure, which does not involve an adjustment for degrees of freedom and has Type 1 error rates very similar to the original version, is now the default in M*plus* when WLSMV is invoked (Asparouhov & Muthén, 2010a).

2009; Rhemtulla et al., 2012). ULS is *not* asymptotically efficient as its weight matrix contains less information, and adjustments to improve the standard errors and the goodness-of-fit test statistic are provided in Muthén (1993) and Asparouhov and Muthén (2010a). ULS with robust standard errors and the mean- and variance-adjusted goodness-of-fit test statistic for ordered-categorical variables is implemented in M*plus*, denoted as ULSMV[6], but is not currently implemented in lavaan.

Full information approach. Maximum Likelihood is a full information estimation approach. Currently ML estimation can be implemented under either the CFA framework for ordered-categorical indicators or the item response theory (IRT) framework.

1. CFA framework for ordered-categorical indicators. Within the CFA framework for ordered-categorical indicators, ML estimation generates expected covariances and means of the latent responses based on the observed *multivariate* frequencies of the ordered-categorical indicators. A numerical integration method is then used for integration over a multivariate normal distribution defined by these covariances and means to maximize the likelihood function and obtain estimates of model parameters (Boker et al., 2014; Wirth & Edwards, 2007). ML estimation of CFA models for ordered-categorical indicators is computer intensive and requires large sample sizes and a relatively small number of observed indicators (e.g., a maximum of 20 are permitted in OpenMx, Boker, et al., 2014; see Jöreskog & Moustaki, 2001; Wirth & Edwards, 2007). It is currently implemented in OpenMx, but not M*plus* or lavaan.

2. IRT framework. Within the IRT framework, the graded response model (Samejima, 1969) is often used to handle ordered-categorical indicators. The normal-ogive version of the graded response model, which uses the probit link function, is isomorphically equivalent to the

[6] The default ULSMV method in M*plus* now does not involve an adjustment for degrees of freedom, but a ULSMV method that involves an adjustment for degrees of freedom can be invoked by using the command Satterthwaite=ON (Asparouhov & Muthén, 2010a).

CFA model for ordered-categorical indicators depicted in Figure 3 (Takane & de Leeuw, 1987). The normal-ogive version of the graded response model attempts to estimate all parameters in one step (Jöreskog & Moustaki, 2001). The estimation methods of this approach generally make use of all the information in the data, and are considered full information methods. In contrast to full information ML estimation under the CFA framework, these estimation methods involve integration (marginalization) over the latent common factors (i.e., person specific values on the latent constructs) rather than indicators (Wirth & Edwards, 2007). Among the most commonly used estimation methods within the IRT framework are the marginal maximum likelihood (MML) methods (e.g., MML, MML with the EM algorithm, MML with adaptive numerical integration; Wirth & Edwards, 2007). MML is what M*plus* uses for ordered-categorical indicators when estimator=ML is used, and is also implemented in lavaan (estimator = "MML").

Comparison of estimators. According to simulation studies, WLS estimation generally performs adequately only at large sample sizes with a small number of latent common factors and small number of measured ordered categorical variables (e.g., N = 1000 in Flora & Curran, 2004 for a two-factor model with 10 indicators). Otherwise, WLS is prone to non-convergence, improper solutions, inflated goodness-of-fit test statistics, and biased estimates of factor correlations and factor loadings. Although it has received less attention, full information ML within the CFA framework for ordered-categorical indicators can be expected to show similar performance to WLS given the need to invert a complex weight matrix (the matrix of the model implied values; Browne & Arminger, 1995) in the estimation process. In contrast, MML, DWLS, and ULS have been shown to provide similar results and work well at more modest sample sizes of about 500, especially with indicators that are *not* highly skewed (Flora & Curran, 2004; Forero & Maydeu-Olivares, 2009; Forero et al., 2009; Rhemtulla et al., 2012): They all provide

accurate estimates of the factor loadings[7] (Forero & Maydeu-Olivares, 2009; Forero et al., 2009), and goodness-of-fit statistics with proper Type 1 error rates and sufficient statistical power (Forero & Maydeu-Olivares, 2009; Savalei & Rhemtulla, 2013). At smaller sample sizes of 200 or less, none of these methods have satisfactory performance; they are all prone to non-convergence, improper solutions, large standard errors, biased estimates of factor loadings and thresholds, and problematic goodness-of-fit tests (Forero & Maydeu-Olivares, 2009; Forero et al., 2009; Savalei & Rhemtulla, 2013).

Software implementation. Table 1 summarizes several of the key features of three major software packages (M*plus*, lavaan, and OpenMx)[8] with respect to the different estimators, parameterizations, convergence time, and restrictions on the number of observed indicators. ULS with robust correction for ordered-categorical variables is implemented in M*plus*. DWLS methods with robust corrections are implemented in M*plus* and lavaan. OpenMx[9] implements full information ML estimation within the CFA framework for ordered-categorical indicators, whereas M*plus* and lavaan implement the MML estimation within the IRT framework[10]. Missing data and model fit are discussed below.

Testing Longitudinal Measurement Invariance with Ordered-Categorical Indicators

Our goal is to develop comparisons for longitudinal models with ordered-categorical indicators that parallel, as closely as possible, the comparisons described above that are used for

[7] MML and ULS also tend to provide unbiased estimates of the thresholds when the sample size is 500 and above (Forero & Maydeu-Olivares, 2009). Less, if any, work has been done to investigate the estimation accuracy of the thresholds using DWLS.

[8] The versions of the software packages compared in this paper are: lavaan 0.5-17, M*plus* 7.11, and OpenMx 2.0.0.3838.

[9] WLS, ULS, or DWLS were not implemented in OpenMx (version 2.0.0.3838) when we conducted this work. The latest release of OpenMx, version 2.3.1, implements WLS, DWLS, and ULS. To make use of these estimators in OpenMx (version 2.3.1), the user needs to manually specify the matrices involved in the fit function (see Equation (5)).

[10] In M*plus*, (a) using MML with the probit link for ordered-categorical indicators invokes the normal-ogive version of the graded response model, and (b) using MML with the logit link for ordered-categorical indicators invokes another version of the graded response model. In lavaan, only MML with the probit link is available.

models with continuous indicators (configural vs. weak; weak vs. strong; strong vs. strict invariance). The mathematical development supporting the interpretation of each level of measurement invariance for ordered-categorical indicators is presented in Appendix A.

