Problems with Ignoring Clustering in Confirmatory Factor Analysis: A Monte Carlo Simulation Study

Nicholas Copeland

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Education

University of Washington

2022

Committee:

Elizabeth Sanders

Oscar Olvera Astivia

Program Authorized to Offer Degree:

College of Education

University of Washington

Abstract

Problems with Ignoring Clustering in Confirmatory Factor Analysis:

A Monte Carlo Simulation Study

Nicholas Copeland

Chair of the Supervisory Committee:

Elizabeth Sanders

Department of Measurement & Statistics

The purpose of the current study is to highlight and extend previous research on the consequences of ignoring cluster dependencies in confirmatory factor analysis (CFA) models, also known as measurement models. Although applied researchers are now well aware that they should avoid violating the independence assumption in univariate analyses (e.g., using multilevel models or unilevel regression with cluster-robust standard errors), some may not be aware that the independence assumption applies to multivariate analyses as well. Assuming the same 2-factor, tau-equivalent model at Level 1 (e.g., students) and Level 2 (e.g., schools), a relatively high within-cluster factor reliability of .90, and a modest intraclass correlation of .20, we specifically studied scenarios in which the Level 2 item-factor relations were either the same or weaker than Level 2 item-factor relations. Our Monte Carlo simulation results show that, when a unilevel factor model is fitted to clustered data, factor loading standard errors will be

substantially biased when the Level 1 and Level 2 reliabilities differ, particularly for data structures with large sized clusters. Moreover, irrespective of sample size, a lower Level 2 reliability relative to Level 1 will lead to underestimates of factor reliabilities when clustering is ignored, irrespective of sample size.

Problems with Ignoring Clustering in Confirmatory Factor Analysis: A Monte Carlo Simulation Study

Researchers in education and other social sciences are typically interested in testing hypotheses involving latent constructs that cannot be directly observed, such group comparisons on mathematics knowledge, science teaching self-efficacy, or youth sports coaching quality. In all of these scenarios, researchers must infer that the latent construct of interest is validly measured using manifest self-report or observational scales. Thus, psychometric studies of observed measures are crucial for bridging the gap between measuring a desired construct and the observed measure used to represent the construct. The most common avenue for studying construct validity is confirmatory factor analyses (CFA), which analyzes the variances and covariances among manifest variables (items) to: (a) evaluate the underlying factor structure/dimensionality in the data, (b) estimate each item's specific relationship with the latent factor, and conversely, the amount of error for each item, and (c) estimate reliability using omega in lieu of Cronbach's alpha (that latter assumes tau-equivalence: that the item-factor relations are the same across all items, which is a lower-bound on scale reliability).

Despite its usefulness, like any model, CFA assumes that the raw data used to compute the observed variances and covariances are independent (i.e., the individual scores are not systematically correlated due to clustering). In other words, that each case in the raw data contributes unique information that does not depend on other cases in the sample. And, although prior research has shown that independence violations bias CFA loading standard errors (e.g., Julian, 2001), applied researchers may be ignoring the problem because: (a) multilevel CFA involves added complexity, and (b) they may truly not be aware of the assumption in the first place. Given the importance in obtaining accurate psychometric properties for our observed

scales, the present study sought to re-emphasize the problem of ignoring clustering in CFAs, and additionally, to extend the previous research on the issue by systematically examining how differences in Level 2 factor reliability may impact loadings when clustering is ignored.

CFA in Context: Youth Sports

In the arena of youth sports, latent constructs of interest could include children's participation interest, sport self-efficacy, and moral behavior in sport. For example, Smith et al. (2008) conducted factor analyses of their Motivational Climate Scale for Youth Sports (MCSYS), which is a 12-item scale measuring the extent to which coaches fostered a "mastery climate" for the athletes on their team. A mastery climate was defined as an "emphasis on self-referenced improvement, effort, and a cooperative learning environment," which is in opposition to an ego-driven climate that "is marked by an emphasis on outperforming others, a focus on outcome, preferential attention to top performers, and punishment of mistakes" (Smith et al., 2008). Their analysis results indicated that the best fitting model was involved two factors with six items per factor (one representing mastery and the other representing ego).

Other studies of measures in youth sports have included: the Prosocial and Antisocial Behavior in Sport Scale (Kavussanu & Boardley, 2009), which measures observable moral action in competitive sport; the Leadership Scale for Sports (Chelladurai & Saleh, 1980; Jambor & Zhang, 1997), which measures leadership qualities specific to sports coaching; and the Behavioral Regulation in Sport Questionnaire-6 (Lonsdale et al., 2008), which measures 6 different types of motivational regulation in sport. In each of these instances, factor models were used to explore and gather reliability and validity evidence supporting the use of the scales—a crucial first step before the need for intervention can be established, and before programs can be evaluated for their efficacy.

