



Evaluating fit indices in a multilevel latent growth curve model: A Monte Carlo study

Hsien-Yuan Hsu¹ · John J. H. Lin² · Susan Troncoso Skidmore³ · Minjung Kim⁴

Published online: 10 December 2018
© Psychonomic Society, Inc. 2018

Abstract

The multilevel latent growth curve model (MLGCM), which is subsumed by the multilevel structural equation modeling framework, has been advocated as a means of investigating individual and cluster trajectories. Still, how to evaluate the goodness of fit of MLGCMs has not been well addressed. The purpose of this study was to conduct a systematic Monte Carlo simulation to carefully investigate the effectiveness of (a) level-specific fit indices and (b) target-specific fit indices in an MLGCM, in terms of their independence from the sample size's influence and their sensitivity to misspecification in the MLGCM that occurs in either the between-covariance, between-mean, or within-covariance structure. The design factors included the number of clusters, the cluster size, and the model specification. We used Mplus 7.4 to generate simulated replications and estimate each of the models. We appropriately controlled the severity of misspecification when we generated the simulated replications. The simulation results suggested that applying $RMSEA_{T_S_COV}$, $TLI_{T_S_COV}$, and $SRMR_B$ maximizes the capacity to detect misspecifications in the between-covariance structure. Moreover, the use of $RMSEA_{PS_B}$, CFI_{PS_B} , and TLI_{PS_B} is recommended for evaluating the fit of the between-mean structure. Finally, we found that evaluation of the within-covariance structure turned out to be unexpectedly challenging, because none of the within-level-specific fit indices ($RMSEA_{PS_W}$, CFI_{PS_W} , TLI_{PS_W} , and $SRMR_W$) had a practically significant sensitivity.

Keywords Fit index · Model evaluation · Multilevel latent growth curve model · Multilevel structural equation modeling

A panel study is a powerful longitudinal design in which data are observed or gathered from exactly the same people, group, or organization across multiple time points (Neuman, 2009). Panel studies allow researchers to investigate a moving picture of observed units over time (i.e., a trajectory), rather than a single snapshot, as in cross-sectional studies. During the past few decades, two-stage cluster sampling (TCS) has been widely adopted for most large-scale panel studies (e.g., the Education Longitudinal Study of 2002, Ingels et al., 2013; or the Early Childhood Longitudinal Study, Kindergarten Class of 1998–

1999, Tourangeau, Nord, Lê, Sorongon, & Najarian, 2009). Briefly speaking, TCS is conducted by randomly selecting clusters (e.g., schools), and then randomly selecting individuals (e.g., students) within the selected clusters (Lohr, 2009). Incorporating TCS in panel studies not only makes the research design more cost-efficient (Scheaffer, Mendenhall, & Ott, 2005), but also generates three-level data (e.g., school/cluster, student/individual, and time point) that permit a comprehensive investigation of trajectories at both the individual (e.g., student) and cluster (e.g., school) levels.

The *multilevel latent growth curve model* (MLGCM), which is subsumed by the multilevel structural equation modeling (MSEM) framework, has been advocated as a means of investigating individual and cluster trajectories (for further discussion, see B. O. Muthén & Asparouhov, 2011). In an MLGCM, the time dimension is converted into a multivariate vector, which allows three-level data to be analyzed with a two-level model in which individual-related parameters are estimated in the *within model* and cluster-related parameters are evaluated in the *between model*. An example of using an MLGCM to investigate individual and cluster trajectories can be found in B. O. Muthén's (1997) study, in which he analyzed data drawn from the Longitudinal Study of American Youth (LSAY; Miller, Kimmel, Hoffer, &

✉ Hsien-Yuan Hsu
Hsien-Yuan.Hsu@uth.tmc.edu

¹ Children's Learning Institute, University of Texas Health Science Center, Houston, TX, USA

² Office of Institutional Research, National Central University, Taoyuan City, Taiwan

³ Department of Educational Leadership, Sam Houston State University, Huntsville, TX, USA

⁴ Department of Educational Studies, Ohio State University, Columbus, OH, USA

Nelson, 2000), a national panel study of mathematics and science education in US public schools.

Still, how to evaluate the goodness of fit of MLGCMs has not been well addressed. Model evaluation is required in order to examine the extent to which the hypothesized models, proposed on the basis of solid theories or empirical findings, are representative of the relationships among the variables, given the data (Kaplan, 2009; Kline, 2011). One common approach to model evaluation uses *fit indices* (e.g., the root mean square error of approximation [RMSEA], comparative fit index [CFI], Tucker–Lewis index [TLI], and standardized root mean square residual [SRMR]) to assess the model fit. However, because a traditional MLGCM comprises both between and within models, it has been suggested that the models at different levels should be evaluated separately by *level-specific fit indices* (Hox, 2010; Hsu, Kwok, Acosta, & Lin, 2015; Ryu, 2014; Ryu & West, 2009). Studies contributing to understanding the performance of level-specific fit indices in MSEM have been conducted in the context of multilevel confirmatory factor analysis (MCFA; e.g., Hsu, Lin, Kwok, Acosta, & Willson, 2016; Ryu & West, 2009), multilevel path models (Ryu, 2014), and multilevel nonlinear models (Schermelele-Engel, Kerwer, & Klein, 2014). Among these three approaches, MCFA is the most similar to MLGCMs. Some researchers have recommended using the aforementioned level-specific fit indices, on the basis of simulation studies conducted in the context of MCFA. Particularly, Ryu and West (2009) investigated whether level-specific CFI and RMSEA could detect a lack of fit at both the within and between levels in MCFA, and they found that these fit indices correctly indicated poor model fit in the models at different levels, regardless of the sample size. However, we argue that this recommendation for using level-specific fit indices is hard to generalize from the context of MCFA to MLGCM. The reason is that Ryu and West's simulation study only considered one misspecification, occurring in the covariance structure of MCFA—the covariance between factors was 0.3 in a two-factor population MCFA model, but it was misspecified as 1.0. The designed misspecification was meaningful in MCFA but also limited the generalizability of their recommendations to MLGCMs, because MLGCMs often estimate the covariance structure as well as the mean structure, in order to have a comprehensive understanding of trajectories (W. Wu & West, 2010). Consequently, the current recommendations for using level-specific fit indices cannot effectively guide applied researchers to evaluate their hypothesized MLGCMs. For this reason, in our study we attempted to address this gap in the literature by systematically investigating the sensitivity of level-specific fit indices to misspecifications occurring within the different structures of MLGCMs.

In addition to level-specific fit indices, our study also evaluated the performance of *target-specific fit indices*, originally computed for *single-level latent growth curve models* (SLGCMs) that have both a covariance structure and a mean structure. If any

misspecifications occur in SLGCMs, traditional fit indices cannot tell which structure in the model is misspecified. To obtain more informative model evaluation results, W. Wu and West (2010) suggested that researchers consider evaluating these two structures separately. In their study, W. Wu and West generated and evaluated fit indices targeting the covariance structure and the mean structure separately. In the present study, we extended their investigation from the context of SLGCMs to MLGCMs. In MLGCMs, both the between and within models contain a covariance structure and a mean structure. However, the parameters in the within-mean structure are fixed to zero (dummy zero means; B. O. Muthén, 1997), and therefore within-level-specific fit indices are sufficient to describe the misspecification occurring in the within model. Consequently, we evaluated two kinds of target-specific fit indices in our study: target-specific fit indices for (a) the between-covariance structure (T_S_COV fit indices) and (b) the between-mean structure (T_S_MEAN fit indices). Following W. Wu and West (2010) as well as Ryu and West (2009), we created T_S_COV fit indices by saturating the within model as well as the mean structure of the between model. On the other hand, we created T_S_MEAN fit indices by saturating the within model as well as the covariance structure of the between model.

It should be noted that the extent to which target-specific fit index findings from SLGCMs can be generalized to MLGCMs remains a question. Target-specific fit indices are different in nature in SLGCMs versus MLGCMs. In SLGCMs, there is only a single-level model, and saturation can only occur in either the mean structure or the covariance structure. However, as is shown in Appendix A, which outlines a practical way to derive target-specific fit indices, the computation of target-specific fit indices in MLGCM requires that the within model be saturated as well. To the best of our knowledge, the sensitivity of target-specific fit indices in MLGCMs has not been investigated, and this study was intended to close this gap in the literature.

In summary, the purpose of this study was to conduct a systematic Monte Carlo simulation in order to carefully investigate the effectiveness of (a) level-specific fit indices and (b) target-specific fit indices in MLGCMs across varying conditions. We evaluated the extent to which fit indices could be independent of sampling error due to small sample sizes when a hypothesized model was correctly specified (Gerbing & Anderson, 1992; Marsh, Hau, & Grayson, 2005). Moreover, we evaluated the extent to which different fit indices could reflect the discrepancy between correctly specified models and misspecified hypothesized models (i.e., the indices' sensitivity). We expected desirable fit indices to be less impacted by sampling error and to demonstrate reasonable sensitivity to misspecifications. This study contributes to the MSEM literature in two ways. First, it adds an understanding of the performance of level-specific and target-specific fit indices in MSEM. Second, it makes recommendations regarding model evaluation practices for MLGCMs.

Multilevel latent growth curve models

This section presents a two-level latent growth curve model capturing quadratic growth at both levels as an example. The featured clustered longitudinal data include repeated measures for each individual nested within the groups, thus forming a three-level structure. Consider a multilevel dataset with T waves of repeated measures for each of N individuals nested within G groups. For the i th individual within the g th group, \mathbf{y}_{ig} is a multivariate normally distributed random vector with T elements of repeated measures y_{tig} , $t = 1, 2, \dots, T$, which can be expressed as

$$\mathbf{y}_{ig} = [y_{1ig}, y_{2ig}, \dots, y_{Tig}]'_{T \times 1}, i = 1, 2, \dots, N; g = 1, 2, \dots, G.$$

The random vector \mathbf{y}_{ig} can be decomposed into its between-level (B) and within-level (W) components:

$$\mathbf{y}_{ig} = \mathbf{y}_{B..g} + \mathbf{y}_{W..ig} \\ = \boldsymbol{\mu}_B + \boldsymbol{\Lambda}_B \boldsymbol{\eta}_{B..g} + \boldsymbol{\varepsilon}_{B..g} + \boldsymbol{\mu}_W + \boldsymbol{\Lambda}_W \boldsymbol{\eta}_{W..ig} + \boldsymbol{\varepsilon}_{W..ig}. \quad (1)$$

Here, two random vectors representing the unique variances of repeated measures, $\boldsymbol{\varepsilon}_{B..g}$ and $\boldsymbol{\varepsilon}_{W..ig}$, are specified separately for the two levels and are assumed to be uncorrelated with each other; $\boldsymbol{\varepsilon}_{B..g}$ is multivariately normally distributed with mean zero and variance Θ_B , and $\boldsymbol{\varepsilon}_{W..ig}$ is multivariately normally distributed with mean zero and variance Θ_W . The random vectors of latent growth factors $\boldsymbol{\eta}_{B..g}$ and $\boldsymbol{\eta}_{W..ig}$ comprise the latent growth factors I (intercept factor), L (linear slope factor), and Q (quadratic slope factor), and the corresponding factor loading matrices $\boldsymbol{\Lambda}_B$ and $\boldsymbol{\Lambda}_W$ for T (e.g., five) waves of measurements are set as:

$$\boldsymbol{\eta}_{B..g} = \begin{bmatrix} I_B \\ L_B \\ Q_B \end{bmatrix}_{3 \times 1} \quad \text{with } \boldsymbol{\Lambda}_B = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix}_{5 \times 3}, \text{ and } \boldsymbol{\eta}_{W..ig} = \begin{bmatrix} I_W \\ L_W \\ Q_W \end{bmatrix}_{3 \times 1} \\ \text{with } \boldsymbol{\Lambda}_W = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix}_{5 \times 3}.$$

The variance–covariance matrix for \mathbf{y}_{ig} is presented in Eq. 2:

$$\text{Cov}(\mathbf{y}_{ig}) = \boldsymbol{\Lambda}_B \boldsymbol{\Psi}_B \boldsymbol{\Lambda}_B' + \boldsymbol{\Theta}_B + \boldsymbol{\Lambda}_W \boldsymbol{\Psi}_W \boldsymbol{\Lambda}_W' + \boldsymbol{\Theta}_W. \quad (2)$$

Model evaluation in multilevel structural equation modeling

Level-specific fit indices Previous studies have indicated that traditional global fit indices (e.g., RMSEA) can only reveal the model fit of the within models, and thus cannot be used to evaluate the between models (Hsu et al., 2015; Ryu & West, 2009). Hox (2010) drew attention to the need to develop level-

specific (l-s) fit indices to evaluate the within model and the between model separately. Ryu and West began the evaluation of l-s fit indices. Several recent published works have recommended that researchers apply l-s fit indices to evaluate the corresponding models at different levels in MSEMs (Hsu et al., 2016; Ryu, 2014; Ryu & West, 2009; Schermelleh-Engel et al., 2014).

According to Ryu and West (2009), the partially saturated model method (PS method) can be used straightforwardly to compute l-s fit indices. For example, using the PS method, between-level-specific (b-l-s) χ^2 test statistics ($\chi^2_{PS_B}$) can be obtained by specifying a hypothesized between model and saturating the within model (Hox, 2010). A saturated model can be seen as a just-identified model with zero degrees of freedom, and thus has a χ^2 test statistic equal to zero. As a result, b-l-s χ^2 test statistics only reflect the model fit of the hypothesized between model (Hox, 2010). After $\chi^2_{PS_B}$ is obtained, b-l-s fit indices (e.g., RMSEA_{PS_B}, CFI_{PS_B}, and TLI_{PS_B}) can be computed, because these fit indices are a function of the χ^2 test statistics. In the same way, within-level-specific (w-l-s) χ^2 test statistics ($\chi^2_{PS_W}$) can be also derived by using the PS method, reflecting the model fit of the hypothesized within model. After $\chi^2_{PS_W}$ is obtained, w-l-s fit indices (e.g., RMSEA_{PS_W}, CFI_{PS_W}, and TLI_{PS_W}) can be computed. In addition, alternative l-s fit indices, SRMR_B and SRMR_W, which are not computed on the basis of l-s χ^2 test statistics, are also available to evaluate models at different levels in some statistical packages (e.g., Mplus). The formulas for computing l-s fit indices are introduced in Appendix B. Generally, b-l-s and w-l-s fit indices are expected to detect any misspecifications occurring in the between model and the within model, respectively.