Model 1. The baseline model

The baseline model tests the hypothesis that the same general pattern of factor loadings holds across time. This baseline model should provide a good fit to the data in order to continue evaluation of loading (weak), threshold (strong), and unique factor (strict) invariance models.

Neither the latent common factors nor the underlying latent responses X^* have inherent scales (i.e., unit of measurement); constraints on parameters must be imposed to identify the scales of latent common factors and the underlying latent responses. The following constraints serve to identify this baseline model (adapted from Millsap & Tein, 2004):

1. At all measurement occasions, the latent intercepts $\boldsymbol{\tau}_t$ are fixed to zero.
2. At one measurement occasion (the reference measurement occasion, typically the first or last), the common factor mean κ_t is constrained to zero, and the unique factor covariance matrix $\boldsymbol{\Theta}_{tt}$ is constrained to be $\boldsymbol{\Theta}_{tt} = \boldsymbol{I}$, the identity matrix[11]. At all other measurement occasions, the unique factor covariance matrix $\boldsymbol{\Theta}_{tt}$ is a diagonal matrix with the diagonal elements freely estimated.
3. At all measurement occasions, the same observed measure is chosen as the marker variable, and the factor loading of the marker variable is constrained to be 1.00.
4. One threshold for each indicator (and a *second* threshold for the marker variable) is constrained to be equal across measurement occasions.

[11] Alternatively, the total variances of all latent responses at the reference measurement occasion can be constrained to 1.0, instead of constraining the unique variances to 1.00 (adapted from Millsap & Tein, 2004).

This model identification strategy makes it possible to freely estimate the unique factor variances at occasions other than the reference occasion, and allows for the estimation of a model representing the invariance of the unique factor variances (discussed below) that parallels the strict invariance model in the continuous case. This unique factor invariance model can be implemented using the WLS and DWLS estimators with the theta parameterization[12] in M*plus* or lavaan, using the ULS estimator with the theta parameterization in M*plus*[13], or using the ML estimator in OpenMx. However, it *cannot* be implemented in M*plus* or lavaan using the MML estimator with the probit link, as the unique scores in probit regressions are constrained to follow a standard normal distribution. We discuss practical considerations in choosing marker variables and in specifying constraints for factor loadings and thresholds, as well as specific issues involved in fitting the model when applied to empirical data in the *Empirical Illustration* section.

Model 2. The loading invariance model

The baseline model is compared to the loading invariance model that adds the constraint that factor loadings are identical across time: $\lambda_{11} = \lambda_{12} = \cdots = \lambda_{1T}$, $\lambda_{21} = \lambda_{22} = \cdots = \lambda_{2T}$, $\lambda_{31} = \lambda_{32} = \cdots = \lambda_{3T}$, …. The first subscript represents the latent response underlying the observed indicator and the second subscript represents time. The loading invariance model parallels the weak invariance model for continuous indicators. Of importance, since all intercepts are fixed to zero for model identification, when loading invariance holds for ordered-categorical indicators, changes over time in the expected *means* of the *continuous latent responses*

[12] For ordered-categorical CFA models, two parameterizations are available: delta and theta (Muthén & Asparouhov, 2002). These two parameterizations have a scaling factor for each indicator. Using the delta parameterization, the scaling factors for each indicator are the inverse of the standard deviations of the underlying latent responses, and are treated as model parameters; in contrast, the unique variances are not treated as model parameters, but instead are calculated as a function of the model explained variances of the latent responses and the scaling factors. The theta parameterization considers the unique variances as model parameters, and calculates the scaling factors as a function of the model explained variances of the latent responses and the unique variances. Thus, using the theta parameterization, invariance constraints can be directly put on the unique variances.

[13] Currently ULS for ordered-categorical indicators is not implemented in lavaan.

underlying the measured ordered-categorical indicators are entirely attributable to changes in the latent common factors over time (see Appendix A). However, since the continuous latent responses are not observed but are instead inferred based on distributional assumptions (multivariate normality), this condition is not sufficient to guarantee invariance of the observed responses.

Model 3. The threshold invariance model

The loading invariance model is compared to the threshold invariance model that adds the constraint that for each indicator, the threshold level of going from one response category to the next is identical over time: $\nu_{111} = \nu_{121} = \cdots = \nu_{1T1}$, $\nu_{112} = \nu_{122} = \cdots = \nu_{1T2}$, $\nu_{113} = \nu_{123} = \cdots = \nu_{1T3}$, …, with the first subscript representing the latent response underlying the observed indicator, the second subscript representing time, and the third subscript representing threshold. The threshold invariance model parallels the strong invariance model for continuous indicators. Of importance, unlike the continuous case, having invariant factor loadings and invariant thresholds does not translate into the nice situation in which changes in the means of the *measured* ordered-categorical indicators are solely attributed to changes in the latent common factor. To achieve this, the unique factor variance must also be invariant over time (see Appendix A).

Model 4. The unique factor invariance model

The threshold invariance model is compared to the unique factor invariance model that adds the constraint that the corresponding elements in $\boldsymbol{\Theta}_{tt}$ (all unique factor variances[14]) are equal over time. The lagged unique factor covariances over time (diagonal elements in $\boldsymbol{\Theta}_{t,t+k}$) are freely estimated. This unique factor invariance model is estimable when the unique variances

[14] Within-wave unique factor covariances are normally assumed to be 0. If not, the corresponding within wave factor covariances should be constrained to be equal at each measurement wave.

of the latent responses are freely estimated in the earlier models. Recall that unique variances at the reference occasion were fixed to 1.0 for model identification, so in this unique factor invariance model, all unique variances are fixed to 1.0.

The unique factor invariance model parallels the strict invariance model for continuous indicators. When longitudinal unique factor invariance is achieved, changes over time in the expected *means, variances,* and *within-wave covariances* of the *continuous latent responses* are entirely attributable to changes in the latent common factors over time. Moreover, changes in the *within-wave characteristics* of the *measured ordered-categorical indicators* over time are entirely attributable to changes in the latent common factors over time. The mathematical development supporting these conclusions is presented in Appendix A.

