*Multivariate Behavioral Research*, 49:518–543, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 0027-3171 print / 1532-7906 online
DOI: 10.1080/00273171.2014.933762

# Ignoring Clustering in Confirmatory Factor Analysis: Some Consequences for Model Fit and Standardized Parameter Estimates

Sunthud Pornprasertmanit and Jaehoon Lee
*Texas Tech University*

Kristopher J. Preacher
*Vanderbilt University*

In many situations, researchers collect multilevel (clustered or nested) data yet analyze the data either ignoring the clustering (disaggregation) or averaging the micro-level units within each cluster and analyzing the aggregated data at the macro level (aggregation). In this study we investigate the effects of ignoring the nested nature of data in confirmatory factor analysis (CFA). The bias incurred by ignoring clustering is examined in terms of model fit and standardized parameter estimates, which are usually of interest to researchers who use CFA. We find that the disaggregation approach increases model misfit, especially when the intraclass correlation (ICC) is high, whereas the aggregation approach results in accurate detection of model misfit in the macro level. Standardized parameter estimates from the disaggregation and aggregation approaches are deviated toward the values of the macro- and micro-level standardized parameter estimates, respectively. The degree of deviation depends on ICC and cluster size, particularly for the aggregation method. The standard errors of standardized parameter estimates from the disaggregation approach depend on the macro-level item communalities. Those from the aggregation approach underestimate the standard errors in multilevel CFA (MCFA), especially when ICC is low. Thus, we conclude that MCFA or an alternative approach should be used if possible.

In many circumstances researchers collect data that have an internal multilevel structure. For example, researchers may administer questionnaires to students in different classrooms, to participants across several states, or to employees from many departments. In doing so, they often ignore the multilevel structure of the data and use analytic methods designed for single-level data. This practice can cause inaccurate parameter estimates and standard errors (*SE*s) of target parameters as well as inflated or deflated Type I error (Moerbeek, 2004),[1] depending on which level researchers ignore. They might lose an opportunity to capture relationships among variables at different levels. Moreover, if the relationships differ across levels (e.g., if the effect of socioeconomic status on academic achievement is stronger at the school level than at the student level), not only can researchers not acquire any information about the effect at the ignored level but also the effect in the desired level might be distorted (Raudenbush & Bryk, 2002).

The differences in parameter estimates yielded by single-level methods compared with the methods accounting for clustering (e.g., multilevel models) depend on several aspects of the data. The proportion of overall variability in a variable explained by macro (Level 2) units, indicated by intraclass correlation (ICC), is one of the most important factors (Chen, Kwok, Luo, & Willson, 2010; Julian, 2001; Kim, Kwok, & Yoon, 2012; Moerbeek, 2004). If ICC is higher,

Correspondence concerning this article should be addressed to Sunthud Pornprasertmanit, Institute for Measurement, Methodology, Analysis, and Policy, Texas Tech University, Lubbock, TX 79409. E-mail: sunthud.pornprasertmanit@ttu.edu

[1] Moerbeek (2004) illustrated the consequences of ignoring a level in three-level data. Moerbeek's findings can be applied to two-level data by fixing Level 1 error variance ($\widehat{\sigma}_e^2$) to 0 and Level 1 sample size ($n_1$) to 1. The class level in her study can be thought of as the micro level and the school level can be thought of as the macro level.

ignoring the macro level (disaggregation) will be more detrimental. For example, the *SE*s of regression coefficients tend to be greater or less under disaggregation, which leads to deflated or inflated Type I error (Moerbeek, 2004; Opdenakker & Van Damme, 2000). In the case of confirmatory factor analysis (CFA), Julian (2001) showed that parameter estimates—including factor loadings, unique variances, factor variances, and factor covariances—are higher than the parameter values at the micro level and their *SE*s are lower under disaggregation. Chi-square model fit statistics are also increased, leading to inflated Type I error such that true models are rejected at a rate greater than the nominal level. Note that inflated Type I error is an indication that clustering is ignored in disaggregated analysis (Stapleton, 2006).