Target-specific fit indices We also evaluated target-specific (t-s) fit indices in this study. The idea of t-s fit indices originated in W. Wu and West's (2010) study, which investigated the performance of fit indices in SLGCMs. SLGCMs contain a covariance structure and a mean structure. As W. Wu and West pointed out, traditional fit indices, such as RMSEA, can reflect the overall fit of an SLGCM, but fail to detect the structure in which the misspecification occurs. Thus, the results from an SLGCM evaluation using global fit indices do not provide sufficient information to substantive researchers for further model modification. Accordingly, W. Wu and West asserted that there is a need for fit indices that target evaluating the fit of one specific structure (covariance or mean structure) of the model.

More specifically, t-s fit indices for the mean structure only (e.g., RMSEA_{T_S_Mean}, CFI_{T_S_Mean}, TLI_{T_S_Mean}, and SRMR_{T_S_Mean}) can be derived by saturating the covariance structure of the SLGCM, whereas t-s fit indices for the covariance structure only (e.g., RMSEA_{T_S_Cov}, CFI_{T_S_Cov}, TLI_{T_S_Cov}, and SRMR_{T_S_Cov}) can be derived by saturating the mean structure of the SLGCM. W. Wu and West (2010) found that RMSEA_{T_S_MEAN}, CFI_{T_S_MEAN}, TLI_{T_S_MEAN}, and SRMR_{T_S_MEAN} can be used to evaluate the fit of the mean structure, while RMSEA_{T_S_COV}, CFI_{T_S_COV}, TLI_{T_S_COV}, and SRMR_{T_S_COV} can be used to evaluate the fit of the covariance structure.

MEAN, and $SRMR_{T_S_MEAN}$ were more sensitive to misspecifications in the mean structure than are traditional fit indices. However, $RMSEA_{T_S_COV}$, $CFI_{T_S_COV}$, $TLI_{T_S_COV}$, and traditional fit indices performed similarly in terms of their sensitivity to misspecifications in the covariance structure. Leite and Stapleton (2011) concurred with W. Wu and West's findings, by confirming that $SRMR_{T_S_MEAN}$ has greater power for rejecting misspecifications in the mean structure than does traditional SRMR, whereas $RMSEA_{T_S_COV}$ could not improve the power of detection of traditional RMSEA. W. Wu and West (2010) explained that saturating the covariance structure could dramatically reduce the degrees of freedom, which in turn increases the power of the t-s-mean fit indices to detect misspecified mean structures. On the other hand, saturating the mean structure usually decreases the degrees of freedom by only a small amount, and thus cannot substantially change the power of t-s-cov fit indices.

Although t-s fit indices were recommended as a useful strategy for evaluating SLGCMs, as of yet there has been no empirical evidence in the literature to support the use of t-s fit indices for MLGCMs. To contribute to the understanding of how t-s fit indices perform, we examined the effectiveness of t-s fit indices in the context of MLGCMs. As previously mentioned, we evaluated two kinds of t-s fit indices in this study: the target-specific fit indices for (a) the between-covariance structure ($RMSEA_{T_S_COV}$, $CFI_{T_S_COV}$, $TLI_{T_S_COV}$, and $SRMR_{T_S_COV}$) and (b) the between-mean structure ($RMSEA_{T_S_MEAN}$, $CFI_{T_S_MEAN}$, $TLI_{T_S_MEAN}$, and $SRMR_{T_S_MEAN}$). We note that we did not evaluate t-s fit indices for the within model, because the means of growth factors are fixed at zero (dummy zero means; B. O. Muthén, 1997), and it appears self-evident that any misspecifications detected by the w-l-s fit indices could be attributed to the within-covariance structure.

Following W. Wu and West (2010) and Ryu and West (2009), we created $RMSEA_{T_S_COV}$, $CFI_{T_S_COV}$, $TLI_{T_S_COV}$, and $SRMR_{T_S_COV}$ by saturating the within model as well as the mean structure of the between model. More specifically, to saturate the mean structure, we freely estimated the intercepts for all repeated measures and fixed the means of the growth factors (i.e., IB, LB, and QB in Fig. 1) to zero. On the other hand, we created $RMSEA_{T_S_MEAN}$, $CFI_{T_S_MEAN}$, $TLI_{T_S_MEAN}$, and $SRMR_{T_S_MEAN}$ by saturating the within model as well as the covariance structure of the between model. Appendix A outlines a practical way to derive t-s fit indices.

More considerations when computing CFI- and TLI-related fit indices Both CFI and TLI are used to evaluate model fit by comparing the hypothesized model to the independence model (Bentler, 1990; Tucker & Lewis, 1973). Note that the independence model must be nested within the hypothesized model. We created CFI- and TLI-related fit indices based on Widaman and Thompson's (2003) approach, in which the independence model is an intercept-only growth model in

which only the mean of the intercept factor and the residual variances are freely estimated.

Method

We conducted a Monte Carlo study to assess the performance both of l-s fit indices ($RMSEA_{PS_B}$, CFI_{PS_B} , TLI_{PS_B} , $SRMR_B$, $RMSEA_{PS_W}$, CFI_{PS_W} , TLI_{PS_W} , $SRMR_W$) and of t-s fit indices ($RMSEA_{T_S_COV}$, $CFI_{T_S_COV}$, $TLI_{T_S_COV}$, $SRMR_{T_S_COV}$, $RMSEA_{T_S_MEAN}$, $CFI_{T_S_MEAN}$, $TLI_{T_S_MEAN}$, and $SRMR_{T_S_MEAN}$) in an MLGCM in terms of their independence from the sample size's influence and their sensitivity to misspecifications occurring in the between-covariance, between-mean, or within-covariance structures. The design factors we considered included the number of clusters, the cluster size, and the model specification. We used Mplus 7.4 (L. K. Muthén & Muthén, 1998–2017) to generate simulated replications and estimate each of the models.

Population model

We adopted an MLGCM, shown in Fig. 1, as the population model. In line with previous simulation studies (W. Wu & West, 2010; W. Wu, West, & Taylor, 2009), we used a quadratic trajectory population model to generate simulated data. The repeated measures, denoted as V1–V5, are assumed to be on a standardized scale (i.e., $M = 0$ and $SD = 1$). The quadratic growth pattern is modeled at both the within and the between levels. The factor loadings of the intercept factors (IW and IB) are fixed at 1.0, and those of the linear slope factors (LW and LB) are set to 0, 1, 2, 3, and 4 as part of the growth model parameterization. To model a quadratic growth pattern, we specified the quadratic slope factors (QW and QB) with factor loadings set to 0, 1, 4, 9, and 16.

To mimic more realistic conditions from an empirical dataset, we adopted parameter settings from LSAY (Miller et al., 2000), which used a two-stage stratified probability sampling approach—representative schools were randomly selected, and students within the selected school were then randomly sampled. The LSAY has been widely used to study the growth in mathematics and science performance (e.g., Ma & Ma, 2004; Ma & Wilkins, 2007). Following B. O. Muthén's (2004) study, we analyzed the Cohort 2 data, which contain 3,102 students nested within 52 schools. We used an MLGCM, as presented in Fig. 1, to analyze the students' grade 7 to grade 11 mathematics achievement scores obtained by item response theory equating. The intraclass correlation coefficients (ICCs) of five repeated measures ranged from .15 to .19. The ICC herein is a cluster statistic (e.g., school level) defined as the ratio between cluster-level variance and total variance (Cohen, Cohen, West, & Aiken, 2003; B. O. Muthén & Satorra, 1995). The identified magnitudes for the

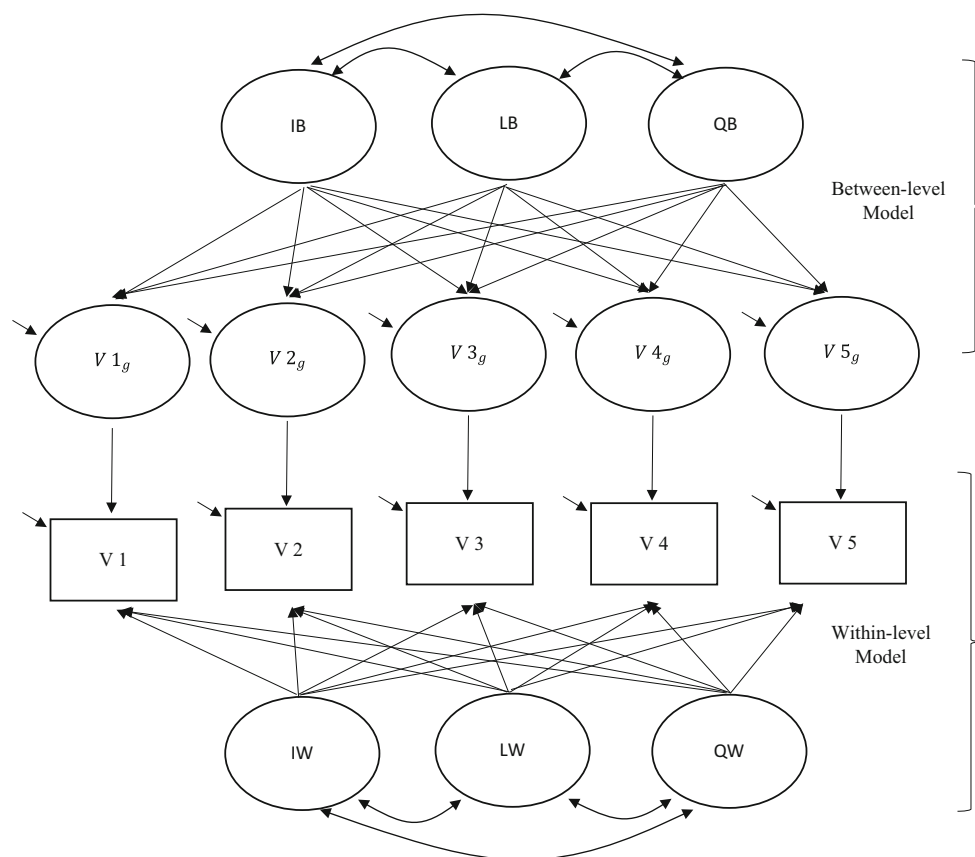


Fig. 1 A two-level MLGCM

ICCs are common in educational research and suggested that clustering should not be ignored (Hox, 2010).

The parameter settings for the mean structure (α_W) and covariance structure (Φ_W) of the population within model are presented in the following matrices:

$$\alpha_W = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \Phi_W = \begin{bmatrix} 71.453 & 6.762 (\tau_{01}) & 0 \\ 6.762 (\tau_{10}) & 14.755 & 0 \\ 0 & 0 & 0.703 (\tau_{22}) \end{bmatrix}.$$

In α_W , the means of *IW*, *LW*, and *QW*, referred to as *dummy zero means* by B. O. Muthén (1997), are fixed at zero. In Φ_W , the diagonal values are the variances of *IW* (71.453), *LW* (14.755), and *QW* (0.703), and the nondiagonal values are the covariances among the three factors. Following W. Wu and West's (2010) simulation design, the covariances between *IW* and *QW* and between *LW* and *QW* are constrained to be zero for simplicity. The error variances are 11.906, 15.249, 10.321, 12.592, and 1.931 for the repeated measures V1–V5, respectively, and are uncorrelated over time.

In this simulation, the between model has the same structure as the within-level model. The parameter settings for the

mean structure and covariance structure in the between model are presented in matrices α_B and Φ_B , respectively:

$$\alpha_B = \begin{bmatrix} 49.956 \\ 4.324 \\ -0.127 (\gamma_{200}) \end{bmatrix}, \quad \Phi_B = \begin{bmatrix} 16.200 & 2.819 (\beta_{01}) & 0 \\ 2.819 (\beta_{10}) & 0.609 & 0 \\ 0 & 0 & 0.018 (\beta_{22}) \end{bmatrix}.$$

The means of *IB*, *LB*, and *QB* are set to 49.956, 4.324, and -0.127 , respectively. In Φ_B , the diagonal values are the variances of *IB* (16.200), *LB* (0.609), and *QB* (0.018), and the nondiagonal values are the covariances among the three factors. The covariances between *IB* and *QB* and between *LB* and *QB* are constrained to be zero. The error variances of the repeated measures V1–V5 are independent and set to 1.800, 1.277, 0.059, 0.541, and 0.305, respectively.

Design factors

We took three related design factors into account: the number of clusters (NC), cluster size (CS), and model specification. NC is a critical factor when estimating MSEMs (Hox & Maas, 2001; Hox, Maas, & Brinkhuis, 2010). Therefore, we

considered different NCs in accordance with Hox and Maas's and J.-Y. Wu, Kwok, and Willson's (2015) simulation studies: 50, 100, and 200. The CS was manipulated into three levels, 5, 10, and 20, in line with the simulation design in Hox and Maas's (2001) study. This CS range is also consistent with the CSs found in two large-scale educational databases: the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (CS = 18; Tourangeau et al., 2009) and the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (CS = 15; Tourangeau et al., 2015).

The last design factor, the model specification, had two scenarios: correct specification and misspecification. *Correct specification* meant that a hypothesized model identical to the population model was specified to fit each of the simulated replications. An ideal fit index would successfully identify the correctly specified models. Conversely, if a hypothesized model with an intentionally imposed misspecification were fitted to the simulated replications, the ideal fit index would indicate that the hypothesized model was misspecified. The following section presents detailed information regarding the different types of intentional misspecifications.

Intentional misspecifications in the hypothesized models We created a total of five types of model misspecifications in this study. In line with previous simulation studies (W. Wu & West, 2010; W. Wu et al., 2009), we considered three intentional misspecifications in the hypothesized between model for this simulation study: (1) misspecification of the covariance between the intercept factor and the linear slope factor at the between level as 0 ($\beta_{10}/\beta_{01} = 0$; MIS_COV_B), (2) misspecification of the variance of the quadratic slope factor at the between level as 0 ($\beta_{22} = 0$; MIS_VAR_B), and (3) misspecification of the mean of the quadratic factor at the between level as 0 ($\gamma_{200} = 0$; MIS_MEAN_B). Additionally, we considered two intentional misspecifications in the hypothesized within model: (4) misspecification of the covariance between the intercept factor and the linear slope factor at the within level as 0 ($\tau_{10}/\tau_{01} = 0$; MIS_COV_W), and (5) misspecification of the variance of the quadratic slope factor at the within level as 0 ($\tau_{22} = 0$; MIS_VAR_W). Consistent with previous studies, we only considered underparameterized misspecifications in the study. We implemented an underparameterized misspecification by fixing at zero targeted parameter values whose population values were nonzero (Hu & Bentler, 1998). In each case, only one misspecification was imposed in the hypothesized model. Between-model-related fit indices were expected to detect intentional misspecifications occurring in the between model (MIS_COV_B, MIS_VAR_B, and MIS_MEAN_B), and within-model-related fit indices were expected to detect intentional misspecifications occurring in the within model (MIS_COV_W and MIS_VAR_W).