CFA: Unilevel Model

In any univariate regression, it is assumed that there is no measurement error in the predictors and that all error is captured in residual variance of the outcome. In contrast, in a standard CFA, each variable, which we'll term as an "item" from a scale, is assumed to have some amount of error in its measurement, attributable to either the nature of the variable itself or the manner in which it was collected. In addition, a CFA assumes that the shared variance among the manifest (observed) variables is caused by a latent (unobserved) variable, which is inferred to be the construct of interest. The CFA model for a given item is specifically as follows.

$$Y_i = \lambda_{ij}\xi_j + \delta_i \quad (1)$$

In the model above, the i^{th} manifest item (Y_i) is a function of the strength of its relationship with the j^{th} latent factor (i.e., factor loading, λ_{ij}), the j^{th} latent factor (ξ_j), and the manifest item's residual (δ_i). To estimate the model parameters, the model-implied variance-covariance matrix is estimated as:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda^T} + \boldsymbol{\Theta} \quad (2)$$

where the model-implied variance-covariance matrix ($\boldsymbol{\Sigma}$) is a function of the product of the item-factor loading matrix ($\boldsymbol{\Lambda}$), the latent factor variance-covariance matrix ($\boldsymbol{\Phi}$), and the transpose of the variable-factor loading matrix ($\boldsymbol{\Lambda^T}$), plus the variable residual variance-covariance matrix, assumed to be a diagonal matrix ($\boldsymbol{\Theta}$) (Kline, 2016).

An example of a 2-factor, 5-item-per-factor path diagram is given in Figure 1. As can be seen, this model would scale the latent factor variance to the first item in the set of items associated with the factor, and then would allow the remaining eight item-factor relations (loadings) and residual variances to be estimated; additionally, one structural path would also be freely estimated (i.e., the correlation between the two latent factors). Note in this specification

that the relations among the first set of items and the second factor would be constrained to zero, and vice versa (this can be relaxed to allow for cross loading, if desired). Last, it is also noteworthy that, if all item loading estimates are desired (i.e., estimating the loading for the first item associated with each factor), then the two latent factors would need to be constrained to a scale of some form, typically unit normal.

The assumptions of the standard unilevel model include linearity and multivariate normality (both of which can be relaxed for models involving nominal, ordinal, or skewed data, so long as they are appropriately specified), population homogeneity (i.e., that the sample variance-covariance matrix represents a single population and that subgroups do not have different matrices), and independence of observations (the variance-covariance matrix is computed from scores that are not correlated due to clustering).

Consequences of Ignoring Clustering (Non-Independence)

It has been known by quantitative methodologists for many decades that the independence assumption is one of the most crucial assumptions underlying any statistical model (e.g., Cochran, 1976). In addition to biased variance estimates (and therefore inflated Type I error and undercoverage of confidence intervals), parameter estimates themselves may be biased when within- and between-cluster relations are not properly decomposed (e.g., Snijders & Bosker 2012, pp. 13-37). With respect to CFA models in particular, violation of independence has been far less studied. One of the only studies on this topic was conducted by Julian (2001). His work, which used a large-scale Monte Carlo simulation to study the effects of ignoring clustering on data with varied Level 1 and Level 2 factor structures, cluster sizes, and numbers of clusters, showed that factor variance estimates and fit statistics became biased across most scenarios when clustering was ignored, particularly for conditions involving large sized clusters

relative to the number of clusters. On the other hand, loading parameter estimates themselves only became biased when the factor structures at the different levels varied.

CFA: Multilevel Model

When individuals are sampled from clusters like schools and districts, dependencies in individuals' scores due to cluster membership will naturally arise if the construct being measured is related to cluster membership. For example, in the case of youth sports coaching, the extent to which a construct like mastery is encouraged by a coach may be dependent on district-level policy and initiatives, the attitudes and beliefs around sport in their community, and/or levels of access to recreational facilities and resources. Thus, if coaches from multiple districts were sampled, a unilevel CFA would be inappropriate; instead, a multilevel CFA could be used to analyze the data correctly. The model would be:

$$Y_{gi} = \mu + \alpha_g + \beta_{gi}\,, \tag{3}$$

where Y_{gi} represents the vector of observations taken from an individual i in group/cluster g, μ is an overall mean vector, and α_g and β_{gi} represent vectors of independent, random, normal variables corresponding to the between-group (or between-cluster, Level 2) and within-group (or within-cluster, Level 1) models, respectively. In essence, Level 2 data involve aggregated item means (and their variances and covariances), and level 1 data involve cluster-mean centered item values (and their variances and covariances). The model-implied variance-covariance matrices for each level are then computed separately using the same decomposition shown in (Figure 2), as within-cluster (Level 1) and between-cluster (Level 2) model-implied variance-covariance matrices, as follows:

$$\boldsymbol{\Sigma_w = \Lambda_w \Phi_w \Lambda_w^T + \Theta_w} \tag{4}$$

and

$$\boldsymbol{\Sigma_b = \Lambda_b \Phi_b \Lambda_b^T + \Theta_b}. \quad (5)$$

where the model-implied variance-covariance matrix ($\boldsymbol{\Sigma}$) *at each level* is a function of the product of the item-factor loading matrix ($\boldsymbol{\Lambda}$), the latent factor variance-covariance matrix ($\boldsymbol{\Phi}$), and the transpose of the variable-factor loading matrix ($\boldsymbol{\Lambda^T}$), plus the variable residual variance-covariance matrix, assumed to be a diagonal matrix ($\boldsymbol{\Theta}$) (Julian, 2001).