Although it has been 13 years since Julian's (2001) article was published, from our brief search we could still find many recent applications of factor analysis that do not take clustering into account when the data may show nontrivial ICCs (e.g., Babakus, Bienstock, & Van Scotter, 2004; Cassidy, Hestenes, Hegde, Hestenes, & Mims, 2005; C. J. Collins & Smith, 2006; Detert, Schroeder, & Cudeck, 2003; Ebesutani, Okamura, Higa-McMillan, & Chorpita, 2011; Garb, Wood, & Fiedler, 2011; Glisson & James, 2002; González-Romá, Peiró, & Tordera, 2002; Han, Chou, Chao, & Wright, 2006; Hatami, Motamed, & Ashrafzadeh, 2010; Keller, 2001; Law, Shek, & Ma, 2011; Merrell, Felver-Gant, & Tom, 2011; Nelson, Canivez, Lindstrom, & Hatt, 2007; Oliver, Jose, & Brough, 2006; Patterson et al., 2005; Philips et al., 2006; Raspa et al., 2010; Riordan, Vandenberg, & Richardson, 2005; Robert & Wasti, 2002; Salanova, Agut, & Peiró, 2005; Schaubroeck, Lam, & Cha, 2007; Takeuchi, Lepak, Wang, & Takeuchi, 2007; C. M. Tucker et al., 2011; van der Vegt & Bunderson, 2005; Zohar & Tenne-Gazit, 2008) and only a few studies that properly account for clustering (Breevaart, Bakker, Demerouti, & Hetland, 2012; Jackson, Levine, & Furnham, 2003; Mathisen, Torsheim, & Einarsen, 2006; Subramony, Krause, Norton, & Burns, 2008; Wang, Willett, & Eccles, 2011). Throughout this article, we show that a few methods accounting for clustering in CFA are readily applicable in such situations, and we encourage researchers to avoid disaggregation. Note that there are some articles that illustrate those methods for applied researchers (Cheung & Au, 2005; Cheung, Leung, & Au, 2006; Dyer, Hanges, & Hall, 2005; B. O. Muthén & Satorra, 1995; Stapleton, 2006; Zyphur, Kaplan, & Christian, 2008).

The American Psychological Association (APA; 2010) and Wilkinson and the Task Force on Statistical Inference (1999) encouraged researchers to use effect sizes along with test statistics. APA (2010) stressed that "it is almost always necessary to include some measure of effect size in the Results section" (p. 34). Therefore, we also investigate the effects of disaggregation on the effect sizes in CFA using *standardized* parameter estimates (i.e., standardized factor loadings and factor correlations), which may be different from Julian's (2001) results pertaining to *unstandardized* parameter estimates. Standardized factor loadings and factor correlations are computed as ratios of unstandardized parameters (e.g., a correlation is the ratio of a covariance and two standard deviations). In this article, we show that in some circumstances the differences in unstandardized parameters are canceled out and, as a result, single-level analyses can yield *unbiased* standardized parameter estimates. Standardized factor loadings are almost always used in interpreting results of a factor analysis (Floyd & Widaman, 1995) and factor correlations directly describe the relationships among factors in a universally understandable way. Approximately two thirds of applied factor-analytic studies cited in this article reported standardized factor loadings.[2] Therefore, this study builds on the existing literature by exploring the influences of ignoring nested data on CFA parameter estimates that are commonly used and reported. Applied researchers will want to know the consequences of ignoring the nested data structure on standardized parameter estimates.

Julian (2001) demonstrated the effects of disaggregation, or ignoring the macro level and analyzing only the micro-level units. However, the aggregation method, which involves averaging variables within each macro-level unit and analyzing only these macro-level averages, is another way that researchers might ignore the multilevel structure of nested data. For example, students could be assessed within each classroom, but the student data could be averaged within each classroom and the averaged classroom-level data used in the analysis. In regression analysis, ignoring the micro level does not affect the *SE*s of regression coefficients in the macro level (Moerbeek, 2004). From our brief search we found many applications of factor analysis using aggregated data (Bell, Mengüç, & Stefani, 2004; Ehrhart, 2004; Gibson & Birkinshaw, 2004; Gong, Chang, & Cheung, 2010; Hoegl, Parboteeah, & Munson, 2003; King & Figueredo, 1997; Mathisen, Einarsen, Jørstad, & Brønnick, 2004; Zhou, Gao, Yang, & Zhou, 2005). We also found two articles that ran separate factor analyses on the same but disaggregated or aggregated data (Håvold, 2007; Hoegl & Gemuenden, 2001). In most cases, multilevel CFA (MCFA), which is one way of accounting for nested data structure, should be applied to any data sets with nested structure. Instead of using aggregation, MCFA accurately estimates the macro-level parameters of