In this study, we did not consider misspecifications in the residual (co) variances at the between or within levels, for two

reasons. First, the residual variances at the between level are often trivial (Hox, 2010). Therefore, misspecification in the residual (co) variances at the between level is less likely to be of practical concern for researchers. Second, the structure of residual (co) variances at the within level (i.e., within-subject residuals) can be very complicated, and an independent simulation study on this issue is therefore warranted (e.g., Kwok, West, & Green, 2007).

Population parameters To ensure that the severities of the five intentional misspecifications were the same (Fan & Sivo, 2005), we adjusted the magnitudes of key parameters in the population models in order to achieve a severity of misspecification equal to a power of .80, given the number of clusters = 100 and the cluster size = 10 (for further discussion, see W. Wu & West, 2010, pp. 427–428), before generating the simulated replications. Appendix C presents the key parameter settings in the population models for each of the five misspecified conditions. For example, population model M1 in Appendix C was used to generate simulated data for the MIS_COV_B condition (i.e., the intentional misspecification $\beta_{10}/\beta_{01} = 0$). The value of the population parameter β_{10}/β_{01} was adjusted to 4.390 rather than 2.819, and hence the severity of the misspecification $\beta_{10}/\beta_{01} = 0$ reached a power of .80. As a result, the severity of the misspecifications would not confound the performance of the fit indices of interest.

Analysis

The two design factors related to the sample size (NC: 50, 100, and 200; CS: 5, 10, and 20) and six specification conditions (correct specification, MIS_COV_B, MIS_VAR_B, MIS_MEAN_B, MIS_COV_W, and MIS_VAR_W) were integrated into 54 conditions. For each condition, replications with convergence problems or improper solutions (e.g., negative unique variances) were excluded until 1,000 replications had been generated. The fit indices of interest produced by both the correctly specified models and the misspecified models were saved for further analyses.

We conducted two sets of analyses. The first set of analyses evaluated whether the fit indices of interest could be independent of sampling errors due to small sample sizes when a hypothesized model was correctly specified. Following Marsh, Hau, and Grayson (2005), we analyzed the values of the fit indices (i.e., outcome variables) under the condition that the hypothesized models were correctly specified (i.e., correct specification condition). We conducted a series of factorial analyses of variance (ANOVAs) on the fit indices values to examine the effects of the design factors NC and CS on the performance of the fit indices. Fit indices with lower effect sizes (indicated by eta-squared, discussed below) of the factors NC and CS were less influenced by sampling errors. Because RMSEA-, CFI-, and

TLI-related fit indices were a function of the χ^2 test statistic, we conducted similar ANOVAs on χ^2 test statistics to inform readers of the extent to which the χ^2 test statistic could be influenced by design factors.

The second set of analyses evaluated the extent to which the fit indices were able to reflect the discrepancy between correctly specified models and misspecified models (i.e., their sensitivity). Ideal fit indices should reflect the misfit arising from the imposed misspecification. Therefore, the sensitivity of fit indices can be captured as a discrepancy between the values derived from the misspecified model and the correctly specified model. A larger magnitude of discrepancy suggests higher sensitivity of a fit index. Analytically, we combined the fit indices values derived from a misspecified model and from the correctly specified model into one dataset and then analyzed the values of fit indices (outcome variables) with ANOVAs to determine sensitivity. Factors in the ANOVAs included sensitivity (SEN; i.e., replications fitted to misspecified models vs. correctly specified models), type of misspecification (MIS; e.g., MIS_COV_B vs. MIS_VAR_B), NC, CS, and all interaction terms. Similar ANOVAs were conducted on χ^2 test statistics for comparison purposes. On the basis of the results of these two analyses, we were able to make recommendations for practical and theoretical research. These recommendations appear in the Results and Conclusion sections.

As was mentioned in the Method section, we adjusted key parameters in the population models (see [Appendix C](#), population models M1–M5) to ensure that the findings derived from the five intentional misspecification conditions would be comparable (i.e., not confounded by the severity of different misspecifications). Population models M1 and M2 were designed to generate simulated replications for evaluating fit indices in terms of their sensitivity to the misspecified between-covariance structure. The evaluation of the b-l-s and t-s-cov fit indices was based on replications generated by population models M1 and M2. In contrast, population model M3 was designed to evaluate the sensitivity of the fit indices to the misspecified between-mean structure. Therefore, we evaluated the t-s-mean fit indices based on replications generated by population model M3. Finally, we evaluated the w-l-s fit indices based on replications generated by population models M4 and M5. Furthermore, for each factorial ANOVA, the total sum of squares (SOS) of each fit index provided the variability of the fit index values across all replications under specific simulation conditions. We computed eta-squared (η^2) by dividing the Type III SOS of a particular predictor or the interaction effect by the corrected total SOS, which provides the proportion of the variance accounted for by a particular design factor or interaction effect term. Following Cohen's (1988, 1992) suggestion, we adopted a moderate η^2 of .0588 in

order to identify influential design factors in the fit indices values (i.e., practically significant). Note that in cases in which a fit index had a corrected total SOS close to 0 (or a variance close to 0), the impact of design factors on the fit index were self-evidently trivial, even though the η^2 s were larger than .0588. In our analysis, when fit indices have extremely low variability, we have further clarified the interpretation of their design factors' η^2 s in the Results section.

Results

In this section, three tables have been created to present the simulation results. Table 1 shows the ANOVA results (η^2) with χ^2 test statistics or fit indices values as the dependent variables to evaluate the sensitivity to sampling errors, whereas Table 2 shows the ANOVA results to investigate the sensitivity to misspecifications. To understand the difference on the performance of the targeted fit indices and χ^2 test statistics, we also present the descriptive statistics of χ^2 test statistics including means, standard deviations, and Type I error rates/power in Table 3. Our intent in presenting Table 3 is to inform researchers under which conditions we encourage that χ^2 test statistics be used (given reasonable Type I error rates or power) along with fit indices.

Convergence rates

The convergence rates for each simulation condition were approximately 100%. The results suggest that the smallest sample size we considered in this study (NC = 50, CS = 5) was unlikely to encounter convergence problems if a MLGCM with five repeated measures was specified and analyzed.

First set of analyses: Sensitivity to sampling error

The first set of analyses we conducted aimed at evaluating whether fit indices of interest could be independent of sampling errors due to small sample sizes when a hypothesized model was correctly specified. The left side of Table 1 presents the ANOVA results (η^2) with χ^2 test statistics or fit index values as the dependent variables in order to evaluate the sensitivity of each approach to sampling errors. Note that the main effects of NC and CS were practically significant for some fit indices (described below), but none of interaction effects of NC and CS were practically significant. We provide a visual representation of the main effects of NC and CS with boxplots for each fit index in Figs. 2 and 3, respectively. The horizontal dashed lines noted on the figures denote the traditional cutoff criteria

Table 1 ANOVA results (η^2) with χ^2 test statistics and fit indices values as the dependent variable to indicate the sensitivity to sampling error (left side) and required NC and CS to accurately identify correct models (right side)

Dependent Variable	Number of Cluster (NC)	Cluster Size (CS)	NC \times CS	Required NC and CS to Accurately Identify Correct Models ^b	
χ^2 Test Statistics				NC	CS
$\chi^2_{PS_B}$.025	.037	.005	Cannot reasonably identify correct models given a sample of 4,000 (NC/CS = 200/20).	
$\chi^2_{PS_W}$.016	.025	.007	2,000 or above (NC/CS = 100/20, 200/10, or 200/20)	
$\chi^2_{T_S_COV}$.005	.013	.006	Cannot reasonably identify correct models given a sample of 4,000 (NC/CS = 200/20).	
$\chi^2_{T_S_MEAN}$.000	.007	.011	200	20
Between-level Specific Fit Indices					
RMSEA _{PS_B}	.124	.041	.008	200	> 20
CFI _{PS_B}	.050	.063	.031	50	5
TLI _{PS_B}	.045	.059	.029	100	10
SRMR _B ^a	.088 ^a	.189 ^a	.019	50	5
Within-level Specific Fit Indices					
RMSEA _{PS_W}	.086	.124	.029	50	5
CFI _{PS_W} ^a	.026	.033	.021	50	5
TLI _{PS_W} ^a	.025	.032	.021	50	5
SRMR _W ^a	.090 ^a	.152 ^a	.016	50	5
Target-specific Fit Indices for the Between-Covariance Structure					
RMSEA _{T_S_COV}	.158	.051	.019	200	> 20
CFI _{T_S_COV}	.069	.071	.050	50	5
TLI _{T_S_COV}	.062	.064	.048	50	5
SRMR _{T_S_COV} ^a	.086 ^a	.190 ^a	.019	50	5
Target-specific Fit Indices for the Between-Mean Structure					
RMSEA _{T_S_MEAN}	.133	.114	.021	200	> 20
CFI _{T_S_MEAN}	.047	.094	.033	50	5
TLI _{T_S_MEAN}	.061	.120	.038	100	10
SRMR _{T_S_MEAN} ^a	.031	.078 ^a	.051	50	5

RMSEA = root mean square error of approximation. CFI = comparative fit index. TLI = Tucker–Lewis index. SRMR = standardized root mean square residual. Subscripted PS = partially saturated model method. Subscripted TS = target-specific fit indices. Subscripted B = between-level model. Subscripted W = within-level model. Subscripted COV = fit index for evaluating between-covariance structure. Subscripted MEAN = fit index for evaluating between-mean structure. ^aFit index demonstrated extremely low variability and thus its magnitude of η^2 is less practically meaningful. We highlight (gray shaded cells) $\eta^2 \geq .0588$. ^bThe nominal α level (.05) and traditional cutoff criteria of the fit indices (RMSEA-related fit indices < .06; CFI- and TLI-related fit indices > .95; SRMR-related fit indices < .08; Hu & Bentler, 1999) were applied in order to determine the required NC and CS to correctly identify correct models for the χ^2 test statistics and fit indices, respectively.

of the fit indices (RMSEA-related fit indices < .06; CFI- and TLI-related fit indices > .95; SRMR-related fit indices < .08; Hu & Bentler, 1999). Those lines are expected to facilitate a better understanding regarding whether these fit indices were able to accurately identify correct models across NCs and CSs if the cutoff criteria were applied.

χ^2 test statistics Provided in Table 1 are the various χ^2 test statistics, including $\chi^2_{PS_B}$, $\chi^2_{PS_W}$, $\chi^2_{T_S_COV}$, and $\chi^2_{T_S_MEAN}$, which had η^2 s ranging from .000 to .037. In other words, these χ^2 test statistics were not

practically significantly impacted by NC, CS, or NC \times CS. Presented in Table 3 (top of table) are the means, standard deviations, and Type I errors of these four χ^2 test statistics. In general, $\chi^2_{PS_B}$ (means ranging from 6.897 to 21.886), $\chi^2_{PS_W}$ (means ranging from 1.464 to 17.477), $\chi^2_{T_S_COV}$ (means ranging from 5.082 to 13.893), and $\chi^2_{T_S_MEAN}$ (means ranging from 9.520 to 10.097) had means approaching the degrees of freedom when NC = 200 and CS = 20.

The Type I error rates of the four χ^2 test statistics tended to be closer to the nominal α level (.05) when the sample size

Table 2 ANOVA results (η^2) with χ^2 test statistics and fit indices values as the dependent variable to indicate the sensitivity to misspecifications

	Dependent Variable	Sensitivity (SEN)	Number of Cluster (NC)	Cluster Size (CS)	Type of Misspecification (MIS)	SEN \times NC	SEN \times CS	SEN \times MIS
Misspecified Between-Covariance Structure	$\chi^2_{PS_B}$.179	.004	.008	.012	.028	.047	.000
	$RMSEA_{PS_B}$.148	.079	.002	.032	.003	.046	.002
	CFI_{PS_B}	.026	.051	.051	.007	.000	.001	.001
	TLI_{PS_B}	.012	.045	.047	.012	.001	.003	.002
	$SRMR_B$.307	.026	.069	.012	.000	.001	.006
	$\chi^2_{TS_COV}$.189	.010	.012	.012	.022	.033	.002
	$RMSEA_{TS_COV}$.280	.050	.007	.031	.002	.052	.006
	CFI_{TS_COV}	.054	.042	.043	.005	.001	.000	.000
	TLI_{TS_COV}	.082	.029	.031	.016	.001	.001	.006
	$SRMR_{TS_COV}$.305	.026	.069	.013	.000	.001	.008
	$\chi^2_{PS_B}$.645	.079	.015	NA ^b	.090	.021	NA ^b
	$RMSEA_{PS_B}$.805	.007	.004	NA ^b	.001	.013	NA ^b
	CFI_{PS_B}	.605	.005	.052	NA ^b	.000	.025	NA ^b
	TLI_{PS_B}	.601	.004	.049	NA ^b	.000	.025	NA ^b
Misspecified Between-Mean Structure	$SRMR_B^a$.164 ^a	.082 ^a	.112 ^a	NA ^b	.001	.000	NA ^b
	$\chi^2_{TS_MEAN}$.393	.030	.001	NA ^b	.045	.004	NA ^b
	$RMSEA_{TS_MEAN}$.827	.008	.000	NA ^b	.000	.002	NA ^b
	CFI_{TS_MEAN}	.335	.014	.074	NA ^b	.005	.049	NA ^b
	TLI_{TS_MEAN}	.403	.011	.072	NA ^b	.006	.058	NA ^b
	$SRMR_{TS_MEAN}^a$.223 ^a	.044	.136 ^a	NA ^b	.003	.009	NA ^b
	$\chi^2_{PS_W}$.195	.002	.006	.016	.032	.036	.007
	$RMSEA_{PS_W}$.056	.083	.117	.021	.004	.007	.010
Misspecified Within-Covariance Structure	$CFI_{PS_W}^a$.020	.064	.094	.011	.000	.000	.003
	$TLI_{PS_W}^a$.009	.055	.083	.016	.001	.002	.006
	$SRMR_W^a$.006	.073	.169	.012	.000	.002	.000