Prevalence of Ignoring Clustering in CFAs

Multilevel models and their alternatives for dealing with clustering in univariate analyses are widespread and now considered ubiquitous in education and psychology (Bauer & Sterba, 2011; Luo et al., 2021; McNeish et al., 2017). In contrast, multilevel multivariate methods may be less prevalent. The reasons for this may be threefold. First, although multilevel CFA has been available in *Mplus* (Muthén & Muthén, 2021) software for quite some time, it is currently not as easily or flexibly implemented in more widely available (and popular) lavaan structural equation modeling package in *R* (Rosseel, 2012) which is free. Second, researchers may be under the (mis)impression that the independence assumption is not of critical importance to CFA models, particularly given the lack of methodological papers highlighting the issues involved around the effects of clustering on CFA results. A third reason that multilevel CFA may not be used as frequently as one would expect in education and psychology studies is that it requires added complexity in that the measurement and structural parts of the models may need different specifications at Level 1 and Level 2. Last but not least, given that Julian's (2001) study of the consequence of ignoring clustering in CFAs found that the loading parameter estimates were relatively unaffected by cluster dependencies, applied researchers focused on the measurement portion of a model (i.e., factor reliability) may be unconcerned with the issue in truth.

To obtain some sense of the prevalence of researchers employing unilevel CFAs when clustering is present in their data, we conducted an online library search for any peer-reviewed journal articles using keywords “Education” and “CFA” published in the Academic Search Complete, APA PsycInfo, and Education Source databases, from years 2021 to 2022. The search yielded 1,160 articles, and we selected the most recent 20 non-methodological articles for review. The selected 20 articles were published in a variety of journals, including *Biomed Central*, *Journal of Psychology of Education*, *Journal of Psychopathology of Education*, and the *Journal of General Internal Medicine*, on topics that included online teaching, physical literacy, school engagement, and emotion regulation. The articles were reviewed for (a) what type of CFA (unilevel or multilevel) was used, and (2) potential sources of clustering. Of the 20 articles, potential clustering was found in 15 (75%) but none employed multilevel CFA. Types of clustering that were ignored included teachers or students within different schools, internal medicine students within from different residency programs, and service staff nested within different hotels.

Current Study

While extensive, Julian’s (2001) study findings on bias arising in unilevel CFA results when the data involves cluster dependencies were limited to loading values (and hence factor reliabilities) that were assumed to be the same at each of the levels, and further given that applied researchers appear to be largely ignoring clustering in their data when specifying their CFA models, we set out to evaluate the impact of ignoring clustering in CFAs for scenarios in which the truth is that there are differing factor reliabilities at Level 1 and Level 2. Specifically, we ask the following research question: What are the consequences for loading parameter estimates and their standard errors when clustering is ignored using unilevel CFA when (a) the Level 1 and

Level 2 factor reliabilities are the same, and (b) the Level 1 and Level 2 factor reliabilities differ? Further, do the effects of the ignored Level 2 factor reliability on bias depend on number of clusters or cluster size?

Method

To answer our research questions, a small-scale Monte Carlo simulation was conducted using *Mplus*. We generated sample data from a population 2-level, 2-factor confirmatory factor analysis (CFA) model in which the factor structure at the within-cluster (Level 1) and between-cluster (Level 2) were the same but Level 2 factor reliabilities and sample sizes were varied. Those samples were then analyzed the data using the (correct) 2-level model and the (incorrect) unilevel model. Below we describe the data generation and analysis procedures more fully.

Data Generation

Conditions held constant. Data were generated as a multilevel 2-factor model. To produce a realistic situation for education research while still isolating the conditions of interest, the following conditions were held constant.

1. Each of the two factors are measured by five indicators (items), with 10 items total; for brevity, the same items are set to load onto the same factors at both levels. This again is similar to other studies of CFA models that use 3-5 loadings per factor.
2. Level 1 factors were set to a variance of 1 and Level 2 factors were set to a variance of 0.25 so that the intraclass correlation of the factors and indicators would be set to .20, which is typical of multilevel dependencies in education research (Hedges & Hedberg, 2007).
3. The factors were moderately correlated (r = .30) at both Level 1 and Level 2, similar to many studies of CFA models. (Note: at Level 2, the correlation was set as a

covariance as .075 (correlation multiplied by the standard deviations of the two Level 2 latent variables.)

4. Residual variances of the observed indicators were set to 1 – the squared loading values; at Level 2, these values were further scaled according to the Level 2 factor variances.
5. The within-cluster Level 1 factor reliability for both factors was set at .90 by setting the loadings to a value that would produce this reliability level for a factor with $n = 5$ items (only the between-cluster Level 2 factor reliability was varied). Specifically, the loadings were solved algebraically for a given factor reliability, or omega (Ω), resulting in: $\lambda = \sqrt{\frac{1}{(1-n+(\frac{n}{\Omega}))}}$, where n = the number of items for the factor, and Ω = the desired factor reliability. With this formula, we assume that the factor and item scales are unit normal, and that the population generation model is a parallel test model since all factor loadings were set to the same value at a given level, and all residual variances were the same at a given level. We also note that we set the reliability to be higher than the typical CFA study as a conservative approach to the study of how Level 2 reliabilities might impact estimates when clustering is ignored; in future research we will vary this as well.