[2] Almost all substantive studies that we classified as reporting standardized loadings used the term *loadings* or *factor loadings* to refer to standardized factor loadings. We checked that the reported "loadings" or "factor loadings" were in fact standardized. If researchers provided any guidelines on the nontrivial values of factor loadings, we assumed that they implied standardized factor loadings. Also, if all values of "loadings" or "factor loadings" were not over 1, we assumed that they reported standardized loadings. In the marker variable method of scale identification, at least one loading is fixed at 1. In the fixed factor method of scale identification, all unstandardized factor loadings would be less than 1 when standard deviations of observed variables are close to or lower than 1. In such cases, we assumed the observed variables had been standardized and thus the reported loadings were standardized.

interest. The only circumstances under which it is appropriate to apply CFA to aggregated data are when (a) the target construct is in a formative measurement model (see the next two sections for more details about formative and reflective measurement models) and (b) the sampling ratio (i.e., the proportion of the observed cluster size and total cluster size of a group) is close to 1 (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011), such as when measuring job stress of team members (nested in teams) when almost all members answered the questionnaires (Keller, 2001). Unfortunately, not many analyses fall into this category. Thus, we focus on the more common case where MCFA is required and explore the effects of ignoring the micro level in CFA. Therefore, we show that the standardized parameter estimates and their *SE*s can be biased due to aggregation.

Although ignoring one level can lead to inaccurate results depending on researchers' target parameters, ignoring clustering is sometimes necessary. Researchers may not have information on macro-level units that is necessary for multilevel analyses. For example, the school variable may be inadvertently lost when researchers merge data sets. Also, they may find that multilevel analyses do not converge (Ryu & West, 2009), especially in data characterized by low ICCs. In addition, specifying a multilevel structure in the analysis model creates additional parameters that may not be directly relevant to the goal of the study—for instance, researchers may be interested in the factor structure at only the individual level. In such cases, researchers often choose single-level CFA instead of MCFA, thereby ignoring either the micro or macro level. Thus, using Monte Carlo simulation, we investigate when the effect of ignoring clustering is minimal or significant, focusing on ICC and other important factors. At the end of this article, we also discuss several alternatives to avoid ignoring the nested data structure.

The goal of this study is to extensively investigate the effects of ignoring the nested structure of data (either the macro or micro level) by using single-level CFA rather than MCFA on standardized measures. The organization of this article is as follows. First, we explain CFA and MCFA and the techniques used for model fit evaluation. We then discuss the meanings of parameters from disaggregated CFA, aggregated CFA, and different types of MCFA. We show the effects of disaggregation and aggregation on standardized parameter estimates analytically. Next, we provide the results of our simulation studies that examine the effects of ignoring clustering. We also illustrate the consequences of ignoring nested data structure in a real data set. Finally, we summarize the study findings and provide recommendations for handling nested data in CFA.

## CONFIRMATORY FACTOR ANALYSIS MODEL

Factor analysis is a statistical technique used to identify latent variables that explain relationships among observed variables. This technique is based on partitioning the variance of each variable into two major parts: common factor variance (that part of a variable's variance explained by common factors) and unique variance (uniqueness; that part of a variable's variance not explained by common factors). In CFA, researchers have hypotheses about the common factor structure in advance (i.e., the number of factors and the assignment of observed variables to factors), whereas in exploratory factor analysis, the goal often is to identify the number of interpretable factors that explain covariances among observed variables. We focus on the case of CFA. The results from CFA include model fit indices showing the degree to which a hypothesized model accurately represents relationships among observed variables. Because CFA is a special case of structural equation modeling (SEM), fit indices from SEM are often used, such as the chi-square statistic, root mean square error of approximation (RMSEA; Steiger & Lind, 1980), standardized root mean square residual (Jöreskog & Sörbom, 1981), comparative fit index (CFI; Bentler, 1990), and Tucker-Lewis Index (TLI; L. R Tucker & Lewis, 1973).