RMSEA = root mean square error of approximation. CFI = comparative fit index. TLI = Tucker-Lewis index. SRMR = standardized root mean square residual. Subscripted PS = partially saturated model method. Subscripted TS = target-specific fit indices. Subscripted B = between-level model. Subscripted W = within-level model. Subscripted COV = fit index for evaluating between-covariance structure. Subscripted MEAN = fit index for evaluating between-mean structure. ^a Fit index demonstrated extremely low variability and thus its magnitude of η^2 is less practically meaningful. ^b Since there was only one misspecification in the between-mean structure (i.e., $\gamma_{200} = 0$; MIS_MEAN_B), the η^2 s for interaction effect $SEN \times$ Type of Misspecification cannot be computed. Note the η^2 s for remaining second-, third-, or higher-order interaction effects were approximately zero and thus are not reported in the table. We highlight (gray-shaded cells) $\eta^2 \geq .0588$

increased; however, we found that Type I error rates were inflated across various NC and CS conditions. More specifically, the Type I error rates of three of the between-level-related χ^2 test statistics ($\chi^2_{PS_B}$, $\chi^2_{TS_COV}$, $\chi^2_{TS_MEAN}$) were not satisfactory: $\chi^2_{PS_B}$ had Type I error rates ranging from .094 to .421 across all sample conditions; $\chi^2_{TS_COV}$'s error rates ranged from .116 to .354; and $\chi^2_{TS_MEAN}$'s error rates ranged from .065 to .285, where .065 occurred when the sample size was 4,000 (NC/CS = 200/20). On the other hand, $\chi^2_{PS_W}$ had Type I error rates ranging from .051 to .426, where increasing the sample size to 2,000 or above resulted in a reduced Type I error rate. For example, given a sample size of 2,000 (NC/CS = 100/20 or 200/10), the Type I error rates of $\chi^2_{PS_W}$ were between .062 and .077; given a sample size of 4,000 (NC/CS = 200/20), the Type I error rate was .051.

Between-level-specific fit indices Practically, $RMSEA_{PS_B}$ was significantly influenced by NC only ($\eta^2 = .124$), whereas CFI_{PS_B} and TLI_{PS_B} were by CS only (η^2 s = .063 and .059, respectively). As is suggested by Fig. 2a, $RMSEA_{PS_B}$ approached values indicative of good model fit (i.e., 0) when NC increased. Thus, NC = 200 is recommended if it is being used to identify a correct between model. On the other hand, in Fig. 3b, CFI_{PS_B} and TLI_{PS_B} approached values indicative of good model fit (i.e., 1) when CS increased. CFI_{PS_B} was able to identify correct between models given CS = 5. However, as compared with CFI_{PS_B} , the values of TLI_{PS_B} were more spread out and thus needed CS = 10 in order to correctly identify correct between models.

Table 3 Descriptive statistics of χ^2 test statistics by NC and CS for the correctly specified and misspecified conditions

Specification Condition	Fit Index	NC		50	50	50	100	100	100	200	200	200
		CS		5	10	20	5	10	20	5	10	20
Correctly Specified Condition	$\chi^2_{PS,B}$ (df = 6)	Mean		21.886	16.012	10.306	18.283	9.770	7.585	11.442	7.424	6.897
		SD		34.048	24.609	12.937	28.438	11.725	4.860	13.778	4.793	4.158
		Type I Error (%)		.421	.348	.222	.407	.211	.133	.276	.122	.094
	$\chi^2_{PS,W}$ (df = 6)	Mean		17.477	13.599	8.026	15.613	8.509	6.404	10.464	6.544	5.969
		SD		19.291	14.301	5.844	15.777	6.820	3.655	10.020	3.853	3.491
		Type I Error (%)		.426	.350	.139	.423	.161	.062	.259	.077	.051
	$\chi^2_{T,S,COV}$ (df = 4)	Mean		11.094	12.625	10.226	13.893	9.203	6.407	12.444	6.401	5.082
		SD		24.207	24.095	20.832	24.850	15.445	7.581	23.971	7.969	4.080
		Type I Error (%)		.219	.310	.255	.354	.256	.176	.334	.173	.116
	$\chi^2_{T,S,MEAN}$ (df = 2)	Mean		5.134	5.515	6.159	6.276	9.566	3.421	10.097	5.999	2.238
		SD		14.141	11.958	17.794	13.017	29.456	6.287	26.553	16.150	2.355
		Type I Error (%)		.165	.199	.204	.206	.274	.124	.285	.207	.065
MIS_COVB	$\chi^2_{PS,B}$ (df = 7)	Mean		32.763	33.714	31.604	32.022	29.231	38.975	28.069	37.702	60.599
		SD		41.489	39.285	25.191	35.836	19.863	18.269	20.299	15.957	18.631
		Power (%)		.574	.726	.865	.732	.843	.987	.810	.968	1.000
	$\chi^2_{T,S,COV}$ (df = 5)	Mean		17.070	29.530	37.914	31.595	37.042	44.327	37.278	43.207	64.843
		SD		25.907	37.250	36.274	40.421	36.872	27.637	39.361	25.411	24.483
		Power (%)		.364	.689	.910	.699	.892	.994	.867	.986	1.000
MIS_VARB	$\chi^2_{PS,B}$ (df = 9)	Mean		23.292	21.453	26.380	19.360	23.871	37.573	20.832	33.537	62.649
		SD		19.740	12.899	13.179	11.338	11.566	13.320	10.385	12.756	17.554
		Power (%)		.556	.580	.768	.496	.709	.969	.587	.930	1.000
	$\chi^2_{T,S,COV}$ (df = 7)	Mean		27.806	22.193	25.232	20.949	22.715	35.802	19.837	32.009	60.681
		SD		33.496	19.862	14.966	21.105	12.893	13.584	11.687	13.108	17.513
		Power (%)		.589	.627	.792	.577	.762	.977	.652	.949	1.000
MIS_MEANB	$\chi^2_{PS,B}$ (df = 7)	Mean		35.717	34.114	25.338	40.200	31.863	34.147	43.114	45.531	57.009
		SD		38.885	33.918	14.509	35.563	16.895	11.684	27.220	15.250	13.682
		Power (%)		.699	.824	.826	.904	.938	.981	.967	.997	1.000
	$\chi^2_{T,S,MEAN}$ (df = 3)	Mean		13.524	25.294	28.110	33.023	44.284	33.816	68.375	55.414	53.360
		SD		17.698	27.703	24.797	44.310	44.376	19.979	71.310	37.417	15.041
		Power (%)		.602	.824	.931	.896	.971	.991	.988	1.000	1.000
MIS_COVW	$\chi^2_{PS,W}$ (df = 7)	Mean		22.165	22.884	21.130	24.507	21.631	27.067	23.984	26.853	44.022
		SD		20.046	17.952	11.460	18.297	12.447	10.709	13.903	11.619	12.753
		Power (%)		.554	.651	.724	.700	.710	.912	.766	.877	.998
MIS_VARW	$\chi^2_{PS,W}$ (df = 9)	Mean		22.214	17.355	16.481	17.950	16.414	20.642	17.203	20.547	31.292
		SD		16.579	10.761	7.437	10.621	7.948	7.809	8.889	8.371	10.302
		Power (%)		.512	.398	.415	.453	.405	.656	.431	.646	.939

MIS_COVB = misspecification of the covariance between the intercept factor and linear slope factor at the between level as 0 ($\beta_{10}/\beta_{01} = 0$). MIS_VARB = misspecification of the variance of the quadratic slope factor at the between level as 0 ($\beta_{22} = 0$). MIS_MEANB = misspecification of the mean of the quadratic factor at the between level as 0 ($\gamma_{200} = 0$). MIS_COVW = misspecification of the covariance between the intercept factor and the linear slope factor at the within level as 0 ($\tau_{10}/\tau_{01} = 0$). MIS_VARW = misspecification of the variance of the quadratic slope factor at the within level as 0 ($\tau_{22} = 0$). All Type I error rates of the χ^2 test statistics were inflated ($\geq .05$). Power of the χ^2 test statistics $\geq .80$ is highlighted in gray

At first glance, SRMR_B appears to be strongly impacted by both NC ($\eta^2 = .088$) and CS ($\eta^2 = .189$). However, the boxplots in Figs. 2a and 3a clearly depict that the variability of SRMR_B is extremely low. Indeed, the variance of SRMR_B, computed on the basis of all simulated replicates, was close to 0. Therefore, the noted magnitude of the η^2 's of NC and CS did not necessarily mean that NC and CS had any practical impact on SRMR_B.

Within-level-specific fit indices RMSEA_{PS_W} was impacted by both NC ($\eta^2 = .086$) and CS ($\eta^2 = .124$). On the basis of the magnitude of η^2 , RMSEA_{PS_W} was more influenced by CS than by NC. We found that both CFI_{PS_W} and TLI_{PS_W} had low variabilities and, practically, were not significantly influenced by NC or CS. On the other hand, although SRMR_W had practically significant values for the η^2 's of NC and CS, SRMR_W had very little variability, so NC and CS did not have practical impacts on SRMR_W. For all w-l-s fit

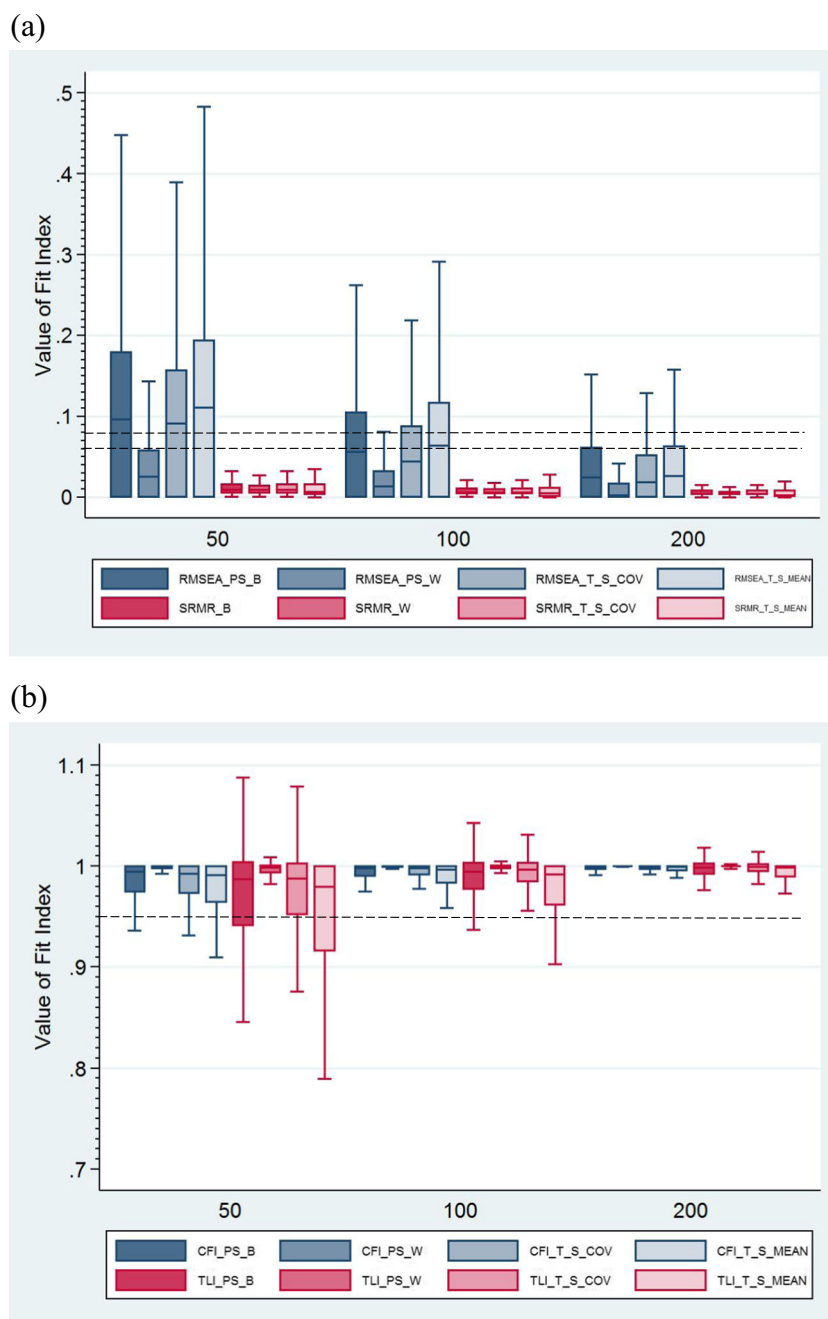


Fig. 2 (a) Box plot of RMSEA- and SRMR-related fit indices values derived from correctly specified models by number of clusters (50, 100, and 200). (b) Box plot of CFI- and TLI-related fit indices values derived from correctly specified models by number of clusters. Horizontal dashed

lines indicate the traditional cutoff criteria of fit indices (RMSEA-related fit indices $< .06$; SRMR-related fit indices $< .08$; CFI- and TLI-related fit indices $> .95$, Hu & Bentler, 1999)

indices, $NC = 50$ (as indicated by Figs. 2a and 2b) and $CS = 5$ (as indicated by Figs. 3a and 3b) were sufficient to correctly identify correct within models.