Conditions that were varied. Of importance, we were interested in (a) how Level 2 factor reliability (loadings) differences from Level 1 factor reliabilities might impact loading estimates when clustering was ignored, and (b) whether cluster-to-cluster size ratios might moderate bias, if any. Given that Julian (2001) found more bias in factor variance estimates in scenarios with relatively larger sized clusters, we created three conditions for each sample size

level: relatively large cluster sizes, cluster sizes equal to the number of clusters, and relatively small cluster sizes. Below are the specific conditions that were varied.

1. Two levels of cluster sample sizes (*J*) were set to either 30 or 100 to represent small and large sample sizes, respectively, and three levels of cluster sizes (*m*) per sample size were employed for each *J* size: for *J* = 30 clusters, *m* = 10, 30, and 100; for *J* = 100 clusters, *m* = 30, 100, and 300. Taken together, we had approximate *J*:*m* ratios of 0.33, 1, and 3 for each *J* to represent relatively large clusters, equality, and relatively small clusters.
2. Last but not least, factor reliability for both factors at Level 2 was set at the following three conditions: .50 (far lower than Level 1), .70 (somewhat lower than Level 1), or .90 (same as Level 1). As with the Level 1 factor loadings, we assumed tau-equivalence.

In total, we simulated 6 (*J:m* ratio) * 3 (Level 2 factor reliability level) = 18 conditions. For each of the 18 conditions, *N* = 1000 replicates were drawn; for each replicate, a 2-level CFA and a unilevel CFA was estimated. Data were generated from a multivariate normal distribution using *Mplus*, and analyses of the replicates were conducted in *Mplus*.

Simulation Results Analysis

Because the present study is a small-scale Monte Carlo simulation, descriptive statistics and related data visualizations were used to evaluate our research questions. Descriptive statistics focus on comparing the two types of CFA analyses on bias in loading estimates and their standard errors as Level 2 reliability decreases, as well as whether the bias (if any) depends on number of clusters (*J*) or cluster size (*m*). To this end, across the 1,000 replicates for each of the 18 simulation conditions, for each type of CFA analysis, we computed the mean of the 10 item loading parameter estimates across both factors (focusing on Level 1 in the 2-level model for

comparison with the unilevel model), as well as the mean of the model-based loading standard error estimates and the mean of the loading parameter standard deviations.

For factor loading parameter estimates, we report **relative bias**, defined as the difference between the mean estimate across 1,000 replications and the true value, divided by the true value. A relative bias of zero indicates no bias, whereas positive values indicate overestimation and negative values indicate underestimation, and Hoogland and Boomsma (1998) suggest that relative bias exceeding ±5% for point estimates is cause for concern.

Because standard errors are unique to each condition and analysis approach, to obtain factor loading standard error bias, we computed **empirical bias**, defined as the ratio of the mean of the observed estimated standard errors to the expected standard error, with the expected standard error based on the empirical standard deviation of the slope parameters. In other words, the expected standard errors are equal to the standard deviation of the slope estimates across replications (we note that this metric has been termed ‘‘overconfidence’’ by Beck and Katz (1995), ‘‘optimism’’ by Bell and Jones (2015), and “efficacy” by McNeish (2019)). Values of 1 indicate no bias, whereas values above 1 indicate upwardly biased standard errors, and values below 1 indicate downwardly biased standard errors. Hoogland and Boomsma (1998) suggest that relative bias exceeding ±10% (values below 0.90 or above 1.10)

Last but not least, we also report the **mean estimated factor reliability** across both factors (again, focusing on Level 1 in the 2-level model for comparison with the unilevel model) to demonstrate how bias in the loadings or their standard errors, if any, contributes to bias in overall factor reliability.

Results

In all simulation conditions, the multilevel confirmatory factor analysis (CFA) model recovered exhibited no bias (as was expected). As such, for reader interest we report the multilevel and unilevel CFA results, but focus our discussion on the unilevel CFA that ignores clustering.

Loading parameter estimate relative bias. Table 1 reports simulation condition mean relative bias in average Level 1 loading estimates by condition and analysis type, and similarly, Figure 3 highlights the relationships based on Level 2 factor reliability (our focal condition) and number of clusters relative to cluster size (*J:m* ratio). As can be seen, we found no meaningful bias in the loading parameter estimates based on our design conditions. This said, there appears to be a slight over-estimation in the loadings as Level 2 factor reliability increases for the unilevel confirmatory factor analysis (CFA) model.

Loading standard error empirical bias. Table 2 reports simulation condition mean empirical bias in average Level 1 loading standard errors by condition and analysis type, and Figure 3 highlights the relationships based on Level 2 factor reliability (our focal condition) and number of clusters relative to cluster size (*J:m* ratio). As is dramatically clear, the standard error is greatly underestimated by the unilevel CFA for nearly every condition; however, the distortion is more pronounced in samples with relatively large sized clusters, especially for samples with the largest number of clusters.

Factor reliability estimates. Table 3 and Figure 4 display results for mean model-estimated Level 1 factor reliabilities. Although the results largely mirror the pattern found for the standard errors—that is, as Level 2 factor reliability decreases, the unilevel CFA underestimates the true Level 1 factor reliability—what differs from the standard error bias findings is that the

reliability distortion is unrelated to sample size conditions. In other words, the factor reliabilities are solely dependent on the blend of the Level 1 and Level 2 factor reliabilities.