As with other statistical models, CFA has several assumptions. One of the most important assumptions is independence of observations. This assumption is violated when cases are organized into natural clusters (e.g., students can be grouped based on schools), which leads to a nested data structure. In such cases, using CFA that is designed for single-level data can lead to biased parameter estimates and *SE*s (Julian, 2001). MCFA, which is a submodel of the multilevel SEM (MSEM) framework, was developed to address these issues (Ansari, Jedidi, & Dube, 2002; Ansari, Jedidi, & Jagpal, 2000; Chou, Bentler, & Pentz, 2000; Jedidi & Ansari, 2001; B. O. Muthén & Asparouhov, 2009; Rabe-Hesketh, Skrondal, & Pickles, 2004; Rabe-Hesketh, Skrondal, & Zheng, 2007; Raudenbush & Sampson, 1999; Skrondal & Rabe-Hesketh, 2004). MCFA can account for the covariation of observed variables in more than one level of the data hierarchy. The next sections illustrate models and assumptions underlying both single- and multilevel CFA.

### Single-Level CFA

The CFA model separates observed scores of an individual into three components: a measurement intercept, a part due to common factor, and a part due to unique factor:

$$\mathbf{Y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\mathbf{Y}_i$ is a *p*-dimensional vector of observed variables for individual *i*; $\boldsymbol{\mu}$ is a *p*-dimensional vector of observed means; $\boldsymbol{\Lambda}$ is a $p \times m$ factor loading matrix, where *m* indicates the number of latent variables; $\boldsymbol{\eta}_i$ is an *m*-dimensional vector of latent variable scores for individual *i*; and $\boldsymbol{\varepsilon}_i$ is a *p*-dimensional vector of unique factors. In this model, $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are constant across individuals, $\boldsymbol{\eta}_i \sim N(\mathbf{0}_m, \boldsymbol{\Psi})$, and $\boldsymbol{\varepsilon}_i \sim N\left(\mathbf{0}_p, \boldsymbol{\Theta}\right)$, where $\mathbf{0}_m$ and $\mathbf{0}_p$ are *m*- and *p*-dimensional zero vectors, $\boldsymbol{\Psi}$ is an $m \times m$ matrix of latent variable covariances, and $\boldsymbol{\Theta}$ is a $p \times p$ matrix of unique factor covariances. Therefore, $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Theta}$ contain parameters to be estimated. The most popular way to estimate these parameters is via maximum likelihood (ML).

The CFA model can be reparameterized to estimate standardized coefficients by

$$\mathbf{Y}_i - \boldsymbol{\mu} = \mathbf{D}_Y^{1/2}\mathbf{Y}_i^* \tag{2}$$

and

$$\mathbf{Y}_i^* = \boldsymbol{\Lambda}^*\boldsymbol{\eta}_i^* + \boldsymbol{\varepsilon}_i^*, \tag{3}$$

where $\mathbf{Y}_i^*$ contains standardized indicators; $\mathbf{D}_Y$ is a $p \times p$ diagonal matrix containing variances of indicators on the diagonal; $\boldsymbol{\Lambda}^*$ is a $p \times m$ matrix of standardized factor loadings; $\boldsymbol{\eta}_i^*$ is an $m$-dimensional vector of factor scores in the standardized scale; and $\boldsymbol{\varepsilon}_i^*$ is a $p$-dimensional vector of the rescaled unique factors, the variances of which are nonlinearly constrained by

$$diag\left(\boldsymbol{\Theta}^*\right) = diag\left(\mathbf{I} - \boldsymbol{\Lambda}^{T*}\boldsymbol{\Psi}^*\boldsymbol{\Lambda}^*\right), \tag{4}$$

where the *diag* function extracts only the unique variances and fixes the covariances to 0, $\boldsymbol{\Theta}^*$ is the $p \times p$ scaled unique factor covariance matrix, and $\boldsymbol{\Psi}^*$ is the $m \times m$ factor correlation matrix.