Model fit evaluation

Once the fit of the baseline model is established, nested model tests should be used to compare the fit of the less restricted model to the next model in the hierarchy. With ML estimation, the likelihood ratio test (Bentler & Bonett, 1980) or its robust version (Satorra & Bentler, 2001; 2010) may be used for model comparison. With DWLS or ULS estimation with robust corrections in M*plus*, the DIFFTEST function (Asparouhov & Muthén, 2006) provides a proper evaluation of the difference between nested ordered-categorical CFA models. Using DWLS estimation with robust correction in lavaan, a difference test akin to the DIFFTEST in M*plus* can be requested. Previous research suggests that the M*plus* DIFFTEST results from ordered-categorical CFA models assessing multiple-group measurement invariance using DWLS may exhibit inflated Type 1 error rates (Sass, Schmitt, & Marsh, 2014). Thus, it is important to also examine local fit indices (residuals; modification indices; cf McDonald & Ho, 2002).

Some authors (e.g., Chen, 2007; Cheung & Rensvold, 2002) have suggested that changes in practical fit indices like the RMSEA and CFI might potentially be useful indices for the comparison of nested models. However, these indices have not been systematically studied in the context of ordered-categorical CFA models. In the only study to date, Sass et al. (2014) found that changes in practical fit indices performed poorly with DWLS estimation with mean- and variance-adjustments to the goodness-of-fit test statistic, especially for misspecified models. At this point, the use of changes in practical fit indices cannot be recommended for the evaluation ordered-categorical CFA models.

Three Common Issues That Arise in Practice

Sparse data problem

If different numbers of categories are observed at each measurement point (i.e., the number of categories C+1 changes over time), problems are created in specifying invariance constraints on the threshold parameters. The thresholds that statistical software packages recognize are based on the *observed* response categories at each occasion, *not* the scaling of the measurement instrument. Take the example of a five-point Likert scale indicator with response categories 1 through 5. Suppose in a longitudinal study participants endorse response categories 1 through 5 at Time 1 but only response categories 2 through 5 at Time 2. At Time 1 the first threshold that statistical software packages recognize is the threshold between response categories 1 and 2, whereas at Time 2 the first threshold that statistical software packages could recognize is the threshold between response categories 2 and 3. Thus, to correctly specify invariance constraints on the threshold parameters, the second observed threshold at Time 1 should be constrained to be equal to the first observed threshold at Time 2 (between the response categories 2 and 3 on the Likert scale, Grimm, Ram, & Estabrook, in press). Depending on how

often sparse data occur, this could be a tedious process. Moreover, with sparse data there may be low or zero expected cell frequencies in the observed contingency table, which could lead to inaccurate estimation of the polychoric correlations (Brown & Bendetti, 1977; Olsson, 1979), which in turn may lead to inaccurate parameter estimates in the CFA models (Flora & Curran, 2004). An alternative approach to specifying equality constraints on the threshold parameters in the face of the sparse data is to collapse some of the adjacent sparse response categories. After collapsing the adjacent sparse response categories, all indicators would have the same number of observed response categories at each of the measurement occasions, so that the specification of equality constraints on the threshold parameters would be straightforward.

Missing data

Most longitudinal studies have missing data. Missing data theory (Rubin, 1976) distinguishes between data that are missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). MCAR makes the strongest assumption--assuming that data are missing due to completely random reasons and that the missing data patterns are independent of the observed data. MAR assumes that the probability of missing data on a variable is related to observed data on other analysis variables, but is unrelated to the would-be values of the variable itself. MNAR assumes that the missing data are related to the unobserved, would-be values of the variable. Under the strong MCAR assumption, unbiased estimates of all model parameters and tests of model fit can be achieved through complete case analysis (listwise deletion) or available case analysis (pairwise deletion), but statistical power of the nested model tests will be reduced. Using either the complete or available case approach under the weaker MAR assumption can lead to biased parameter estimates and tests of model fit.

Improved estimates of model parameters and tests of model fit under the weaker MAR assumption can be achieved in three ways. First, full information maximum likelihood estimation (available in OpenMx) or marginal maximum likelihood estimation (Mislevy, in press; available in M*plus*[15]) adjusts estimates based on all of the observed data in the measurement model. However, when using ML or MML, common fit statistics (e.g., χ^2 goodness-of-fit statistic, CFI, RMSEA) are not readily available, although researchers can calculate them manually by fitting a separate null or saturated model. Second, for both the full information and the limited information estimators, auxiliary variables, variables that are not part of the measurement model but that have a substantial correlation with both missingness and the variables in the measurement model (Graham, 2003; 2009), can be added to the model. This approach adjusts estimates for the observed values of the auxiliary variables. M*plus* (e.g., Asparouhov & Muthén, 2010b), lavaan, and OpenMx all permit estimation of models with auxiliary variables, although auxiliary variables will influence the model fit statistics. Third, multiple imputation can be performed followed by analyses using ML/ MML, WLS, DWLS or ULS (Asparouhov & Muthén, 2010c). Using this approach, the parameter estimates and standard errors may be combined across multiple imputed data sets following Rubin's (1987) procedure. However, to our knowledge no published article addresses the mechanism of pooling the overall model fit statistics like CFI, RMSEA, and the χ^2 goodness-of-fit statistic across multiple imputed data sets, which is critical to the evaluation of longitudinal measurement invariance.

Gauging the Practical Significance of the Violations of Invariance

If the statistical tests of measurement invariance fail to be supported for one or more of the four models, it is useful to have a method with which to gauge the *practical significance* of

[15] Although MML is implemented in lavaan, missing data handling using MML is not available in the current version (0.5-17). Lavaan uses listwise deletion with MML.

the differences in measurement model parameters over time (i.e. whether the differences have practical implications, Kirk, 1996). The primary concern for researchers is whether the mean/variance change in the observed indicators between measurement occasions is due to true change in the mean/variance of the latent construct, or an artifact of the different values of the parameters in the measurement model across different measurement occasions. We propose that the practical significance of the differences in measurement model parameters over time can be evaluated through sensitivity analyses that compare the model-predicted probabilities of choosing specific response categories at specific measurement occasions, which were calculated based on measurement models with different levels of invariance constraints.