Discussion

Measuring latent constructs, such as children's sports interest or youth coaching leadership skills, with high reliability is crucial for evaluating research questions with a sufficient accuracy and confidence. However, researchers may be unaware that measurement models such as confirmatory factor analysis (CFA) can be prone to bias when multilevel data structures are used in single-level analyses. The current study sought to replicate and extend seminal work by Julian (2001) on the consequences of ignoring clustering in confirmatory factor analysis (CFA) models. Using a limited set of conditions for a 2-factor, 2-level CFA in which the Level 1 and Level 2 factor structures were assumed to be the same but the factor reliabilities (i.e., loadings) were either the same or lower for Level 2 factors compared to Level 1 factors, we generated clustered sample data and fit a unilevel 2-factor CFA to the data. Consistent with Julian (2001), our results showed that, although there was no substantial bias found in the loading parameters themselves, there was extreme downward bias in the loading standard errors. Further, we found that the standard error distortion was more marked in samples with relatively large sized clusters—a situation that would arise when collecting data from a small number of districts that have fairly large sample sizes of children or coaches.

Importantly, the unique contribution of our study to the literature is in the systematic manipulation of the factor reliabilities (loading values) at Level 1 and Level 2. We found that the standard error was increasingly underestimated as the reliability of Factor 2 decreased when clustering was ignored using the unilevel CFA. In other words, when the Factor 2 reliability was the same as the Factor 1 reliability (both at .90), there was only slight standard error

underestimation. This indicates that the effects of the clustering dependencies (intraclass correlations) on model-based standard errors in the unilevel model are amplified when the underlying (ignored) Level 2 factors have increased random error. Further, when the cluster sizes are relatively large, the deleterious effects on the loading standard errors are bolstered further.

Limitations and Future Research

There are several limitations in the current study. First, our $J = 30$ clusters conditions were problematic for the number of parameters estimated in the current model (i.e., with five indicators per factor) which caused estimation problems; in the future we will likely decrease the number of indicators per factor to three instead of five. Second, we used a fairly high Level 1 factor reliability level and did not vary this as a condition; the natural next step is to use the same levels as our Level 2 reliability conditions and to fully cross Level 1 and Level 2 reliabilities. A third major limitation is that we used only one intraclass correlation level (.20). In future work we will be sure to include lower and higher levels. Fourth, in the spirit of using a scenario similar to other methodological research, we used a 2-factor model for this study; however, we believe that the pattern of results should also be tested with single-factor models to better disentangle the number of factors from the factor reliabilities. Last, we used a tau-equivalent measurement model at Level 1 and Level 2 (i.e., loadings of equal value in each factor), which is unrealistic in real-world settings. Nevertheless, we do not believe using a random selection of loadings that culminate in the same reliability levels will meaningfully change the pattern of results observed.

Despite the limitations, the present study does offer insight into the effects of different item-factor relations at Level 1 (the individual level) and Level 2 (the aggregate cluster level) on CFA models that ignore the multilevel nature of the data. Although the loading estimates were unaffected, loading standard errors were seriously biased when the Level 2 factor reliabilities

were weaker than those of Level 1, which will naturally lead to overly liberal hypothesis tests as well as undercoverage for loading estimate 95% confidence intervals. But perhaps the most important result we found was that the overall factor reliabilities can be seriously affected: in the case of (true) lower Level 2 factor reliabilities relative to (true) Level 1 factor reliabilities, researchers who use unilevel CFA on clustered data will come away from their analysis believing their measure has lower individual-level reliability than it really does.

References

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods, 16*(4), 373-390. http://dx.doi.org/10.1037/a0025813

Beck, N., & Katz, J. N. (1995) What to do (and not to do) with time-series cross-section data. *American Political Science Review, 89*(3), 634-47. https://doi.org/10.2307/2082979

Bell, A., & Jones, K. (2015). Explaining fixed effects: Random effect modeling of time-series cross-sectional and panel data. *Political Science Research and Methods, 3*(1), 133-153. http://dx.doi.org/10.1017/psrm.2014.7

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulations of questions, design and analysis.* Stanford, CA: Stanford Evaluation Consortium. Retrieved from https://eric.ed.gov/?id=ED135801

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87 https://doi.org/10.3102/0162373707299706.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329–367. https://doi.org/10.1177/0049124198026003003

Jambor, E. A., & Zhang, J. J. (1997). Investigating leadership, gender, and coaching level using the Revised Leadership for Sport Scale. *Journal of Sport Behavior, 20*(3), 313-321. https://link.gale.com/apps/doc/A20139670/AONE?u=anon~2396dee3&sid=googleScholar&xid=2d554f73

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling, 8*(3), 325-352. doi: 10.1207/S15328007SEM0803_1

Kavussanu, M., & Boardley, I. D. (2009). The Prosocial and Antisocial Behavior in Sport Scale, *Journal of Sport and Exercise Psychology, 31*(1), 97-117. Retrieved from: journals-humankinetics.com.offcampus.lib.washington.edu/view/journals/jsep/31/1/article-p97.xml