In ML, scores are assumed independent across individuals. If any observations are correlated and this correlation is not modeled, the ML likelihood function will be inflated, resulting in bias in fit indices and *SE*s. One way to account for dependency among observations is to use MCFA or, in general, MSEM.

## CFA Accounting for the Nested Data Structure

There are two general ways to account for nested data structure: design-based or model-based approaches (B. O. Muthén & Satorra, 1995; Stapleton, 2002; Sterba, 2009). The design-based framework is based on defining how a sample has been drawn from a target population in a nested data structure (Asparouhov, 2005; Kaplan & Ferguson, 1999; B. O. Muthén & Satorra, 1995; Stapleton, 2002, 2006, 2008). Then, selection probabilities are created for each micro-level unit (accounting for the size of each cluster; see Sterba, 2009 for an example). The selection probabilities are used to create sampling weights such that micro-level units with low selection probabilities will be given more weight. For example, with large-scale survey data sets, such as the National Assessment of Educational Progress or the Program for International Student Assessment (PISA), how each micro-level unit is drawn is carefully described, and sampling weights are provided for all micro-level units. Finally, CFA is implemented while incorporating sampling weights, which are the inverse selection probabilities of each case (Kaplan & Ferguson, 1999).

Macro-level units are accounted for in the data analysis by estimation procedures. Parameter estimates and *SE*s are aggregated across clusters, such as with Taylor series linearization (Asparouhov, 2005; Asparouhov & Muthén, 2005). The aggregation accounts for the fact that the probabilities of all micro and macro units are drawn from a finite population through the use of sampling weights. Note that if the probabilities of selecting each macro-level unit and each micro-level unit are equal, the sampling weights are equal and some programs (e.g., Mplus) can appropriately normalize equal weights and provide accurate parameter estimates and *SE*s. However, results from the design-based framework represent neither micro- nor macro-level factor structure. We discuss the meanings of the parameters from design-based CFA in the next section. Because the factor structure at both micro and macro levels cannot be examined by this approach, we do not focus on the design-based approach in this study.

MCFA is a model-based framework that includes higher level units into the models. MCFA is the extension of CFA to the analysis of nested data. A score in a nested data structure can be influenced by two latent variables: the latent cluster score and latent individual score (Lüdtke et al., 2008):

$$\mathbf{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\upsilon}_j + \boldsymbol{\upsilon}_{ij}, \tag{5}$$

where $i$ indexes individual cases and $j$ indexes clusters, $\mathbf{Y}_{ij}$ is a $p$-dimensional vector of observed variables for individual $i$ in cluster $j$, $\boldsymbol{\mu}$ is a $p$-dimensional vector of grand means, $\boldsymbol{\upsilon}_j$ is a $p$-dimensional vector containing latent cluster (macro-level deviation) scores for cluster $j$, and $\boldsymbol{\upsilon}_{ij}$ is a $p$-dimensional vector containing latent individual (micro-level deviation) scores for individual $i$ in cluster $j$.

Similar to single-level CFA, both latent cluster scores and latent individual scores also can be classified into two sources of variation: common factors and unique factors. The micro-level (or *Within*) measurement model is

$$\boldsymbol{\upsilon}_{ij} = \boldsymbol{\Lambda}_W\boldsymbol{\eta}_{Wij} + \boldsymbol{\varepsilon}_{ij}, \tag{6}$$

where $\boldsymbol{\Lambda}_W$ is a $p \times m$ micro-level factor loading matrix, where $m$ indicates the number of micro-level latent variables, $\boldsymbol{\eta}_{Wij}$ is an $m$-dimensional vector of micro-level latent variable scores for individual $i$ in cluster $j$, and $\boldsymbol{\varepsilon}_{ij}$ is a $p$-dimensional vector of micro-level unique factors. $\boldsymbol{\eta}_{Wij}$ is multivariate normally distributed with zero means and $m \times m$ micro-level covariance matrix $\boldsymbol{\Psi}_W$. $\boldsymbol{\varepsilon}_{ij}$ is multivariate normally distributed with zero means and $p \times p$ micro-level covariance matrix $\boldsymbol{\Theta}_W$. The macro-level (or *Between*) measurement model is