Consider the equation for the mean structure of the latent responses at time *t*

$$\boldsymbol{\mu}_{X_t^*} = \boldsymbol{\tau}_t + \boldsymbol{\Lambda}_t \kappa_t, \tag{6}$$

where $\boldsymbol{\mu}_{X_t^*}$ is a $J \times 1$ vector of the population means of the *J* latent responses at time *t*, $\boldsymbol{\tau}_t$ is a vector of the latent intercepts at time *t*, $\boldsymbol{\Lambda}_t$ is the factor loading vector at time *t*, and κ_t is the latent common factor mean at time *t*. As mentioned earlier, all latent intercepts are fixed to zero to identify the model, so Equation (8) can be simplified to

$$\boldsymbol{\mu}_{X_t^*} = \boldsymbol{\Lambda}_t \kappa_t. \tag{7}$$

Then consider the equation for the population variance-covariance matrix of the *latent* responses at time *t*:

$$\boldsymbol{\Sigma}_{X_t^* X_t^*} = \boldsymbol{\Lambda}_t \varphi_t \boldsymbol{\Lambda}_t' + \boldsymbol{\Theta}_{tt}, \tag{8}$$

where φ_t is the common factor variance at time *t*, and $\boldsymbol{\Theta}_{tt}$ is the unique factor covariance matrix at time *t*. All the elements on the right-hand side of Equations (9) and (10), $\boldsymbol{\Lambda}_t$, κ_t, φ_t, and $\boldsymbol{\Theta}_{tt}$, can be obtained from the ordered-categorical CFA output, and thus the model-predicted means and variances of the latent responses at each time can be calculated. Assuming a normal

distribution of the latent responses, the model-predicted probabilities of choosing each response category on each indicator at each time point can be calculated based on the estimated thresholds (which are also obtained from the ordered-categorical CFA output). An R function (presented in Supplemental Material 1A) was developed to calculate the estimated probabilities of choosing each response category for each indicator. An example of applying the R function is illustrated in the *Empirical Illustration* section and in Supplemental Material 1B.

When loading invariance (Model 2) holds but threshold invariance (Model 3) does not, the researcher can use this R function to calculate the model-predicted probabilities based on the loading invariance model results and the model-predicted probabilities based on the threshold invariance model results. Differences in the model-predicted probabilities from these two models can serve as an estimate of the effect size of the violation of longitudinal measurement invariance at the threshold level. Likewise, when the baseline model (Model 1) holds but the loading invariance model (Model 2) does not, differences in the model-predicted probabilities from these two models can serve as an estimate of the effect size of the violation of longitudinal measurement invariance at the factor loading level. If the threshold invariance model (Model 3) holds, but the unique factor invariance model (Model 4) does not, differences in the model-predicted probabilities from these two models can serve as an estimate of the effect size of the violation of longitudinal measurement invariance at the unique factor level.

Empirical Illustration

Background

Many theories identify enculturation processes, or changes in Mexican American youths' heritage cultural values, as an important mechanism through which later educational and psychosocial outcomes are influenced. However, few researchers have directly examined how

Mexican American youths' heritage cultural values change over time. To examine such changes in heritage cultural values, it is important to first evaluate the measurement invariance of the measures of the heritage cultural values across time. Without an investigation of longitudinal measurement invariance, it is difficult to understand the degree to which changes over time in the observed scores of a measure of Mexican American youths' heritage cultural values can be attributed to *true* changes in the latent constructs of interest. The current empirical illustration highlights the utility of the methods to evaluate both longitudinal measurement invariance of ordered-categorical variables and violations of such assumptions on real data drawn from an on-going longitudinal study examining Mexican American youths and their families (Roosa et al., 2008).

Sample

The data used in this illustration come from the first four measurement occasions (5^{th}, 7^{th}, 10^{th}, and 12^{th} grades) of an on-going longitudinal study examining the changes in cultural values, contexts, and mental health problems of Mexican American youths and their families (Roosa et al., 2008). The analyses were based on a sample of 749 Mexican American youths.

Measures

Our empirical illustration will focus on one of the subscales of the Mexican American Cultural Values Scale (MACS; Knight et al., 2010), familism obligations, which was measured by five indicators. Youths were instructed to indicate their degree of agreement with each indicator by responding to a 5-point Likert-type scale (1 = not at all to 5 = very much).

Sparse Data Problem

In this sample, Mexican American youth typically endorsed high levels on each indicator on the familism obligations subscale. One indicator reads, "If a relative is having a hard time

financially, one should help him or her out if possible." At Time 1, only one of 749 youths responded 1 (not at all), and only 14 responded 2 (a little), with the rest of the youths responding 3 (somewhat) or higher. At Time 3, no youth responded 1, and only four responded 2. As mentioned above, the thresholds that statistical software packages recognize are based on the *observed* response categories at each occasion. When different numbers of categories are observed at each measurement point (i.e., the number of categories C +1 changes over time), problems are created in specifying invariance constraints on the threshold parameters. We merged categories 1 and 2 for all of the indicators. Each indicator in the familism obligations subscale then had four observed response categories at each of the four measurement occasions, making specification of equality constraints on the threshold parameters straightforward.

Analyses

Item responses were rescaled from (1 or 2 collapsed), 3, 4, and 5 to 0, 1, 2, and 3, respectively, so that the lowest response category was zero in accordance with the notation we have used in this paper. The code used to test longitudinal baseline, loading invariance, threshold invariance, and unique factor invariance models of familism obligations is presented in Supplemental Materials 2A – 2D (using DWLS[16] in M*plus*), 3A – 3D (using DWLS in lavaan), and 4A – 4D (using ML in OpenMx), respectively. The corresponding factors and the unique factor scores for each indicator were allowed to correlate across measurement occasions. The constraints for scaling purposes described earlier were imposed. Specific issues regarding the choice of thresholds to be constrained, the choice of marker variables, and the calculation of CFI are discussed below.

[16] The ULS estimator with robust correction in M*plus* failed to produce converged and proper solutions for the baseline, loading invariance, and threshold invariance models in this data set. The M*plus* code using the ULS estimator with robust correction is the same with the M*plus* code using the DWLS estimator with robust correction with one exception: The estimator is ULSMV instead of WLSMV.