Target-specific fit indices for the between-covariance structure RMSEA_{T_s_COV} practically was significantly

influenced by NC ($\eta^2 = .158$) but not by CS ($\eta^2 = .051$). As is suggested in Fig. 2a, RMSEA_{T_s_COV} required $NC = 200$ in order to accurately identify correct between-covariance structures. On the other hand, both CFI_{T_s_COV} and TLI_{T_s_COV} were practically significantly affected by both NC (η^2 s = .069 and .062, respectively) and CS

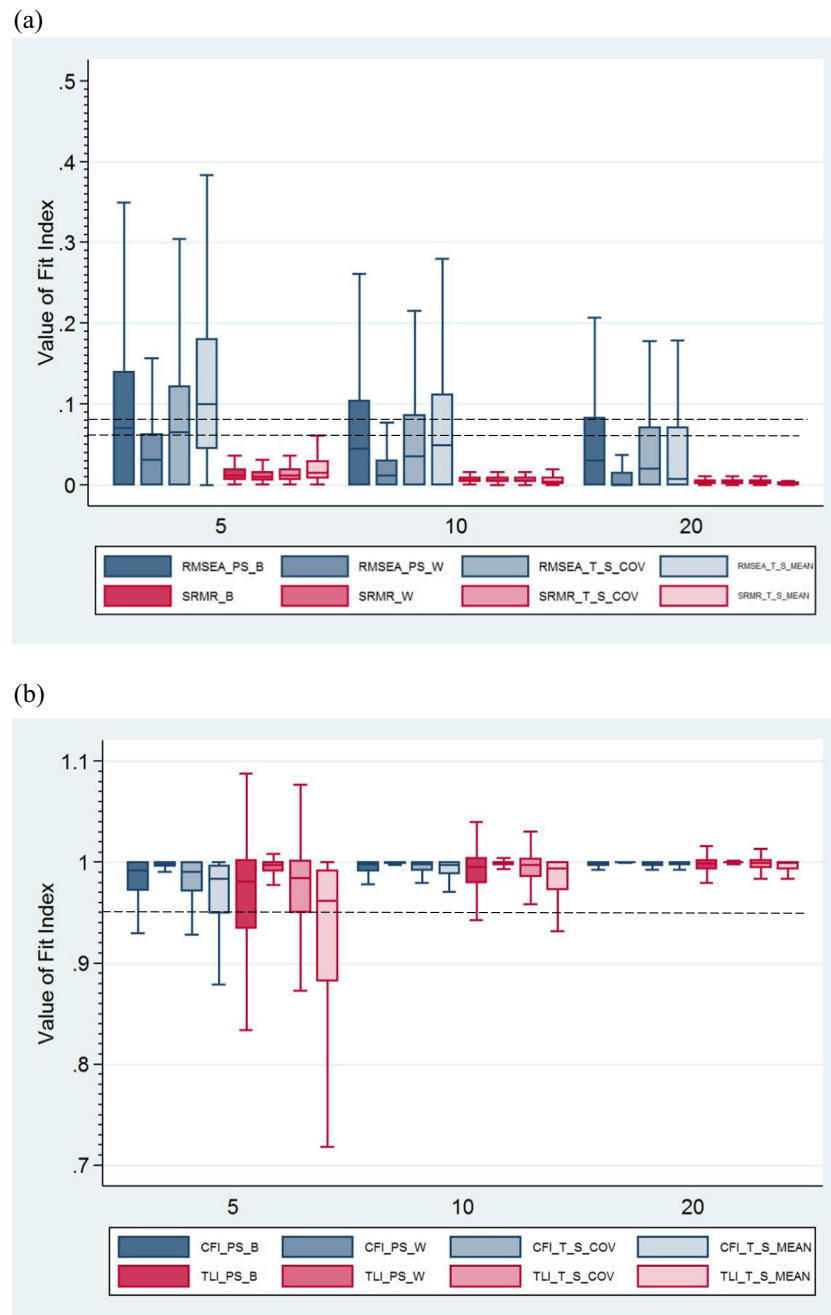


Fig. 3 (a) Box plot of RMSEA- and SRMR-related fit indices values derived from correctly specified models by cluster size (5, 10, and 20). (b) Box plot of CFI- and TLI-related fit indices values derived from correctly specified models by cluster size. Horizontal dashed lines

indicate the traditional cutoff criteria of fit indices (RMSEA-related fit indices < .06; SRMR-related fit indices < .08; CFI- and TLI-related fit indices > .95, Hu & Bentler, 1999)

(η^2 s = .071 and .064, respectively). Nevertheless, we found that NC = 50 (indicated by Fig. 2b) and CS = 5 (indicated by Fig. 3b) were sufficient to accurately identify correct between-covariance structures. As for $\text{SRMR}_{T_S_COV}$,

we found that its variance computed on the basis of all simulated replicates was close to 0. Therefore, $\text{SRMR}_{T_S_COV}$'s larger noted values of η^2 for NC and CS did not mean it was practically affected by NC and CS.

Target-specific fit indices for the between-mean structure RMSEA_{T_S_MEAN} practically was significantly influenced by both NC ($\eta^2 = .133$) and CS ($\eta^2 = .114$), and so was TLI_{T_S_MEAN} (η^2 : NC = .061, CS = .120). RMSEA_{T_S_MEAN} had a higher demand on both NC (200, as indicated by Fig. 2a) and CS (> 20, as indicated by Fig. 3a) in order to accurately identify correct between-mean structures, whereas TLI_{T_S_MEAN} had a moderate demand (NC = 100, as indicated by Fig. 2b, and CS = 10, as indicated by Fig. 3b). CFI_{T_S_MEAN} was practically significantly affected by CS ($\eta^2 = .094$), but nonetheless we found that CS = 5 was sufficient. On the other hand, SRMR_{T_S_MEAN} had a variance close to 0, and thus NC and CS had no practical effect on it.

Required NC and CS in order to accurately identify correct models/structures We have summarized the required values of NC and CS for the χ^2 test statistics and each fit index on the right side of Table 1. As we mentioned earlier, the Type I error rates of the four χ^2 test statistics were inflated across all sample conditions (see Table 3). The results provided that $\chi^2_{PS_B}$ and $\chi^2_{T_S_COV}$ could not reasonably identify correct models even when the sample size was as large as 4,000. In contrast, $\chi^2_{PS_W}$ and $\chi^2_{T_S_MEAN}$ had Type I error rates close to .05 when sample sizes were 2,000 or above (NC/CS = 100/20, 200/10, or 200/20) and 4,000 (NC/CS = 200/20), respectively. In general, all fit indices had low demand on NC (50) and CS (5), except for the following fit indices. Specifically, three of the RMSEA-related fit indices—namely RMSEA_{PS_B}, RMSEA_{T_S_COV}, and RMSEA_{T_S_MEAN}—had high demand on NC (= 200) and CS (> 20) in comparison to the other fit indices. Moreover, two of the TLI-related fit indices, TLI_{PS_B} and TLI_{T_S_MEAN}, had medium demand of NC (100) and CS (10). Our results suggested that the fit indices (except for RMSEA-related fit indices) were more effective in identifying correct models/structures than were the χ^2 test statistics.

Second set of analyses: Sensitivity to misspecification

In this section, we evaluate the performance of the fit indices under various scenarios in which the hypothesized models were misspecified. We present η^2 s for the main effects of sensitivity (SEN), NC, CS, and type of misspecification (MIS), as well as for some of their second-order interaction effects, in Table 2. The remaining second-, third-, or higher-order interaction effects are not presented in this table because the results showed that none of η^2 s of the interaction effects were practically significant. The means of the fit indices

across different sample size conditions can be requested by contacting the first author.

Misspecified between-covariance structure The top ten rows of Table 2 contain the η^2 s of the design factors for both the b-l-s and t-s-cov χ^2 test statistics and fit indices. Note that the factor MIS included two types of misspecifications, MIS_COV_B and MIS_VAR_B, and a practically significant η^2 of MIS indicated that the fit index was more sensitive to one type of misspecification (e.g., MIS_COV_B) than to the other (e.g., MIS_VAR_B).

The results suggested that both $\chi^2_{PS_B}$ and $\chi^2_{T_S_COV}$ had practically significant η^2 s for sensitivity (.179 and .189, respectively) and were not impacted by the other design factors. The means, standard deviations, and powers of $\chi^2_{PS_B}$ and $\chi^2_{T_S_COV}$ for the MIS_COV_B and MIS_VAR_B conditions are reported in Table 3. Although the values of $\chi^2_{PS_B}$ and $\chi^2_{T_S_COV}$ were not practically sensitive to our design factors, Table 3 suggests that these two χ^2 test statistics were able to detect a misspecified between-covariance structure when the sample sizes were large. Specifically, in the MIS_COV_B condition, both $\chi^2_{PS_B}$ and $\chi^2_{T_S_COV}$ had adequate power ($\geq .800$) given a sample size of 1,000 (NC/CS = 50/20, 100/10, or 200/5), whereas in the MIS_VAR_B condition, a sample size over 1,000 was required in order to reach an adequate power level. The results also suggested that $\chi^2_{T_S_COV}$ outperformed $\chi^2_{PS_B}$ in most sample size scenarios. In summary, a sample size over 1,000 was suggested for $\chi^2_{PS_B}$ and $\chi^2_{T_S_COV}$ to appropriately detect misspecified between-covariance structures. Moreover, $\chi^2_{T_S_COV}$ was favored over $\chi^2_{PS_B}$ due to its relatively higher power in most sample size conditions.

On the other hand, the t-s-cov fit indices (saturating both the within model and the between-mean structure) had relatively larger sensitivity η^2 s than did the b-l-s fit indices (saturating the within model only), except for SRMR. For example, the η^2 of sensitivity for RMSEA_{T_S_COV} was .280, which was larger than that of RMSEA_{PS_B} (.148). Similarly, TLI_{T_S_COV} had a larger η^2 of sensitivity (.082) than did TLI_{PS_B} (.012). Although CFI_{T_S_COV} had a larger η^2 of sensitivity (.054) than did CFI_{PS_B} (.026), both η^2 s were not practically significant.

Alternatively, SRMR_B and SRMR_{T_S_COV} had similarly high η^2 s of sensitivity (.307 and .305, respectively), suggesting that both fit indices performed equally well in terms of their sensitivity to the misspecified between-covariance structure. Nevertheless, we noticed that both SRMR_B and SRMR_{T_S_COV} were also practically significantly affected by CS (η^2 s = .069). In other words, when the between-covariance structure was misspecified, the values of SRMR_B and SRMR_{T_S_COV}

COV reflected not only the severity of the misspecification, but also the size of CS. Our further data analysis showed that the means of SRMR_B and SRMR_{T_S_COV} approached the values indicative of poor model fit (i.e., 1) when CS decreased.

In summary, SRMR_B and SRMR_{T_S_COV} acted comparably—both had the highest sensitivity to misspecified between-covariance structures, relative to all other fit indices, but they were also influenced by CS. Therefore, researchers are encouraged to use either SRMR_B or SRMR_{T_S_COV}, but they need to be aware that a small CS (e.g., 5) can contribute to the values of SRMR_B and SRMR_{T_S_COV} (leading to values indicative of poor model fit). Computing t-s-cov fit indices for RMSEA and TLI (i.e., RMSEA_{T_S_COV} and TLI_{T_S_COV}) was a favorable strategy for between-covariance structure evaluation. Strategically, researchers should rely more on RMSEA_{T_S_COV} than on TLI_{T_S_COV}, because RMSEA_{T_S_COV} had a larger η^2 of sensitivity. The results did not support CFI_{PS_B} or CFI_{T_S_COV} having a practically significant sensitivity. In addition, on the basis of the performance of $\chi^2_{PS_B}$ and $\chi^2_{T_S_COV}$ discussed earlier, we encourage researchers to use $\chi^2_{PS_B}$ or $\chi^2_{T_S_COV}$ (especially the latter, because of its higher power) in combination with fit indices when the sample size is over 1,000.

Misspecified between-mean structure The middle ten rows in Table 2 contain η^2 s of the design factors for both the b-l-s and t-s-cov χ^2 test statistics and fit indices. The factor MIS for the between-mean structure by design included only one type of misspecification (MIS_MEAN_B), and thus this factor had no variation, and its η^2 cannot be computed. Our original simulation design included three levels of NC (50, 100, and 200) and of CS (5, 10, and 20). Our preliminary analysis resulted in 12.77% of the replications for CS = 5 producing TLI_{T_S_MEAN} < 0 (the descriptive statistics of TLI_{T_S_MEAN} for these 12.77% of the replications were $M = -4.52$, $SD = 18.75$, $\min = -369.56$, and $\max = 0.00$). Moreover, we also found that a high percentage (7.11%) of the replications for CS = 10 produced TLI_{T_S_MEAN} < 0 ($M = -4.25$, $SD = 19.70$, $\min = -230.47$, and $\max = 0.00$), whereas most replications (57.36%) were with NC = 50. To more comprehensively evaluate the impacts of NC and CS on the b-l-s and t-s-mean fit indices, we expanded on our original work by considering larger sizes for NC and CS: NC = 100, 200, and 300; and CS = 10, 20, and 30. Changing the levels of the NC and CS design factors did not compromise our findings because we analyzed the values of the fit indices produced by the M3-model-simulated replications separately.

The results indicated that both $\chi^2_{PS_B}$ and $\chi^2_{T_S_MEAN}$ had practically significant η^2 s of sensitivity (.645 and .393,

respectively). However, $\chi^2_{PS_B}$'s sensitivity was impacted by NC ($\eta^2 = .079$) and also moderated by NC ($\eta^2 = .090$). The results suggested that increasing NC from 200 to 300 caused a larger growth in $\chi^2_{PS_B}$'s means than did increasing NC from 100 to 200. On the other hand, $\chi^2_{T_S_MEAN}$ was not practically significantly impacted by either NC or CS.

We report the means, standard deviations, and power of $\chi^2_{PS_B}$ and $\chi^2_{T_S_MEAN}$ in the MIS_MEAN_B condition in Table 3. For simplicity, we do not report that information for NC = 300 or CS = 30 in the table, given that the patterns of the means, standard deviations, and powers of $\chi^2_{PS_B}$ and $\chi^2_{T_S_MEAN}$ exhibited across NC = 100 to 200 and CS = 10 to 20 were quite clear. The results suggested that $\chi^2_{PS_B}$ and $\chi^2_{T_S_MEAN}$ had adequate power to detect misspecified between-mean structures when the sample size was 500 (NC/CS = 50/10 or 100/5) or above. In addition, we found that the two χ^2 test statistics performed similarly in most sample size conditions. In sum, considering that $\chi^2_{PS_B}$ is less computationally complicated, researchers might use $\chi^2_{PS_B}$ when the sample size is approximately 500 or above.

Both RMSEA_{PS_B} and RMSEA_{T_S_MEAN} performed comparably: Both had outstanding sensitivity (η^2 s = .805 and .827, respectively) and were not significantly impacted practically by the design factors. On the basis of the magnitudes of sensitivity η^2 values, we found that using RMSEA_{T_S_MEAN} instead of RMSEA_{PS_B} did not gain much sensitivity. On the other hand, CFI_{PS_B} was superior in comparison to CFI_{T_S_MEAN}—CFI_{PS_B} had a larger η^2 of sensitivity (.605) than did CFI_{T_S_MEAN} (.335). In addition, unlike CFI_{T_S_MEAN}, which was practically significantly impacted by CS ($\eta^2 = .074$), CFI_{PS_B} was not practically significantly impacted by any design factor. A similar pattern was found when we compared the performance of TLI_{PS_B} against that of TLI_{T_S_MEAN}.