$$\boldsymbol{\upsilon}_j = \boldsymbol{\Lambda}_B\boldsymbol{\eta}_{Bj} + \boldsymbol{\zeta}_j, \tag{7}$$

where $\boldsymbol{\Lambda}_B$ is a $p \times h$ macro-level factor loading matrix, where $h$ indicates the number of macro-level latent variables, $\boldsymbol{\eta}_{Bj}$ is an $h$-dimensional vector of macro-level latent variable scores for cluster $j$, and $\boldsymbol{\zeta}_j$ is a $p$-dimensional vector of macro-level unique factors. $\boldsymbol{\eta}_{B_j}$ is multivariate normally distributed with zero means and $h \times h$ micro-level covariance matrix $\boldsymbol{\Psi}_B$. $\boldsymbol{\zeta}_j$ is multivariate normally distributed with zero means and $p$

$\times\, p$ micro-level covariance matrix $\mathbf{\Theta}_B$. Note that the factors in both levels can represent the same or different constructs across levels (Klein & Kozlowski, 2000). As an example of different constructs, self-efficacy might represent the relative efficacy of individuals (to their peers within a group) at the micro level but collective team efficacy at the macro level.

The model described earlier is referred to as a random intercept model. Almost any parameters, however, can be allowed to vary across clusters. For example, factor loading values can be normally distributed across clusters. A model with different path coefficients across clusters is referred to as a random effect model (B. O. Muthén & Asparouhov, 2009). We, however, concentrate only on models with random intercepts.

The MCFA model can be reparameterized to estimate standardized coefficients by rescaling the micro- and macro-level deviation scores:

$$\boldsymbol{v}_{ij} = ((\mathbf{I} - \mathbf{P})\,\mathbf{D}_Y)^{1/2}\,\boldsymbol{v}_{ij}^{*} \tag{8}$$

$$\boldsymbol{v}_{j} = (\mathbf{P}\mathbf{D}_Y)^{1/2}\,\boldsymbol{v}_{i}^{*}, \tag{9}$$

where $\mathbf{I}$ is a $p \times p$ identity matrix, $\mathbf{P}$ is a $p \times p$ diagonal matrix containing ICCs of all variables, $\mathbf{D}_Y$ is a $p \times p$ diagonal matrix containing the total observed variances of all variables, $\boldsymbol{v}_{ij}^{*}$ is a vector of the standardized micro-level scores of individual $i$ in cluster $j$, and $\boldsymbol{v}_{i}^{*}$ is a vector of the standardized macro-level scores of cluster $j$. The measurement models of the standard scores of both levels are

$$\boldsymbol{v}_{ij}^{*} = \mathbf{\Lambda}_W^{*}\boldsymbol{\eta}_{Wij}^{*} + \boldsymbol{\varepsilon}_{ij}^{*} \tag{10}$$

$$\boldsymbol{v}_{j}^{*} = \mathbf{\Lambda}_B^{*}\boldsymbol{\eta}_{Bj}^{*} + \boldsymbol{\zeta}_{j}^{*}, \tag{11}$$

where $\mathbf{\Lambda}_W^{*}$ and $\mathbf{\Lambda}_B^{*}$ are micro- and macro-level standardized factor loading matrices, $\boldsymbol{\eta}_{Wij}^{*}$ and $\boldsymbol{\eta}_{Bj}^{*}$ are the micro- and macro-level factor scores in the standardized scale, and $\boldsymbol{\varepsilon}_{i}^{*}$ and $\boldsymbol{\zeta}_{j}^{*}$ are the micro- and macro-level unique factors. The micro- and macro-level unique variances are nonlinearly constrained by