Lonsdale, C., Hodge, K., & Rose, E. (2009). Athlete burnout in elite sport: A self-determination perspective. *Journal of Sports Sciences, 27*(8), 785-795. https://doi.org/10.1080/02640410902929366

Luo, W., Li, H., Baek, E., Chen, S., Lam, K. H., & Semma, B. (2021). Reporting practice in multilevel modeling: A revisit after 10 Years. *Review of Educational Research, 91*(3), 311-355. https://doi.org/10.3102/0034654321991229

McNeish, D. M. (2019). Effect partitioning in cross-sectionally clustered data without multilevel models. *Multivariate Behavioral Research, 54*(6), 906-925. https://doi.org/10.1080/00273171.2019.1602504

McNeish, D. M., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*(1), 114-140. https://doi.org/10.1037/met0000078

Muthén, L. K., & Muthén, B. O. (1998/2017). Mplus *User's Guide (8th Ed.).* Los Angeles, CA: Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. doi: 10.18637/jss.v048.i02

Smith, R. E., Cumming, S. P., & Smoll, F. L. (2008). Development and validation of the motivational climate scale for youth sports. *Journal of Applied Sport Psychology, 20*(1), 116-136. doi: 10.1080/10413200701790558

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis (2nd Ed.)*. Thousand Oaks, CA: Sage Publications, Inc.

Table 1

Mean Level 1 Loading Parameter Estimate Relative Bias across Conditions, by Analysis Type

	Multilevel CFA (Correct Model)			Unilevel CFA (Incorrect Model)		
Sample Sizes	L2 F Reliab = .5	L2 F Reliab = .7	L2 F Reliab = .9	L2 F Reliab = .5	L2 F Reliab = .7	L2 F Reliab = .9
$J = 30$						
$M = 10$	0.00	0.00	0.00	0.00	0.01	0.03
$M = 30$	0.00	0.00	0.00	0.01	0.01	0.03
$M = 100$	0.00	0.00	0.00	0.01	0.01	0.03
$J = 100$						
$M = 30$	0.00	0.00	0.00	0.01	0.01	0.03
$M = 100$	0.00	0.00	0.00	0.01	0.01	0.03
$M = 300$	0.00	0.00	0.00	0.01	0.02	0.03

Note. $N = 1{,}000$ replications per cell. Values closer to 0 are better; values greater than 0±.05 are of concern.

Table 2

Mean Level 1 Loading Standard Error Empirical Bias across Conditions, by Analysis Type

	Multilevel CFA (Correct Model)			Unilevel CFA (Incorrect Model)		
Sample Sizes	L2 F Reliab = .5	L2 F Reliab = .7	L2 F Reliab = .9	L2 F Reliab = .5	L2 F Reliab = .7	L2 F Reliab = .9
$J = 30$	1.00	1.00	1.00	0.72	0.77	0.83
$M = 10$	0.99	0.99	0.98	0.89	0.92	0.94
$M = 30$	1.00	1.00	1.00	0.74	0.80	0.85
$M = 100$	1.01	1.01	1.01	0.53	0.61	0.70
$J = 100$	1.01	1.01	1.01	0.54	0.60	0.68
$M = 30$	1.00	1.00	1.00	0.75	0.80	0.86
$M = 100$	1.01	1.01	1.01	0.53	0.60	0.69
$M = 300$	1.01	1.01	1.01	0.34	0.39	0.48

Note. $N = 1{,}000$ replications per cell. Values closer to 1 are better; values greater than $1 \pm .10$ are of concern.

Table 3

Mean Estimated Level 1 Factor Reliability across Conditions, by Analysis Type

	Multilevel CFA (Correct Model)			Unilevel CFA (Incorrect Model)		
Sample Sizes	L2 F Reliab = .5	L2 F Reliab = .7	L2 F Reliab = .9	L2 F Reliab = .5	L2 F Reliab = .7	L2 F Reliab = .9
$J = 30$						
$M = 10$	0.90	0.90	0.90	0.85	0.86	0.88
$M = 30$	0.90	0.90	0.90	0.85	0.86	0.88
$M = 100$	0.90	0.90	0.90	0.85	0.86	0.88
$J = 100$						
$M = 30$	0.90	0.90	0.90	0.85	0.86	0.88
$M = 100$	0.90	0.90	0.90	0.85	0.86	0.88
$M = 300$	0.90	0.90	0.90	0.85	0.86	0.88

Note. $N = 1{,}000$ replications per cell. Values closer to 0.90 are better.