As for SRMR, the results suggested that SRMR_{T_S_MEAN} (η^2 of sensitivity = .223) outperformed SRMR_B ($\eta^2 = .164$) regarding their sensitivity to the misspecified between-mean structure. However, we also recognize that SRMR_B and SRMR_{T_S_MEAN} had both means and variances close to 0 across different combinations of NC and CS, given the misspecified between-mean structure. In other words, both SRMR_B and SRMR_{T_S_MEAN} were not sensitive to the misspecified between-mean structure, and therefore their η^2 s of sensitivity had no practical meaning and thus can be ignored. On the basis of this finding, we suggest that both SRMR_B and SRMR_{T_S_MEAN} not be used for between-mean structure evaluation.

To sum up, $RMSEA_{T_S_MEAN}$ slightly outperformed $RMSEA_{PS_B}$ in terms of its sensitivity to the misspecification. However, considering that $RMSEA_{PS_B}$ (saturating the within model only) can be computed in a less complicated way than $RMSEA_{T_S_MEAN}$ (saturating both the within model and the between-covariance structure), $RMSEA_{PS_B}$ was favored. For CFI and TLI, their b-l-s forms (CFI_{PS_B} and TLI_{PS_B}) were preferred, too, because they had a larger sensitivity to misspecification and were not practically significantly impacted by other design factors. Although the η^2 s of sensitivity for both $SRMR_B$ and $SRMR_{T_S_MEAN}$ were practically significant (.164 and .223, respectively), the results showed they had both means and variances close to 0 across different combinations of NC and CS, given the misspecified between-mean structure. As a result, these practically significant η^2 s of sensitivity did not have a practical meaning, and neither $SRMR_B$ nor $SRMR_{T_S_MEAN}$ was recommended. Comparing the η^2 s of sensitivity for $RMSEA_{PS_B}$, CFI_{PS_B} , and TLI_{PS_B} that were useful for between-mean structure evaluation, we found that $RMSEA_{PS_B}$ was more sensitive to the misspecified between-mean structure and therefore should be used preferentially. Both CFI_{PS_B} and TLI_{PS_B} , on the other hand, performed alike and thus can be used interchangeably. In addition, according to our findings on $\chi^2_{PS_B}$ and $\chi^2_{T_S_MEAN}$, we suggest that researchers use the aforementioned fit indices along with $\chi^2_{PS_B}$ (because it can be easily computed and was as effective as $\chi^2_{T_S_MEAN}$) when the sample size is 500 or above.

Misspecified within-covariance structure The last five rows in Table 2 contain the η^2 s of the design factors for both the w-l-s χ^2 test statistics and fit indices. Note that the factor MIS included two types of misspecifications, MIS_COV_W and MIS_VAR_W . The results suggested that $\chi^2_{PS_W}$ had a practically significant η^2 s of sensitivity (.195) and was not impacted by the other design factors. The means, standard deviations, and power of $\chi^2_{PS_W}$ for the MIS_COV_W and MIS_VAR_W conditions are reported in Table 3. In the MIS_COV_W condition, $\chi^2_{PS_W}$ had adequate power given a sample size of 2,000 or higher (NC/CS = 100/20, 200/10, or 200/20), whereas in the MIS_VAR_W condition, a sample size larger than 2,000, and closer to 4,000, was needed.

The results showed that none of the w-l-s fit indices had a practically significant η^2 of sensitivity, and all were impacted by CS or by both NC and CS. Nevertheless, a deep investigation of the data showed that CFI_{PS_W} , TLI_{PS_W} , and $SRMR_W$

had variances approaching 0 across different combinations of NC and CS, given the misspecified within-covariance structure. That is, the impacts of NC and CS on CFI_{PS_W} , TLI_{PS_W} , and $SRMR_W$ were not practically significant. Our findings raised a concern that using w-l-s fit indices might yield an invalid conclusion on the fitness of the within-covariance structure. On the basis of our findings regarding $\chi^2_{PS_W}$, researchers might use $\chi^2_{PS_W}$ to evaluate the within-covariance structure. However, in general, a large sample size (larger than 2,000 and closer to 4,000) was needed for $\chi^2_{PS_W}$ to reach an adequate power level, which would not necessarily be practical.

Discussion

Sample size

In this study, we evaluated the performance of l-s and t-s fit indices in terms of their independence from sample size influence and sensitivity to misspecifications in a MLGCM. We expected ideal fit indices to be less influenced by sampling errors arising from a small sample size and to be more sensitive to misspecifications. Accordingly, we investigated the extent to which the fit indices of interest could be influenced by sampling errors, based on simulated data derived from correctly specified models. Table 1 presents the influences (in terms of η^2) of NC and CS (left side). The results showed that most of the fit indices were practically significantly influenced by NC or CS. Specifically, the fit indices indicated poor model fit as the sample size (a function of NC or CS) decreased, even though the hypothesized models were correctly specified (i.e., the hypothesized model was the same as the population model). This finding is in line with previous research (Marsh, Balla, & McDonald, 1988; Marsh et al., 2005; McDonald & Marsh, 1990; W. Wu & West, 2010; W. Wu et al., 2009) that has explored the issue of sample size dependency among fit indices. As was pointed out by Marsh et al. (2005), the discrepancy between the covariance matrix reproduced by a correctly specified model and a sample covariance matrix can vary systematically with the sample size. The reason is that a sample covariance matrix derived from a small sample size no longer approaches the population covariance matrix due to sampling error, and this in turn increases the discrepancy between the two matrices.

We would note that not all small sizes of NC or CS raised practical concerns when fit indices were applied. Specifically, traditional cutoff criteria of the fit indices (RMSEA-related fit indices < .06; CFI- and TLI-related fit indices > .95; SRMR-

related fit indices $< .08$; Hu & Bentler, 1999) were evaluated to determine the NC and CS required in order to accurately identify correct models (see Figs. 2a, 2b, 3a, and 3b). We summarize the required NC and CS for each fit index on the right side of Table 1. We found that the CFI- and SRMR-related fit indices were not practically affected by sampling errors resulting from a small NC or CS, because they were able to identify correct models even if a small NC (50) and CS (5) was given. The performance of SRMR-related fit indices was noteworthy. The ANOVAs on values of the SRMR-related fit indices had a corrected total SOS close to 0 (or a variance close to 0), so the impact of NC and CS on SRMR-related fit indices was self-evidently trivial. In contrast, RMSEA-related fit indices for the between-model evaluation, including $RMSEA_{PS_B}$, $RMSEA_{T_S_COV}$, and $RMSEA_{T_S_MEAN}$, were very likely to be affected by sampling errors in practice. Large NC (200) and CS (> 20) are recommended when these fit indices are used. Two TLI-related fit indices for the between-model evaluation, namely TLI_{PS_B} and $TLI_{T_S_MEAN}$, needed a moderate NC (100) and CS (10). Substantive researchers need to be aware of these characteristics of the fit indices and to strive for a sufficient sample size to obtain a more accurate model evaluation. Future studies can further investigate the necessary sample sizes for different population models—for example, a model with more time-point measures or with different types of trajectories (e.g., a piecewise-linear trajectory).

Furthermore, NC and CS, two parameters related to sample size in MLGCMs, might also impact the performance of fit indices in different ways. In fit indices for between-model evaluation, NC determines the sample size of the between model. As we discussed earlier, the NC decides the magnitude of the sampling error carried into the model—where smaller NC increases the amount of the sampling errors in model estimation. Consequently, some between-level fit indices can fail to identify correctly specified between models unless the NC is large (e.g., NC = 200 for $RMSEA_{PS_B}$, $RMSEA_{T_S_COV}$, and $RMSEA_{T_S_MEAN}$). On the other hand, CS can influence the quality of the scores of indicators (e.g., $V1_g$ – $V5_g$ in Fig. 1) in the between model of MLGCMs (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011). Note that in our simulation (and in practice), the scores of between-level indicators were estimated by the scores of within-level indicators. As was pointed out by Lüdtke et al., the quality of the estimated between-level indicator scores can be influenced by CS—where smaller CS increases the amount of sampling error in the between-level indicators scores. As a result, some between-level fit indices can fail to identify correctly specified between models unless the CS is large (e.g., CS > 20 for $RMSEA_{PS_B}$, $RMSEA_{T_S_COV}$, and $RMSEA_{T_S_MEAN}$).

In contrast to the case of fit indices for between-model evaluation, both NC and CS jointly determine the sample size of the within model, and therefore influence the performance of fit indices in similar ways.

Finally, in the exploration of the performance of four χ^2 test statistics ($\chi^2_{PS_B}$, $\chi^2_{PS_W}$, $\chi^2_{T_S_COV}$, and $\chi^2_{T_S_MEAN}$), we found that their Type I error rates were inflated unless the sample size was extremely large. These findings were consistent with Schermelleh-Engel et al.'s (2014) study, which investigated the effectiveness of l-s χ^2 test statistics ($\chi^2_{PS_B}$ and $\chi^2_{PS_W}$) under 6,000, 15,000, and 30,000 sample size conditions (NC/CS = 200/30, 500/30, and 1,000/30), which were much larger than our greatest sample size condition (NC/CS = 200/20). Schermelleh-Engel et al. found that $\chi^2_{PS_B}$ and $\chi^2_{PS_W}$ had Type I error rates lower than .05 when the sample sizes were 30,000 and 15,000, respectively. As a result, it was not surprising to see inflated Type I error rates for the four χ^2 test statistics in our study. Given our findings, researchers need to be aware that using these χ^2 test statistics will very likely lead to overrejection of correctly specified MLGCMs. Therefore, using fit indices jointly to evaluate model fit is highly recommended.

Fit indices for evaluating between-covariance structures

Our results in Table 2 show that RMSEA, CFI, and TLI in the form of t-s-cov ($RMSEA_{T_S_COV}$, $CFI_{T_S_COV}$, and $TLI_{T_S_COV}$) expressed a higher sensitivity to the misspecified between-covariance structure than did those same statistics in the form of b-l-s ($RMSEA_{PS_B}$, CFI_{PS_B} , and TLI_{PS_B}). This finding supports computing RMSEA, CFI, and TLI by saturating the within model as well as the between-mean structure as a favorable strategy for the evaluation of the between-covariance structure. The results also suggest that researchers should prioritize the utilization of $RMSEA_{T_S_COV}$, $CFI_{T_S_COV}$, and $TLI_{T_S_COV}$. Specifically, researchers can rely more on $RMSEA_{T_S_COV}$ than on $TLI_{T_S_COV}$, because $RMSEA_{T_S_COV}$ had a larger η^2 of sensitivity. $CFI_{T_S_COV}$, on the other hand, had no practically significant sensitivity and was therefore not recommended. The aforementioned finding that fit indices in the t-s-cov form were superior to those in the b-l-s form was expected, because the results presented in Table 3 suggested that $\chi^2_{T_S_COV}$ was favored over $\chi^2_{PS_B}$ due to its relatively higher power in most sample size conditions. Since the t-s-cov fit indices are a function of $\chi^2_{T_S_COV}$, it was not surprising to find that fit indices in the t-s-cov form outperformed those in

the b-l-s form in terms of their sensitivity to misspecified between-covariance structures.

Alternatively, we found that $SRMR_B$ and $SRMR_{T_S_COV}$ acted comparably. That is, computing SRMR in the form of t-s-cov or of b-l-s did not make any difference. On the basis of this finding, we recommend saturating only the within model to obtain $SRMR_B$ as a simple strategy to using SRMR. We did find that both $SRMR_B$ and $SRMR_{T_S_COV}$ had the largest sensitivity to misspecification, in comparison to the other fit indices. However, the results also suggested that $SRMR_B$ and $SRMR_{T_S_COV}$ were also influenced by CS—that is, a small CS (e.g., 5) resulted in $SRMR_B$ and $SRMR_{T_S_COV}$ values indicative of poor model fit. Overall, we recommend that SRMR be used along with other t-s-cov fit indices, especially when the CS is small.

Our findings are not consistent with W. Wu and West's (2010) study, in which they concluded that saturating the mean structure did not influence the sensitivity of fit indices except for SRMR (p. 446) in the context of a single-level latent growth curve model. We consider our findings to be reasonable for the following reasons. First, as W. Wu et al. (2009) mentioned, RMSEA is based only on the chi-squared statistic (χ^2) for the hypothesized model, and CFI and TLI are also defined by χ^2 . Therefore, these three fit indices can reflect the fit of a model to the mean structure. In other words, saturating the between-mean structure can influence the sensitivity of RMSEA, CFI, and TLI. Alternatively, SRMR is a weighted function of the model residuals, and it is not necessary to take into account the residuals of the means (deviations of the sample means from the model-implied means; W. Wu et al., p. 193). That is, SRMR disregards the information from the between-mean structure, so that saturating the between-mean structure therefore has no influence on SRMR.

Fit indices for evaluating between-mean structures

When evaluating misspecified between-mean structures, we found that (a) RMSEA, CFI, and TLI in the form of t-s-mean ($RMSEA_{T_S_MEAN}$, $CFI_{T_S_MEAN}$, and $TLI_{T_S_MEAN}$) did not necessarily display a higher sensitivity to misspecifications than those same statistics in the form of b-l-s ($RMSEA_{PS_B}$, CFI_{PS_B} , and TLI_{PS_B}), and (b) $SRMR_B$ and $SRMR_{T_S_MEAN}$ had both means and variances close to 0 across different combinations of NC and CS, given the misspecified between-mean structure. As a result, their practically significant η^2 s of sensitivity did not have practical meaning, such that both $SRMR_B$ and $SRMR_{T_S_MEAN}$ are not recommended. This initial finding suggested that researchers would not make substantial gains by using t-s-mean fit indices ($RMSEA_{T_S_MEAN}$, $CFI_{T_S_MEAN}$, and

$TLI_{T_S_MEAN}$). As is shown in Table 2, $RMSEA_{PS_B}$ demonstrated high sensitivity to misspecified between-mean structures (.805), and its sensitivity was very close to $RMSEA_{T_S_MEAN}$'s (.827). On the other hand, both CFI_{PS_B} and TLI_{PS_B} outperformed $CFI_{T_S_MEAN}$ and $TLI_{T_S_MEAN}$, respectively, in terms of their sensitivity to misspecifications. As a result, we recommend that researchers use $RMSEA_{PS_B}$, CFI_{PS_B} , and TLI_{PS_B} to evaluate between-mean structures, and $RMSEA_{PS_B}$ should be used with a higher priority, because it was more sensitive to the misspecified between-mean structure. These findings are not consistent with W. Wu and West's (2010) study, which concluded that saturating the covariance structure could increase the sensitivity of fit indices in the context of single-level latent growth curve models. We consider our findings to be reasonable, because we found that $\chi^2_{PS_B}$ and $\chi^2_{T_S_MEAN}$ performed similarly in most sample size conditions, in terms of their power to detect the misspecified between-mean structures (see Table 3). Because RMSEA, CFI, and TLI are computed from χ^2 , the fit indices in the t-s-mean form were expected to perform comparably to those in the b-l-s form. We encourage future studies to validate our findings in a multilevel context.