Choosing the thresholds to be constrained for model identification. As discussed, the model identification strategy requires that one threshold per indicator and a second threshold for the marker variable be constrained to be equal across measurement occasions to identify the latent responses. Note that this strategy makes the assumption that the constrained thresholds are invariant in the population. If a chosen threshold is *not* invariant in the population, then the equality constraint placed on this chosen threshold might result in a misfit of the model to the data. A true violation of measurement invariance at the threshold level can mistakenly lead to the conclusion that the violation is at the factor loading level due to a poor selection of the thresholds to constrain for the identification of the latent responses. A modification index can be calculated for each constrained parameter in the loading invariance model, which can provide useful information with which to diagnose this problem. In this context, if the loading invariance model is rejected, and if the following examination of the modification indices reveals a constrained threshold with a high modification index, then constraining the equality of a different threshold over time should be considered for model identification (Yoon & Millsap, 2007). Since estimates of thresholds from categories with sparse data tend to be less stable, constraining such thresholds to identify the model typically is *not* a good choice.

Choosing the marker variable. A marker variable should have a meaningful metric, or be an indicator of the latent common factor with a high factor loading. For evaluating longitudinal measurement invariance, it is crucial to choose a marker variable whose loading is invariant at all occasions[17]. The model identification strategy requires that two of the thresholds for the marker variable be constrained to be invariant across measurement occasions. Therefore, the marker

[17] If loading invariance does not hold for any one indicator so that no appropriate marker variable exists, then the researcher can conclude that loading invariance does not hold.

variable should not only have an invariant factor loading across all measurement occasions, but also have at least two invariant thresholds[18].

Caution in calculation of CFI. As a relative fit index that assesses goodness of fit, CFI measures the improvement of model fit relative to a baseline (null) model (Bentler, 1990), which should be nested within the most restricted model of interest. In the continuous indicator case, the null model for testing longitudinal measurement invariance used by standard statistical packages to calculate the CFI (and other relative fit indices like the TLI, Tucker & Lewis, 1973) places no constraint on the mean structure or the unique factor structure, and becomes an inappropriate comparison model (cf. Widaman & Thompson, 2003; Wu, West, and Taylor, 2009). Supplemental Materials 5A and 5B adapt this work to the ordered-categorical indicator case, providing the M*plus* and lavaan syntax for specifying an appropriate alternative null model for the test of longitudinal measurement invariance[19]. Supplemental Material 5C provides a SAS macro adapted from Wu et al. (2009) for calculating the corrected CFI.

Calculating probabilities. Supplemental Material 1A contains an R function we created to calculate the estimated probabilities for endorsing each response category for each indicator. Supplemental Material 1B contains the R code for calculating the discrepancies in the estimated probabilities at each occasion between the loading and threshold invariance models. When loading invariance is retained but threshold invariance is rejected, discrepancies in the estimated probabilities between these two models can serve as an estimate of the magnitude of the effect of violating longitudinal measurement invariance at the threshold level. Likewise, when the baseline model holds but loading invariance does not (or when the threshold invariance model

[18] If no indicator has two threshold parameters that are invariant across time, then the researcher can conclude that threshold invariance does not hold.

[19] In addition to constraining all the thresholds to be invariant across occasions, this alternative null model also constrains all unique variances at all measurement occasions to be 1.0 and all within-wave and lagged unique factor covariances to zero. Thus, this null model is also appropriate for the unique variance invariance model.

holds but the unique factor invariance model does not), differences in the predicted probabilities between these two models can serve as an estimate of the effect size of the violation of longitudinal measurement invariance at the factor loading level (or the unique factor level).

The R function in Supplemental Material 1A calculates the model predicted probabilities of endorsing each response category on each indicator at each measurement occasion, based on the assumption of normal distributions of the latent responses underlying the ordered-categorical measures, and the following parameter estimates from a one-factor ordered-categorical CFA model: 1) factor loadings; 2) latent common factor means; 3) latent common factor variances; 4) unique factor variances; and 5) thresholds. This R function has several restrictions: 1) The number of indicators and 2) the number of response categories for each indicator must be the same across measurement occasions; 3) the response categories of the indicators are scaled such that the lowest response category is 0; and 4) unique factors *within* a measurement occasion are uncorrelated. These restrictions are typically met.

Results

Table 2 presents the results from the longitudinal measurement invariance tests for the familism obligation subscale based on M*plus* output using DWLS[20]. The baseline model (Model 1) showed adequate fit. The CFI and RMSEA values for the loading invariance model (Model 2) also suggested adequate fit, and the DIFFTEST indicated that adding the loading invariance constraints did not significantly worsen the model fit when compared to the baseline model. The DIFFTEST comparing the threshold invariance model (Model 3) to the loading invariance model (Model 2), however, indicated that the threshold invariance model fit the data significantly worse than the loading invariance model. Given that the threshold invariance model (Model 3) did not

[20] Model solutions obtained using DWLS in M*plus*, DWLS in lavaan, and ML in OpenMx were similar for our analyses of this data set of $N = 749$.

hold, the unique factor invariance model (Model 4) was not tested for this data set and a sensitivity analysis of the practical significance of the failure of threshold invariance was conducted (see below). However, if the threshold invariance model were to hold, the unique factor invariance model needs to be tested. Of importance, with ordered-categorical indicators longitudinal unique factor invariance is needed to attribute changes in the means of the observed ordered-categorical indicators solely to changes in the latent common factor.

As noted earlier, the M*plus* DIFFTEST results from ordered-categorical CFA models using DWLS may exhibit inflated Type 1 error rates (Sass et al., 2014), thus it is important to also examine local fit indices, such as modification indices. Examination of modification indices suggested that constraining the thresholds of some indicators to be invariant across measurement occasions was problematic. Indicators V1, V2, and V5 for familial obligation did not appear to have invariant thresholds across time, with Times 1 and 3 different from Times 2 and 4. However, these results do not directly inform us about the magnitude (practical significance) of the violation of measurement invariance. We therefore conducted a sensitivity analysis to examine the model-based predicted probabilities of choosing each response for the familism obligations subscale, calculated using the R programs in Supplemental Materials 1A and 1B.