Figure 1

Path Diagram for Example Unilevel 2-Factor CFA

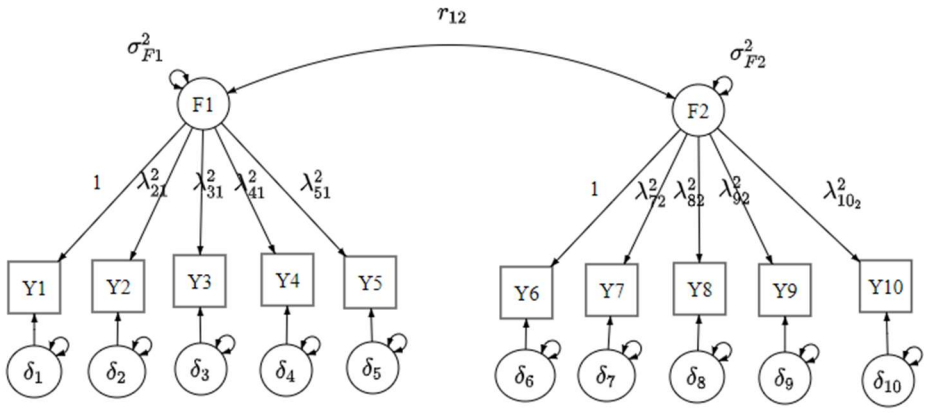

Figure 2

Path Diagram for Example Multilevel 2-Factor CFA

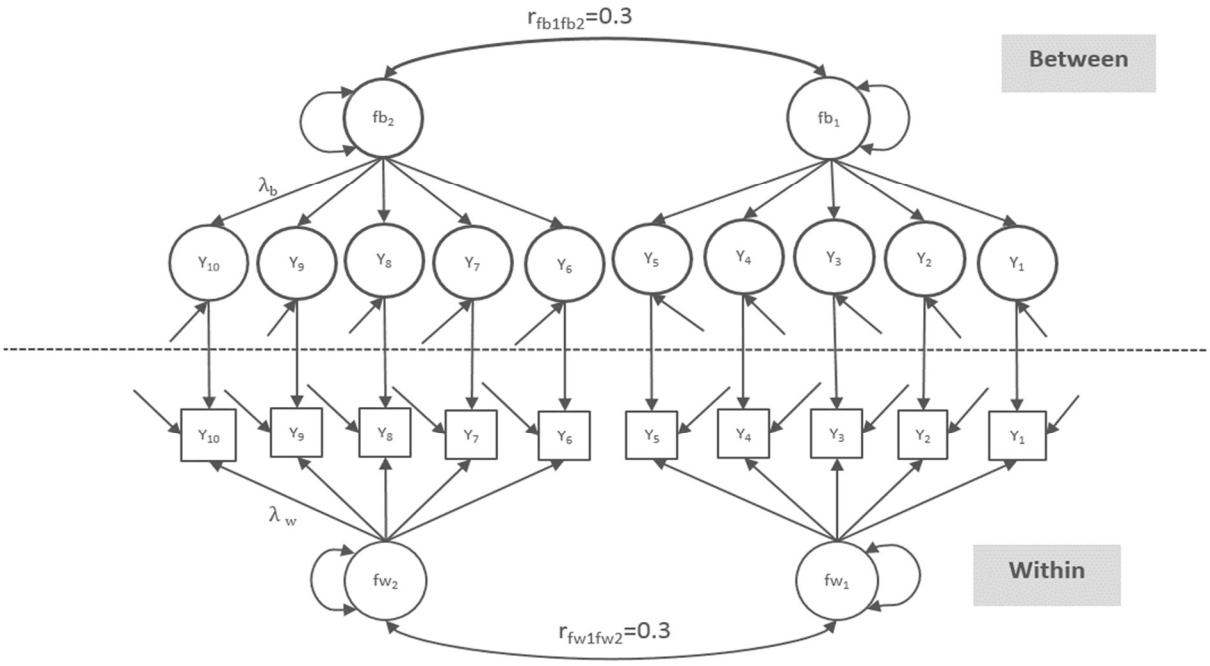

Figure 3

Illustration of Mean Level 1 Loading Estimate Relative Bias by J:m ratio and Analysis Type