On the other hand, SRMR is not necessary for reflecting the model fit for means. Leite and Stapleton (2011) confirmed the superior sensitivity of RMSEA and the limited sensitivity of SRMR for identifying misspecified mean structures in latent growth models. In line with Leite and Stapleton's findings, we found that both $SRMR_B$ and $SRMR_{T_S_MEAN}$ had means close to zero as well as trivial variability across all sample size conditions.

We note that the η^2 s shown in Table 2 were derived from larger sizes for NC (100, 200, and 300) and CS (10, 20, and 30). Our original simulation design (NC = 50, 100, and 200; CS = 5, 10, and 20) resulted in a high percentage of replications producing $TLI_{T_S_MEAN} < 0$, especially when (a) CS = 5, regardless of NC, or (b) CS = 10 and NC = 50. We extracted those replications and found that they produced extremely large chi-squared values when we intended to compute t-s-mean fit indices (i.e., saturating the within model as well as the between-covariance structure) for the model with a misspecified between-mean structure. In fact, saturating the between-covariance structure in order to obtain t-s-mean fit indices raises the number of estimated parameters (e.g., increasing four parameters in our study), and in this case, the small sample size resulted in a larger than expected obtained chi-squared value (Bentler & Dudgeon, 1996; Jackson, 2003). On the basis of our simulation results, NC = 100 and CS = 10 were required for computing t-s-mean fit indices, but we encourage future studies to investigate this issue further.

Last but not least, on the basis of the η^2 s presented in Table 2, there was a tendency for the fit indices of interest to have a higher sensitivity to the misspecified between-mean structure than to the misspecified between-covariance structure. Nevertheless, as we mentioned, the η^2 s in this section were derived from an alternative NC and CS design (NC = 100, 200, and 300; CS = 10, 20 and 30). Therefore, these η^2 s were not necessarily comparable to the η^2 s derived from the original NC and CS design (NC = 50, 100, and 200; CS = 5, 10 and 20). Additional efforts will be needed to verify this tendency.

Fit indices for evaluating within-covariance structures

None of the w-l-s fit indices demonstrated promise in detecting a misspecified within-covariance structure. We did not expect to observe such low or near-zero sensitivity values for w-l-s fit indices, because previous research simulation studies (e.g., Ryu & West, 2009) had shown that w-l-s fit indices can successfully detect misspecified within models in the context of MCFA models. We therefore wondered whether traditional fit indices (e.g., RMSEA, CFI, TLI, and SRMR) could be more effective in identifying misspecified within-covariance structures. After comparing the performance of the w-l-s fit indices with that of traditional fit indices, we found that these two types of fit indices acted almost identically, showing little to no sensitivity to the misspecifications. We validated our finding by comparing it with W. Wu and West's (2010) study, which had evaluated fit indices in the context of a single-level latent growth curve model. More specifically, according to the information presented in Fig. 3 of W. Wu and West (2010, p. 437), they had an RMSEA close to .04, an SRMR close to .06, and a CFI/TLI > .99 across sample sizes from 125 to 1,000, given a moderate severity of misspecification (defined as power = .80, which we also adopted) in the covariance structures. Their findings are consistent with ours, except that they found slightly higher SRMR values. Moreover, the findings on w-l-s fit indices are also confirmed by our findings on χ^2_{PS-W} (see Table 3). Specifically, we found that χ^2_{PS-W} had little power to detect misspecified within-covariance structures unless the sample size was larger (greater than 2,000, and closer to 4,000). For those w-l-s fit indices that were a function of χ^2_{PS-W} (RMSEA_{PS-W}, CFI_{PS-W}, and TLI_{PS-W}), it was not surprising to see that they were not sensitive to the misspecification. In summary, our findings, as well as those of W. Wu and West, suggest that the evaluation of within-covariance structures can be challenging. Substantive researchers might be overly optimistic about the fit of the within model in MLGCMs, and future researchers are encouraged to validate our findings and look for an optimal strategy for within-model evaluation.

Limitations and future research direction

Because it is not possible to consider all plausible scenarios in a single simulation study, generalizations beyond the set of conditions investigated in our study should be made with caution. First, we adopted an MLGCM, as is shown in Fig. 1, for data generation. Therefore, our findings can only be generalized to studies that use similar MLGCMs. Further studies are encouraged to investigate whether our findings can also be replicated using different models (e.g., piecewise-linear trajectory models). Second, in our study, we did not consider misspecifications in the residual (co) variances for the between and within models. Practically, the residual variances at the between level are often low (Hox, 2010). Freely estimating between-level residual variances and constraining the covariances to zero seems to be a reasonable approach. Therefore, we did not consider misspecifications in residual (co) variances at the between level. On the other hand, the structure of residual (co) variances at the within level (i.e., within-subject residuals) can be more complicated (Kwok et al., 2007). Misspecifications in within-subject residuals are possible and deserve further systematic investigation in another simulation study. Future studies could evaluate fit indices in scenarios in which within-subject residuals are misspecified. Third, in our study, we considered a limited number of design factors. Additional scenarios created by adopting different design factors, such as unbalanced designs (unequal group conditions), the number of time-point measures, and ICCs of repeated measures, will be needed in future studies.

Last but not least, in our simulation design, the variance of the quadratic slope factor in the between-level (or within-level) population model was nonzero but was constrained to 0 as a type of misspecification (i.e., conditions MIS_VAR_B or MIS_COV_W). In each misspecification condition, only one parameter was misspecified. However, in practice, when the variance of the quadratic slope factor is misspecified (i.e., constrained to 0), two other parameters would be also automatically constrained to 0: (a) the covariance between the quadratic slope factor and the linear slope factor, and (b) the covariance between the quadratic slope factor and the intercept factor. Consequently, the findings on sensitivity of fit indices to a misspecified quadratic slope factor may be confounded with two additional potential misspecified parameters. To address this issue, W. Wu and West's (2010) specification (the two aforementioned covariance parameters also being set to 0 in the population model) can be applied to control the confounding factors. Nevertheless, W. Wu and West's specification of the population model might decrease the generalizability of our findings to empirical research. Future studies will be needed to validate our findings using a population with the aforementioned covariance parameters not equal to 0.

Conclusion

Previous simulation studies have investigated the performance of level-specific fit indices in the context of MCFA (Hsu et al., 2016; Ryu, 2014; Ryu & West, 2009). Our study has extended this research line by systematically examining the effectiveness of level-specific fit indices (RMSEA_{PS_W}, CFI_{PS_W}, TLI_{PS_W}, SRMR_W, RMSEA_{PS_B}, CFI_{PS_B}, TLI_{PS_B}, and SRMR_B) and target-specific fit indices (RMSEA_{T_S_COV}, CFI_{T_S_COV}, TLI_{T_S_COV}, SRMR_{T_S_COV}, RMSEA_{T_S_Mean}, CFI_{T_S_Mean}, TLI_{T_S_Mean}, and SRMR_{T_S_Mean}) in terms of their independence from the sample size's influence and their sensitivity to misspecification in MLGCMs. We appropriately controlled the severity of misspecification when

we generated the simulated replications. On the basis of our simulation results, we recommend applying RMSEA_{T_S_COV} and TLI_{T_S_COV} along with SRMR_B in order to maximize the capacity to detect misspecifications in the between-covariance structure. On the other hand, we recommend using RMSEA_{PS_B}, CFI_{PS_B}, and TLI_{PS_B} to detect misspecifications in the between-mean structure. Evaluation of the within-covariance structure turned out to be unexpectedly challenging, as none of w-l-s fit indices (RMSEA_{PS_W}, CFI_{PS_W}, TLI_{PS_W}, and SRMR_W) had a practically significant sensitivity. Future researchers are encouraged to validate our findings and to look for an optimal strategy for within-model evaluation.

Appendix A

Appendix A A practical way to derive fit indices of interest

Fit Index	Saturate the Between Model	Saturate the Within Model	Saturate the Covariance Structure of the Between Model	Saturate the Mean Structure of the Between Model
<i>Between-level-specific fit indices</i> (e.g., RMSEA _{PS_B})		✓		
<i>Within-level-specific fit indices</i> (e.g., RMSEA _{PS_W})	✓			
<i>Target-specific fit indices for the between-covariance structure</i> (e.g., RMSEA _{T_S_COV})		✓		✓
<i>Target-specific fit indices for the between-mean structure</i> (e.g., RMSEA _{T_S_MEAN})		✓	✓	

Appendix B. Equations of the partially saturated level-specific and target-specific fit indices^{B1}

Chi-squared statistic Three χ^2 statistics can be computed for between-level model evaluation: (a) level-specific (l-s) χ^2 ($\chi^2_{PS_B}$), (b) target-specific (t-s) χ^2 for the between-mean structure ($\chi^2_{T_S_MEAN}$), and (c) χ^2 for the between-covariance structure ($\chi^2_{T_S_COV}$). According to Ryu and West (2009), the following equation can be used to obtain $\chi^2_{PS_B}$:

$$\chi^2_{PS_B} = F_{ML}[\Sigma_B(\hat{\theta}), \Sigma_W(\hat{\theta}_S)] - F_{ML}[\Sigma_B(\hat{\theta}_S), \Sigma_W(\hat{\theta}_S)], \quad (B1)$$

where $F_{ML}[\Sigma_B(\hat{\theta}), \Sigma_W(\hat{\theta}_S)]$ is the value of fitting function for the two-level model with a hypothesized model between levels and a saturated model within levels (partially

saturated model); $F_{ML}[\Sigma_B(\hat{\theta}_S), \Sigma_W(\hat{\theta}_S)]$ is the value of fitting function when both the within-level and between-level models are saturated (fully saturated model). The degrees of freedom of $\chi^2_{PS_B}$ (denoted by df_{PS_B}) are equal to the difference between the numbers of parameters in the fully saturated model and the partially saturated model.

A different $F_{ML}[\Sigma_B(\hat{\theta}), \Sigma_W(\hat{\theta}_S)]$ in Eq. B1 is required in order to obtain $\chi^2_{T_S_MEAN}$ and $\chi^2_{T_S_COV}$. More specifically, $F_{ML}[\Sigma_B(\hat{\theta}), \Sigma_W(\hat{\theta}_S)]$ for $\chi^2_{T_S_MEAN}$ is the value of the fitting function for a partially saturated model in which the between-covariance structure is saturated. Using our population model shown in Fig. 1 as an example, a saturated between-covariance structure can be specified by freely estimating the (co) variances of $V1_g, \dots, V5_g$ at the between level and constraining the (co) variances of the intercept, linear slope, and quadratic slope factors to be zero (W. Wu & West, 2010). The degrees of freedom of

$\chi^2_{PS_B_Mean}$ (denoted by $df_{T_S_MEAN}$) are equal to the difference between the numbers of parameters in the fully saturated model and in the partially saturated model with the saturated between-covariance structure.

Alternatively, $F_{ML}[\Sigma_B(\hat{\theta}), \Sigma_W(\hat{\theta}_S)]$ for $\chi^2_{T_S_COV}$ is the value of fitting function for a partially saturated model in which the mean structure between levels is saturated. A saturated between-mean structure can be specified by freely estimating the intercepts of V_1, \dots, V_5 between levels and constraining the intercepts of the intercept, linear slope, and quadratic slope factors to be zero (W. Wu & West, 2010). The degrees of freedom of $\chi^2_{T_S_COV}$ (denoted by $df_{T_S_COV}$) are equal to the difference between the numbers of parameters in the fully saturated model and in the partially saturated model with the saturated between-mean structure.

A level-specific (l-s) χ^2 ($\chi^2_{PS_W}$) is computed for the within-level model. $\chi^2_{PS_W}$ can be obtained through Eq. B2:

$$\chi^2_{PS_W} = F_{ML}[\Sigma_B(\hat{\theta}_S), \Sigma_W(\hat{\theta})] - F_{ML}[\Sigma_B(\hat{\theta}_S), \Sigma_W(\hat{\theta}_S)], \quad (B2)$$

where $F_{ML}[\Sigma_B(\hat{\theta}_S), \Sigma_W(\hat{\theta})]$ is the value of the fitting function for the saturated between-level model. The degrees of freedom of $\chi^2_{PS_W}$ (denoted by df_{PS_W}) are equal to the difference between the numbers of parameters in the fully saturated model and the partially saturated model.

RMSEA Given conventional $\chi^2_{PS_B}$ and its corresponding df , the l-s RMSEA for the between-level model (conventional $RMSEA_{PS_B}$) can be derived by Eq. B3:

$$RMSEA_{PS_B} = \sqrt{\text{Max}\left(\frac{\chi^2_{PS_B} - df_{PS_B}}{df_{PS_B}(J)}, 0\right)}. \quad (B3)$$

In Eq. B3, J is the between-level sample size (number of groups), which functions as a penalty for large sample size. The $RMSEA_{PS_B}$ is set to zero, providing that conventional $\chi^2_{PS_B}$ is smaller than df_{PS_B} . Note that $RMSEA_{T_S_MEAN}$ and $RMSEA_{T_S_COV}$ can be also obtained via Eq. B3 by using $\chi^2_{T_S_MEAN}$ (and $df_{T_S_MEAN}$) and $\chi^2_{T_S_COV}$ (and $df_{T_S_COV}$), respectively.