Tables 3 and 4 present the predicted probabilities of choosing each response category for each indicator of youth reported familism obligation at each measurement occasion, based on the *retained* loading invariance model (Model 2) and the *rejected* threshold invariance model (Model 3), respectively. Table 5 presents the *discrepancies* in the predicted probabilities between the threshold and loading invariance models. One indicator for familism obligations, “Parents should be willing to make great sacrifices to make sure their children have a better life”, (V5) had the largest discrepancies in predicted probabilities between the *retained* loading invariance model

(Model 2) and the *rejected* threshold invariance model (Model 3) at Time 3 (V5T3). The predicted probability of endorsing the response category "Very much agree" was estimated to be 0.490 in the accepted loading invariance model, but 0.381 when thresholds were also constrained to be invariant over time. The violation of threshold invariance led to differences in the predicted probabilities of choosing the response categories, with the greatest difference being -0.109 (=0.381-0.490). Given that our illustrative example is based on a sample size of 749 youths, discrepancies in the estimated probabilities that are small (e.g., < .05) may represent relatively few individuals in the sample (e.g., a .04 discrepancy in the estimated probabilities represents about 30 people out of 749), and the researcher might decide to ignore such discrepancies. For the familism obligations subscale which reached loading invariance but failed to achieve threshold invariance, adding threshold invariance constraints led to material differences in the estimated probabilities at Times 1 and 4 for V2, and at Time 3 for V5. Given the nature of this illustrative data set, it is possible that important changes in the nature of the constructs occurred between Time 1 (5^{th} grade) and Time 2 (7^{th} grade). ***Missing data***. In the present longitudinal data set, missing data (0%, 5.2%, 15.0%, and 19.3% at Times 1, 2, 3, and 4, respectively) were primarily due to participant attrition. We reported in Tables 2-5 results based on the default M*plus* procedure to estimate the ordered-categorical CFA models using DWLS, which makes the strong MCAR assumption and uses pairwise deletion (Asparouhov & Muthén, 2010b; Muthén & Muthén, 1998-2012). We performed additional analyses in which we identified potential auxiliary variables that were expected to be related to attrition (e.g., Time 1 mother employment status, youth gender, youth nativity [Mexico-born versus US-born], Time 1 self-identified ethnicity for youth [Mexican versus Mexican American], and Time 1 self-identified ethnicity for mother). We included these variables in the ordered-categorical CFA models for testing

longitudinal measurement invariance using WLSMV in M*plus*. Results of these additional analyses did not materially change the parameter estimates of the initial investigation of longitudinal measurement invariance without auxiliary variables, with most differences occurring at the third decimal place. These additional analyses suggested that the default approach to addressing missing data using DWLS with robust correction in M*plus* was likely not problematic in the present illustration. We report the uncorrected analyses, which also illustrate the calculation of corrected CFI values that are not influenced by auxiliary variables.

Discussion

Given the importance of longitudinal studies in understanding development, it is important to establish measurement invariance of key constructs. The structure of important constructs, such as abilities, personality traits, and psychopathology may change during childhood, adolescence, and adulthood. When the measures of constructs of interest are *not* invariant over time, any apparent change may reflect change in what is being measured rather than in the level of the latent construct of interest. Methods of testing for measurement invariance allow for the examination of this issue. CFA-based methods for testing measurement invariance for constructs measured with continuous indicators have been extensively developed for both cross sectional (Meredith, 1993; Widaman & Reise, 1997) and longitudinal data structures (e.g., Khoo et al., 2006; Meredith & Horn, 2001). However, these approaches are not optimal when ordered-categorical indicators are collected. Available evidence from simulation studies (e.g., Rhemtulla et al., 2012) suggests that the traditional approach of treating ordered-categorical indicators as continuous using standard procedures of testing measurement invariance may result in suboptimal parameter estimates, especially with a small number of response categories or skewed distributions of observed responses due to asymmetric threshold spacing.

The approach presented here builds on key foundational work on testing measurement invariance. For the case of testing measurement invariance between different groups (e.g., gender; language), Meredith (1993) and Widaman and Reise (1997) developed the basic set of hierarchical models, identification constraints, testing procedures, and the interpretation of the models. Meredith and Horn (2001) modified the specification originally developed for models for cross-sectional data to address the over-time covariance of the latent common and unique factors that characterize continuous longitudinal data. Finally, Millsap and Tein (2004) developed an appropriate specification that permitted the identification and testing of models for testing measurement invariance between different groups with ordered-categorical data. We drew on this work, combining the earlier insights to adapt existing models of measurement invariance to be appropriate for use with longitudinal ordered-categorical data. Our goal was to develop a hierarchical series of model tests that paralleled as closely as possible those used in tests of measurement invariance with continuous variables. We also wanted to highlight many of the subtle issues that arise in implementing these tests of measurement invariance in three commonly used statistical software packages (lavaan, M*plus*, and OpenMx).

In an earlier section we presented the specification and interpretation of four longitudinal measurement invariance models for ordered-categorical indicators that parallel the configural, weak, strong, and strict invariance models for continuous data. The baseline model for ordered-categorical data closely parallels the configural invariance model for longitudinal measurement models with continuous data in most respects (cf. Widaman & Reise, 1997). The loading invariance model also closely parallels the weak invariance model, but includes an interesting unique property: Changes over time in the expected *means* of the *continuous latent responses* are entirely attributable to changes in the latent common factor over time. Note however that the

continuous latent responses are not measured indicators, and they are inferred from the ordered-categorical indicators based on distributional assumptions. In the loading invariance model, changes in the *means* of the *measured ordered-categorical indicators* cannot be solely attributed to changes in the latent common factor. The threshold invariance model also parallels the strong invariance model in many aspects, but its interpretation differs in an important way. Unlike in the continuous case, having invariant factor loadings and invariant thresholds does *not* imply that changes in the *means* of the *measured ordered-categorical indicators* can be solely attributed to changes in the latent common factor (see Appendix A). To achieve this, a higher level of invariance termed unique factor invariance must be reached. The unique factor variances (which are equal to 1.0 in this model) and any non-zero within-wave unique factor covariances must also be invariant over time (see Appendix A). In this case, changes over time in the expected means, variances, and within-wave covariances of the *continuous latent responses* are entirely attributable to changes in the latent common factor over time, paralleling the strict invariance model with continuous measured variables. Moreover, changes in the *within-wave characteristics* of the *measured ordered-categorical indicators* over time are entirely attributable to changes in the latent common factor over time. Even if unique factor invariance is achieved, the expected lagged covariances of the *continuous latent responses* over time will not necessarily be invariant. A model with the further restriction that the lagged unique factor covariances are equal over time must be met to achieve this property.