$$diag\left(\mathbf{\Theta}_W^{*}\right) = diag\left(\mathbf{I} - \mathbf{\Lambda}_W^{T*}\mathbf{\Psi}_W^{*}\mathbf{\Lambda}_W^{*}\right) \tag{12}$$

$$diag\left(\mathbf{\Theta}_B^{*}\right) = diag\left(\mathbf{I} - \mathbf{\Lambda}_B^{T*}\mathbf{\Psi}_B^{*}\mathbf{\Lambda}_B^{*}\right), \tag{13}$$

where $\mathbf{\Theta}_W^{*}$ and $\mathbf{\Theta}_B^{*}$ are micro- and macro-level scaled unique factor covariance matrices and $\mathbf{\Psi}_W^{*}$ and $\mathbf{\Psi}_B^{*}$ are micro- and macro-level factor correlation matrices.

This model is estimated by full information maximum likelihood (FIML; see B. O Muthén & Asparouhov, 2009, for further details). A chi-square statistic is provided to determine the amount of misfit in the model. This chi-square statistic reflects misfit in both micro and macro levels (Ryu & West, 2009). Other fit indices, such as CFI, TLI, and RMSEA, are available to evaluate overall model fit as well. These fit indices tend to represent misfit in the micro level much more than the macro level (Ryu & West, 2009). To separate the misfit into components specific to the macro and micro levels, Ryu and West (2009) proposed that saturation in one level is needed. If researchers wish to quantify the misfit in the micro level, researchers need to saturate the macro-level model, such as by specifying all possible covariances among variables to be freely estimated. If researchers wish to find the misfit in the macro level, the micro-level model needs to be saturated. This strategy is useful only in case of random intercept models.

This model can also be estimated by the segregation approach (B. O. Muthén, 1989, 1990, 1994; Yuan & Bentler, 2007) involving two steps. First, the covariance matrices of the micro-level deviations ($\boldsymbol{v}_{ij}$) and macro-level deviations ($\boldsymbol{v}_{j}$) are estimated. Second, the models at each level can be estimated by ML or other estimators. This approach can be used only for random intercept models and when the cluster size is uniform (Yuan & Bentler, 2007). Because the segregation approach is less general than MCFA using FIML, we focus only on the latter approach in this study.

Although MCFA using FIML is an appropriate analytic approach for nested data, the models do not always converge, especially when ICC is low (Ryu & West, 2009). The segregation approach is not easily employed using popular SEM packages. Thus, sometimes researchers may deem it necessary to ignore clustering when MCFA does not converge. Using single-level CFA on multilevel data, however, will violate the assumption of independent observations. Therefore, it is necessary to know the effects of ignoring clustering when MCFA cannot be used.

## THE MEANINGS BEHIND TARGET PARAMETERS FOR DISAGGREGATED, AGGREGATED, AND DESIGN-BASED CFA AND MCFA

A construct can be represented by two types of measurement models: formative or reflective. Note that the terms *formative* and *reflective* are used here in a purely conceptual sense to describe the nature of aggregation (Lüdtke et al., 2008), unlike the terms used to distinguish two opposing measurement models (Bollen, 1989). Sometimes, the terms *formative* and *reflective* in aggregation are referred to as *direct consensus* and *referent-shift consensus,* respectively (van Mierlo, Vermunt, & Rutte, 2009).

In a formative measurement model, latent variables at the micro level represent the properties of micro-level units. Macro-level constructs are the collective properties of micro-level latent variables, such as means or variances of the micro-level latent variables. For example, items measuring individual neuroticism can be considered formative indicators of a neuroticism construct. Any statistic about neuroticism (e.g., average or maximum) can be considered a macro-level formative construct. The relationships among variables at the micro and macro levels can be represented by factor-analytic models.