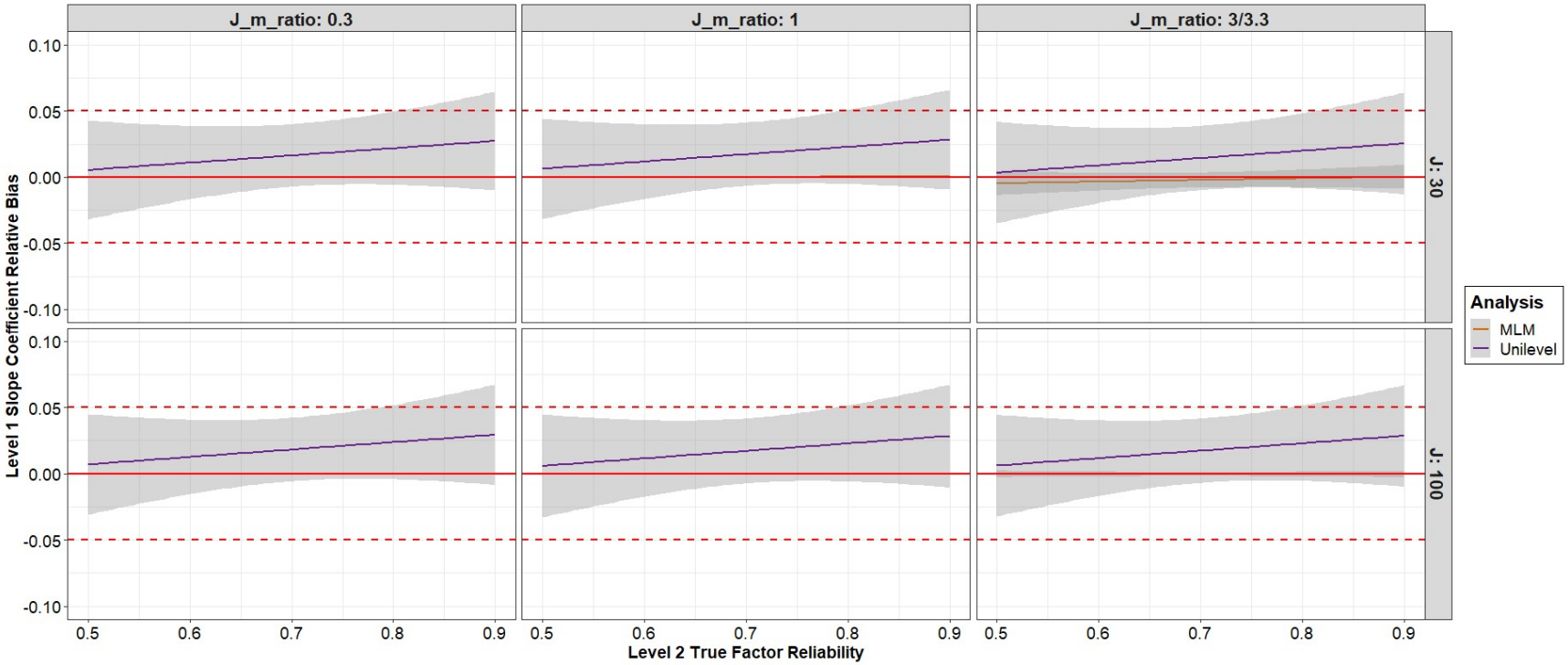

Figure 4

Illustration of Mean Level 1 Loading Standard Error Empirical Bias by J:m ratio and Analysis Type

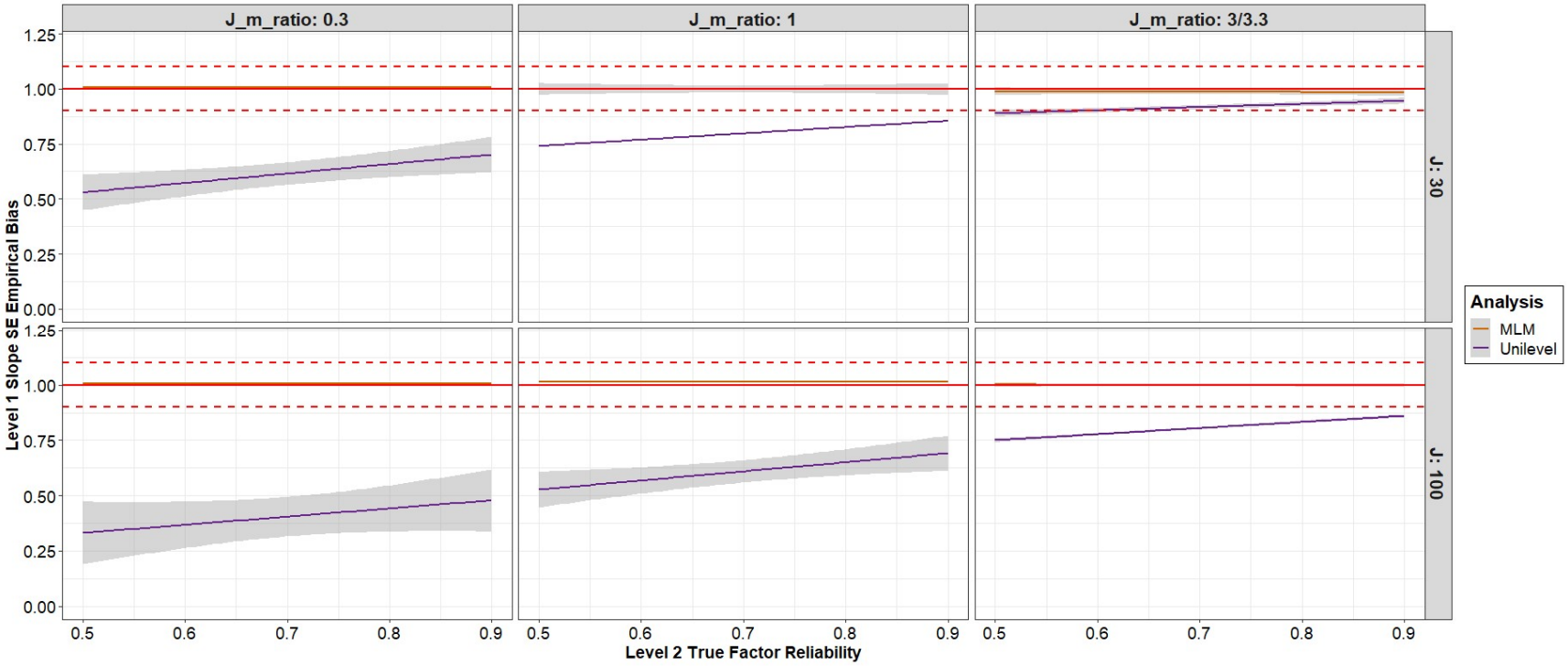

Figure 5

Illustration of Mean Level 1 Estimated Factor Reliability by J:m ratio and Analysis Type