We also considered three important practical issues in the evaluation of longitudinal measurement invariance. Sparse data, particularly in some of the high or low categories, can occur as children mature leading to complexities in equating thresholds. Missing data occur as participants drop out of longitudinal studies. The need to achieve a more stringent level of

measurement invariance (unique factor invariance) in order to compare the observed means over time places a strict requirement that will often not be met in practice. We developed a sensitivity analysis that allows researchers to assess the practical significance of the failure to achieve more advanced levels of measurement invariance.

Of practical importance, we considered three major statistical software packages that can be used to test measurement invariance for ordered-categorical indicators. Special attention was paid to available estimators and treatment of missing data within each package. WLS and ML are theoretically the estimators of choice given very large sample sizes ($N > 1000$) and a limited number of observed ordered-categorical variables. Some current implementations of MML do not permit easy estimation of the models necessary for testing longitudinal measurement invariance because correlated unique factors cannot be directly specified. Alternative specifications of the measurement invariance models must be used. ULS and DWLS appear to have good properties with more moderate sample sizes ($200 < N < 1000$), but default treatment of missing data is not optimal, with multiple imputation providing less biased tests and improved statistical power. Table 1 summarizes the features available in lavaan, M*plus*, and OpenMx.

Testing measurement invariance using ordered-categorical CFA models is dependent on some key distributional assumptions. Ordered-categorical CFA models assume that the *continuous latent responses* underlying the observed ordered-categorical indicators are multivariate-normally distributed (Muthén, 1984), and the relationships estimated by these models are between the latent, multivariate-normally distributed responses (West et al., 1995). If one or more of the true latent variables (e.g., depression) follows a distribution that is markedly non-normal, the parameter estimates and standard errors from ordered-categorical CFA models may be in a transformed metric (of normal latent distributions) that is less meaningful to

researchers than those expressed in the original metric (of non-normal distributions). On a positive note, some simulation studies suggest that the biases in the estimated factor loadings (differences between estimates in the transformed metric versus the population values in the original metric) when latent responses are non-normal tend to be trivial when *robust* WLS estimation (e.g., DWLS, ULS) is used, especially with larger sample sizes (e.g., 500 or greater; Flora & Curran, 2004; Rhemtulla et al., 2012).

Limitations and Implications for Future Research

Given that the high standard of the unique factor invariance must be satisfied to interpret differences in the means of the observed ordered-categorical indicators, we proposed and illustrated a sensitivity analysis based on the model-predicted probabilities to gauge the practical significance of the violations of longitudinal measurement invariance. Simulation studies are needed to examine whether this analysis is equally sensitive to violations of invariance of factor loadings, thresholds, and unique factor variances. Other approaches to sensitivity analysis also need to be developed. Drawing on studies of growth models with continuous indicators (see Kim & Willson, 2014a; 2014b; Leite, 2007; Wirth, 2008), approaches can be developed for ordered-categorical data that examine effects of inappropriate measurement invariance constraints on estimates of key growth parameters and their standard errors. Such sensitivity analyses assume that the form of the latent growth model has been correctly specified. Further research is needed to investigate the performance of such potential sensitivity analysis for ordered-categorical indicators, including the influence of sample size, number of indicators, and improper specification of the growth curve model.

In the test of longitudinal measurement invariance with more than two measurement waves, additional constraints potentially could be put on the stability of the corresponding lagged

covariances of the unique factors across time, which could be considered an additional (higher) level of longitudinal measurement invariance beyond strict invariance. For ordered-categorical indicators, the lagged unique factor covariances influence the lagged covariances of the *continuous latent responses* over time. They also influence the probability that an observed ordered-categorical indicator takes on specific values at two measurement occasions (see Appendix A). The effect of lagged covariances of the unique factors over time is an area that could be examined as part of the test of longitudinal measurement invariance in future research.

Appropriate treatment of missing data is needed to obtain accurate model fit statistics and nested model tests. Although multiple imputation is well understood in the context of ML estimation, less is known about its performance in conjunction with DWLS or ULS estimation that is necessary at smaller sample sizes. To our knowledge, no published article has addressed pooling the overall model fit statistics like CFI, RMSEA, and the χ^2 goodness-of-fit statistic across multiple imputed data sets, which is critical to the evaluation of measurement invariance. A procedure proposed by Meng and Rubin (1992) to combine likelihood ratio test statistics obtained using maximum likelihood estimation across multiply imputed data sets might be extended to combining nested model test statistics obtained using DWLS or ULS with ordered-categorical indicators, but further study is needed to examine its performance. Alternatively, Li, Raghunathan, and Rubin (1991) proposed a procedure in the context of ML estimation for pooling the multivariate-Wald test for a set of parameters across imputed data sets, which may be used to test the longitudinal invariance of parameters (indicated by non-significance of changes in parameters). However, further investigation is needed to evaluate how well these procedures work with ordered-categorical data using DWLS or ULS estimation.

Finally, no satisfactory estimation methods presently exist for data sets with fewer than 200 cases when assessing measurement invariance in longitudinal data with ordered-categorical measures. Bayesian estimation methods offer some promise here, although they will require information (e.g., prior research) to specify reasonable values for the prior distributions. Once again, evaluation of this promise will require further investigation.

Concluding Remarks

In this didactic article we have presented a full summary of the theoretical issues in testing longitudinal measurement invariance with ordered-categorical variables including model specification and testing. We have also addressed many of the common practical problems that arise in testing measurement invariance in longitudinal research and provided comparisons of three popular statistical programs for estimating these models. Finally, we have presented computer scripts in the supplemental materials for estimating the necessary models and programs and macros for computing proper estimates of the CFI and for gauging the practical effects of violations of measurement invariance. While challenges remain and relatively large sample sizes are required, we have provided applied researchers with the necessary foundation and the practical tools to conduct these analyses in their own research.