$RMSEA_{PS_W}$ can be obtained from Eq. B4, where N denotes the within-level total sample size:

$$RMSEA_{PS_W} = \sqrt{\text{Max}\left(\frac{\chi^2_{PS_W} - df_{PS_W}}{df_{PS_W}(N)}, 0\right)}. \quad (B4)$$

CFI CFI is an incremental fit index used to evaluate model fit by comparing the hypothesized model to the independence model (Bentler, 1990). According to Ryu and West (2009), the l-s CFI for the between-level model (CFI_{PS_B}) can be defined as:

$$CFI_{PS_B} = 1 - \frac{\text{Max}[(\chi^2_{PS_B} - df_{PS_B}), 0]}{\text{Max}[(\chi^2_{I_B, S_W} - df_{I_B, S_W}), 0]}, \quad (B5)$$

where $\chi^2_{I_B, S_W}$ represents the χ^2 test statistic with an independence between-level model and a saturated within-level model:

$$\chi^2_{I_B, S_W} = F_{ML}[\Sigma_B(\hat{\theta}_I), \Sigma_W(\hat{\theta}_S)] - F_{ML}[\Sigma_B(\hat{\theta}_S), \Sigma_W(\hat{\theta}_S)]. \quad (B6)$$

Note that the independence between-level model is an intercept-only growth model, in which only the mean of the intercept factor and residual variances are freely estimated. Moreover, the independence between-level model needs to reflect any constraint on the residual variances in the hypothesized between-level model (Widaman & Thompson, 2003). The degrees of freedom of $\chi^2_{I_B, S_W}$ (denoted by df_{I_B, S_W}) are equal to the difference between the numbers of parameters in the fully saturated model and in the model with the independence between-level model and a saturated within-level model.

To compute $CFI_{T_S_MEAN}$ and $CFI_{T_S_COV}$, the numerator of Eq. B5 needs to be substituted with $\chi^2_{T_S_MEAN}$ (and $df_{T_S_MEAN}$) and $\chi^2_{T_S_COV}$ (and $df_{T_S_COV}$), respectively. Regarding the denominator of Eq. B5, $CFI_{T_S_MEAN}$ needs the χ^2 test statistic and df for a model with an independence between-level model with a saturated covariance structure and a saturated within-level model through Eq. B6, whereas $CFI_{T_S_COV}$ needs the χ^2 test statistic and df for a model with an independence between-level model with a saturated mean structure and a saturated within-level model.

The CFI_{PS_W} can be computed by Eq. B7:

$$CFI_{PS_W} = 1 - \frac{\text{Max}[(\chi^2_{PS_W} - df_{PS_W}), 0]}{\text{Max}[(\chi^2_{S_B, I_W} - df_{S_B, I_W}), 0]}. \quad (B7)$$

$\chi^2_{S_B, I_W}$, shown in Eq. B8, represents the χ^2 test statistic with a saturated between-level model and an independence within-level model. The independence within-level model is an intercept-only growth model, in which only residual variances are freely estimated. The independence within-level model needs to reflect any constraint on the residual variances in the hypothesized within-level model (Widaman & Thompson, 2003). The degrees of freedom of $\chi^2_{S_B, I_W}$ (denoted by df_{S_B, I_W}) are equal to the difference between the numbers of parameters in the fully saturated model and in the model with the independence within-level model and a saturated between-level model.

$$\chi^2_{S_B, I_W} = F_{ML}[\Sigma_B(\hat{\theta}_S), \Sigma_W(\hat{\theta}_I)] - F_{ML}[\Sigma_B(\hat{\theta}_S), \Sigma_W(\hat{\theta}_S)] \quad (B8)$$

TLI The TLI is a nonnormed fit index that penalizes for adding parameters in the model (Tucker & Lewis, 1973). TLI_{PS_B} can be

used to evaluate the between-level model by comparing the hypothesized between-level model and the independence between-level model under the condition that the within-level model is saturated. The information for computing conventional CFI_{PS_B} is sufficient for computing conventional TLI_{PS_B} via Eq. B9. In the same manner, the information for computing $CFI_{T_S_MEAN}$ and $CFI_{T_S_COV}$ is also sufficient for computing $TLI_{T_S_MEAN}$ and $TLI_{T_S_COV}$, respectively.

$$TLI_{PS_B} = \frac{\frac{\chi^2_{I_B,S_W}}{df_{I_B,S_W}} - \frac{\chi^2_{PS_B}}{df_{PS_B}}}{\frac{\chi^2_{I_B,S_W}}{df_{I_B,S_W}} - 1} \quad (B9)$$

TLI_{PS_W} can be used to evaluate the within-level model by comparing the hypothesized within-level model and the independence within-level model under the condition that the between-level model is saturated. The equation for TLI_{PS_W} is presented in the following equation (B10).

$$TLI_{PS_W} = \frac{\frac{\chi^2_{S_B,I_W}}{df_{S_B,I_W}} - \frac{\chi^2_{PS_W}}{df_{PS_W}}}{\frac{\chi^2_{S_B,I_W}}{df_{S_B,I_W}} - 1} \quad (B10)$$

SRMR SRMR can be computed for the within-level ($SRMR_W$) and between-level ($SRMR_B$) models. Note that SRMR can be derived from the deviation between the sample and the reproduced variance–covariance matrices. We adopted $SRMR_W$ and $SRMR_B$ as reported by Mplus in the present

study. More specifically, $SRMR_B$ reflects the normed average distance between the sample variance matrix of p observed variables and the model-implied variance matrix at the between level. $SRMR_B$ can be represented as follows:

$$SRMR_B = \sqrt{\frac{2 \sum_{i=1}^p \sum_{j=1}^i \left[\frac{(\Sigma_{Bij} - \Sigma_B(\theta)_{ij})^2}{\Sigma_{Bii} \Sigma_{Bjj}} \right]}{p(p+1)}} \quad (B11)$$

SRMR for evaluating between-mean structure ($SRMR_{T_S_MEAN}$) can be derived using Eq. B11, where the model-implied variance matrix is produced by a model with a saturated between-covariance structure and a saturated within-level model. SRMR for evaluating between-covariance structure ($SRMR_{T_S_COV}$) can be also derived by Eq. B11, where the model-implied variance matrix is produced by a model with a saturated between-mean structure and a saturated within-level model.

$SRMR_W$ reflects the normed average distance between the sample variance matrix and model-implied variance matrix at the within level. $SRMR_W$ can be represented as follows:

$$SRMR_W = \sqrt{\frac{2 \sum_{i=1}^p \sum_{j=1}^i \left[\frac{(\Sigma_{Wij} - \Sigma_W(\theta)_{ij})^2}{\Sigma_{Wii} \Sigma_{Wjj}} \right]}{p(p+1)}} \quad (B12)$$

^{B1} All the χ^2 values in the fit index equations are robust χ^2 values from MLR.

Appendix C

Appendix C Key parameter settings in the population models for each of the five misspecified conditions

Model #	Purpose	Between-Level Model			Within-Level Model	
		Covariance Structure	Mean Structure		Covariance Structure	
		β_{10}/β_{01}	β_{22}	γ_{200}	τ_{10}/τ_{01}	τ_{22}
M1	Generate replications for MIS_COV _B condition	4.390^a	.018	– .127	6.762	.703
M2	Generate replications for MIS_VAR _B condition	2.819	.105^a	– .127	6.762	.703
M3	Generate replications for MIS_MEAN _B condition	2.819	.018	– .184^a	6.762	.703
M4	Generate replications for MIS_COV _W condition	2.819	.018	– .127	6.930^a	.703
M5	Generate replications for MIS_VAR _W condition	2.819	.018	– .127	6.762	.082^a

^a To ensure that the severities of the five intentional misspecifications were the same (Fan & Sivo, 2005), we adjusted the magnitudes of key parameters in the population models to achieve a severity of misspecification equal to a power of .80 given number of cluster = 100 and cluster = 10. MIS_COV_B = misspecification of the covariance between the intercept factor and linear slope factor at the between level as 0 ($\beta_{10}/\beta_{01} = 0$). MIS_VAR_B = misspecification of the variance of the quadratic slope factor at the between level as 0 ($\beta_{22} = 0$). MIS_MEAN_B = misspecification of the mean of the quadratic factor at the between level as 0 ($\gamma_{200} = 0$). MIS_COV_W = misspecification of the covariance between the intercept factor and the linear slope factor at the within level as 0 ($\tau_{10}/\tau_{01} = 0$). MIS_VAR_W = misspecification of the variance of the quadratic slope factor at the within level as 0 ($\tau_{22} = 0$).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246. <https://doi.org/10.1037/0033-2909.107.2.238>
- Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology*, 47, 563–592.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd). Hillsdale: Erlbaum.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd). Mahwah: Erlbaum.
- Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-Index strategy revisited. *Structural Equation Modeling*, 12, 343–367.
- Gerbing, D. W., & Anderson, J. C. (1992). Monte Carlo evaluations of goodness of fit indices for structural equation models. *Sociological Methods and Research*, 21, 132–160.
- Hox, J. J. (2010). Multilevel analysis: Techniques and applications (2nd). New York: Routledge.
- Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, 8, 157–174. https://doi.org/10.1207/S15328007SEM0802_1
- Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, 64, 157–170. <https://doi.org/10.1111/j.1467-9574.2009.00445.x>
- Hsu, H.-Y., Kwok, O., Acosta, S., & Lin, J.-H. (2015). Detecting misspecified multilevel SEMs using common fit indices: A Monte Carlo study. *Multivariate Behavioral Research*, 50, 197–215. <https://doi.org/10.1080/00273171.2014.977429>
- Hsu, H.-Y., Lin, J.-H., Kwok, O.-M., Acosta, S., & Willson, V. (2016). The impact of intraclass correlation on the effectiveness of level-specific fit indices in multilevel structural equation modeling: A Monte Carlo study. *Educational and Psychological Measurement*, 77, 5–31. <https://doi.org/10.1177/0013164416642823>
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424–453. <https://doi.org/10.1037/1082-989X.3.4.424>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55. <https://doi.org/10.1080/10705519909540118>
- Ingels, S. J., Pratt, D. J., Herget, D. R., Dever, J. A., Fritch, L. B., Ottem, R., ... Leinwand, S. (2013). Education Longitudinal Study of 2002: Base-year to first follow-up data file documentation (NCES 2014-361). Washington, DC: National Center for Education Statistics.
- Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling*, 10, 128–141.
- Kaplan, D. (2009). Structural equation modeling: Foundations and extensions (2nd). Thousand Oaks: Sage.
- Kline, R. B. (2011). Principles and practice of structural equation modeling (3rd). New York: Guilford Press.
- Kwok, O.-M., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multi-level models: A Monte Carlo study. *Multivariate Behavioral Research*, 42, 557–592. <https://doi.org/10.1080/00273170701540537>
- Leite, W. L., & Stapleton, L. M. (2011). Detecting growth shape misspecifications in latent growth models: An evaluation of fit indexes. *Journal of Experimental Education*, 79, 361–381.
- Lohr, S. L. (2009). Sampling design and analysis (2nd). Boston: Brooks/Cole.
- Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods*, 16, 444–467. <https://doi.org/10.1037/a0024376>
- Ma, X., & Ma, L. (2004). Modeling stability of growth between mathematics and science achievement during middle and high school. *Evaluation Review*, 28, 104–122. <https://doi.org/10.1177/0193841X03261025>
- Ma, X., & Wilkins, J. M. (2007). Mathematics coursework regulates growth in mathematics achievement. *Journal for Research in Mathematics Education*, 38, 230–257.
- Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103, 391–410.
- Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald*. Multivariate applications book series (pp. 275–340). Mahwah: Erlbaum.
- McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, 107, 247–255. <https://doi.org/10.1037/0033-2909.107.2.247>
- Miller, J. D., Kimmel, L., Hoffer, T. B., & Nelson, C. (2000). Longitudinal study of American youth: User's manual. Evanston: Northwestern University, International Center for the Advancement of Scientific Literacy.
- Muthén, B. O. (1997). Latent variable growth modeling with multilevel data. In M. Berkane (Ed.), *Latent variable modeling and applications to causality* (pp. 149–161). New York: Springer.
- Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345–368). Newbury Park: Sage.
- Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. Hox & J. K. Roberts (Eds.), *The handbook of advanced multilevel analysis*. New York: Routledge.
- Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 25, 267–316. <https://doi.org/10.2307/271070>
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th). Los Angeles: Muthén & Muthén.
- Neuman, W. L. (2009). Social research methods: Qualitative and quantitative approaches (7th). Boston: Allyn & Bacon.
- Ryu, E. (2014). Model fit evaluation in multilevel structural equation models. *Frontiers in Psychology*, 5, 81:1–9. <https://doi.org/10.3389/fpsyg.2014.00081>
- Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, 16, 583–601. <https://doi.org/10.1080/10705510903203466>
- Scheaffer, R. L., Mendenhall, W., III, & Ott, R. L. (2005). Elementary survey sampling (6th). Belmont: Thomson Brooks/Cole.
- Schermelleh-Engel, K., Kerwer, M., & Klein, A. G. (2014). Evaluation of model fit in nonlinear multilevel structural equation modeling. *Frontiers in Psychology*, 5, 181:1–11. <https://doi.org/10.3389/fpsyg.2014.00181>

- Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., Hagedorn, M. C., Daly, P., & Najarian, M. (2015). Early Childhood Longitudinal Study, kindergarten class of 2010–11 (ECLS-K:2011), user's manual for the ECLS-K:2011 kindergarten data file and electronic codebook, public version (NCES 2015-074). Washington, DC: National Center for Education Statistics.
- Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., & Najarian, M. (2009). Early Childhood Longitudinal Study, kindergarten class of 1998–99 (ECLS-K), combined user's manual for the ECLS-K eighth-grade and K–8 full sample data files and electronic codebooks (NCES 2009-004). Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/ecls/data/ECLSK_K8_Manual_part1.pdf
- Tucker, J. S., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10.
- Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, 8, 16–37. <https://doi.org/10.1037/1082-989X.8.1.16>
- Wu, J.-Y., Kwok, O., & Willson, V. L. (2015). Using design-based latent growth curve modeling with cluster-level predictor to address dependency. *Journal of Experimental Education*, 82, 431–454. <https://doi.org/10.1080/00220973.2013.876226>
- Wu, W., & West, S. G. (2010). Sensitivity of fit indices to misspecification in growth curve models. *Multivariate Behavioral Research*, 45, 420–452.
- 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, 183–201. <https://doi.org/10.1037/a0015858>