References

Asparouhov, T., & Muthén, B. O. (2006). *Robust chi square difference testing with mean and variance adjusted test statistics* (M*plus* Web Notes No. 10). Retrieved from http://www.statmodel.com/download/webnotes/webnote10.pdf

Asparouhov, T., & Muthén, B. O. (2010a). *Simple second order chi-square correction* (Technical Report). Retrieved from http://www.statmodel.com/download/WLSMV_new_chi21.pdf

Asparouhov, T., & Muthén, B. O. (2010b). *Weighted least squares estimation with missing data* (Technical Report). Retrieved from http://www.statmodel.com/download/GstrucMissingRevision.pdf

Asparouhov, T., & Muthén, B. O. (2010c). *Bayesian analysis of latent variable models using Mplus* (Technical Report). Retrieved from http://statmodel.com/download/BayesAdvantages6.pdf

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*(2), 186-203. DOI: 10.1207/s15328007sem1302_2

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. DOI: 10.1037/0033-2909.107.2.238

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606. DOI: 10.1037/0033-2909.88.3.588

Boker, S. M., Neale, M. C., Maes, H. H., Wilde, M. J., Spiegel, M., Brick, T. R., Estabrook, R., Bates, T. C., Mehta, P., von Oertzen, T., Gore, R. J., Hunter, M. D., Hackett, D. C., Karch

J., & Brandmaier, A. (2014). *OpenMx 2.0.0-3838 user guide*. Retrieved from http://openmx.psyc.virginia.edu/docs/OpenMx/2.0.0-3838/OpenMxUserGuide.pdf

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

Brown, M. B., & Bendetti, J. K. (1977). On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika*, *42*(3), 347–355. DOI: 10.1007/BF02293655

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

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464-504. DOI: 10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. DOI: 10.1207/S15328007SEM0902_5

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, *9*(3), 327-346. DOI: 10.1207/S15328007SEM0903_2

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*(2), 309-326. DOI: 10.1111/j.2044-8317.1994.tb01039.x

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. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*(3), 275-299. DOI: 10.1037/a0015825

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

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*(1), 80-100. DOI: 10.1207/S15328007SEM1001_4

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549-576. DOI: 10.1146/annurev.psych.58.110405.085530

Grimm, K. J., Ram, N., & Estabrook, R. (in press). *Growth modeling: An illustrative approach*. New York: Guilford Press.

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75*(3), 393-419. DOI: 10.1007/s11336-010-9165-5

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*(3), 347-387. DOI: 10.1207/S15327906347-387

Khoo, S. T., West, S. G., Wu, W., & Kwok, O. M. (2006). Longitudinal methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 301-317). Washington, D.C.: American Psychological Association.

Kim, E. S., & Willson, V. L. (2014a). Measurement invariance across groups in latent growth Modeling. *Structural Equation Modeling*, *21*(3), 408-424. DOI: 10.1080/10705511.2014.915374

Kim, E. S., & Willson, V. L. (2014b). Testing measurement invariance across groups in longitudinal data: Multigroup second-order latent growth model. *Structural Equation Modeling*, *21*(4), 566-576. DOI: 10.1080/10705511.2014.919821

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746-759. DOI: 10.1177/0013164496056005002

Knight, G.P., Gonzales, N.A., Saenz, D.S., Bonds, D.D., Germán, M., Deardorff, J., & Updegraff, K.A. (2010). The Mexican American cultural values scale for adolescents and adults. *The Journal of Early Adolescence*, *30*(3), 444-481. DOI: 10.1177/0272431609338178

Leite, W. L. (2007). A comparison of latent growth models for constructs measured by multiple items. *Structural Equation Modeling*, *14*(4), 581–610. DOI: 10.1080/10705510701575438

Li, K. H., Raghunathan, T. E. & Rubin, D. B. (1991). Large sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, *86*(416), 1065-1073. DOI: 10.1080/01621459.1991.10475152

McDonald, R. P., & Ho, M-H. R. (2002). Principles and practice in reporting structural equation models. *Psychological Methods*, *7*(1), 64-82. DOI: org/10-1037/1082/989X/7.1.64

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009-1020. DOI: 10.1198/016214504000002069

Meng, X. L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika, 79*(1), 103-111. DOI: 10.1093/biomet/79.1.103

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525-543. DOI: 10.1007/BF02294825

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203-240). Washington, D.C.: American Psychological Association.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Millsap, R. E., & Cham, H. (2012). Investigating factorial invariance in longitudinal data. In B. Laursen, T. D. Little, & N.A. Card (Eds.), *Handbook of developmental research methods* (pp. 109-127). New York: Guilford.

Millsap, R. E., & Tein, J. Y. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479-515. DOI: 10.1207/S15327906MBR3903_4

Mislevy, R.J. (in press). Missing responses in item response theory. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (2nd Edition, Volume 2). Boca Raton, FL: Chapman & Hall/CRC Press.

Muthén, B. O. (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. O. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.

Muthén, B. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in M*plus*. *Mplus web note No. 4*. Retrieved from https://www.statmodel.com/download/webnotes/CatMGLong.pdf.

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*(4), 489–503. DOI: 10.1007/BF02294325

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Retrieved from http://www.statmodel.com/bmuthen/articles/Article_075.pdf

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460. DOI: 10.1007/BF02296207

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354-373. DOI: 10.1037/a0029315

Roosa, M.W., Liu, F.F., Torres, M., Gonzales, N.A., Knight, G.P., & Saenz, D. (2008). Sampling and recruitment in studies of cultural influences on adjustment: A case study with Mexican

Americans. *Journal of Family Psychology*, *22*(2), 293-302. DOI: 10.1037/0893-3200.22.2.293

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. URL http://www.jstatsoft.org/v48/i02/.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. DOI: 10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: J. Wiley & Sons.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 167-180. DOI: 10.1080/10705511.2014.882658

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507-514. DOI: 10.1007/BF02296192

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243-248. DOI: 10.1007/s11336-009-9135-y

Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 201-223. DOI: 10.1111/j.2044-8317.2012.02049.x

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393-408. DOI: 10.1007/BF02294363

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1-10. DOI: 10.1007/BF02291170

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Application in substance use domain. In K. Bryant, M. T. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research*, (pp. 281-324). Washington, DC: American Psychological Association.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, *8*(1), 16-37. DOI: 10.1037/1082-989X.8.1.16

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10-18. DOI: 10.1111/j.1750-8606.2009.00110.x

Wirth, R. J. (2008). *The effects of measurement non-invariance on parameter estimation in latent growth models* (Doctoral dissertation). Retrieved from ProQuest dissertation and theses database. (UMI No. 3331053).

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58-79. DOI: 10.1037/1082-989X.12.1.58

Wu, W., West, S. G., & Taylor, A. B. (2009). Evaluating model fit for growth curve models: integration of fit indices from SEM and MLM frameworks. *Psychological Methods, 14*(3), 183-201. DOI: 10.1037/a0015858

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, *14*(3), 435-463. DOI: 10.1080/10705510701301677