On the other hand, the primary unit in a reflective measurement model is the macro-level unit. For example, if employees in a department answer items measuring their supervisor's performance, the measurement target is at the macro level (supervisor). The relationships among variables at the macro level can be represented by a factor-analytic model. The differences within a macro-level unit represent systematic differences unique for each micro-level unit (e.g., differences in rating styles) and measurement errors. If researchers are interested in the micro-level difference, they may construct a factor structure to explain the relationships among variables at the micro level. Otherwise, they can simply estimate all possible covariances among variables at the micro level to avoid model misspecification. For example, suppose that each student rates the hostility level in a classroom. The individual differences within the classroom may indicate classroom dominance hierarchy or bullying behaviors. Thus, researchers may want to model the shared systematic student differences as well. See Lüdtke et al. (2011), Lüdtke et al. (2008), and Marsh et al. (2012) for further details.

Different analytic methods in the previous section treat nested data in different ways resulting in different meanings of the same parameters (e.g., factor correlations). The meanings depend on whether the construct under investigation is formative or reflective. In MCFA, micro-level parameter estimates represent the relations among variables controlled for the influence of clusters. For example, a micro-level standardized factor loading in an independent cluster factor pattern (i.e., each indicator loads on only one factor) represents the correlation between micro-level variations of a variable and scores of a micro-level latent variable. In other words, this loading represents the change in the expected value of a variable in standard deviation units if a latent variable increases by one standard deviation within a macro-level unit. In a pure formative measurement model (Lüdtke et al., 2011; Lüdtke et al., 2008), researchers can interpret the meaning of a micro-level latent variable based on the magnitudes of the micro-level standardized factor loadings. That is, the meanings of micro-level latent variables should be based on the items with high micro-level standardized factor loadings. Note that a micro-level factor structure is not needed in a pure reflective measurement model. If a micro-level structure is specified in a pure reflective model, the interpretation is similar to the micro-level structure in a pure formative model.

Macro-level parameter estimates in MCFA represent the relations among variables at the macro level. For example, a macro-level standardized factor loading for an independent cluster factor pattern represents the correlation between macro-level variations of a variable and scores of a macro-level latent variable. Its squared value corresponds to the proportion of macro-level variance of a variable explained by a macro-level latent variable. In a formative measurement model, factor loadings reflect the degree of differences in measurement intercepts across clusters. If scalar invariance is established across macro-level units, macro-level measurement unique variances are equal to 0 and macro-level standardized factor loadings are equal to 1 (Jak, Oort, & Dolan, 2013). If metric invariance holds but scalar invariance is not established, macro-level measurement errors represent the variances of the intercepts across macro-level units. Then, a macro-level standardized loading is negatively related to the degree of the differences between the intercepts of a variable across macro-level units. Thus, if researchers have a factor structure at the micro level, they should be careful in interpreting macro-level standardized loadings. For example, if scalar invariance is established, all macro-level standardized loadings would be 1 even though a macro-level factor represents the means of micro-level factors. On the other hand, in a pure reflective measurement model, macro-level standardized factor loadings should be used to interpret the meanings behind macro-level latent variables.

Disaggregated CFA investigates the relations of variables uncontrolled for the influences of clusters. For example, a disaggregated standardized factor loading for an independent cluster factor pattern represents the correlation between an observed variable and a latent variable without considering the cluster membership of each case. In other words, this loading represents the change in the expected value of a variable in standard deviation units if a latent variable increases by one standard deviation, regardless of cluster membership. Loosely speaking, target parameters from MCFA are similar to regression coefficients from group-mean centering, whereas those from disaggregated CFA are similar to regression coefficients from grand-mean centering. Aggregated CFA, on the other hand, explains the relations among macro-level units so that the meanings of parameters are the same as macro-level parameters.

Design-based CFA also explains the relations of variables uncontrolled for the influences of clusters (Stapleton, 2006). Design-based CFA and disaggregated CFA differ mostly in terms of the type of population (finite vs. infinite) being targeted. Design-based CFA requires a well-defined finite population (e.g., all high school students in Florida) and parameters represent the relations of variables within the defined finite population. Disaggregated CFA, however, is based on the model-based approach in which the target population is infinite. A model-based analysis can generalize the results derived from a sample to the hypothetical, infinite population under certain conditions (Sterba, 2009). That is, if a model is correct (i.e., all assumptions are met and the conditionality principle is satisfied),[3] researchers can use parameter estimates from a nonrandom or random sample to explain the relations of variables. Estimating parameters in disaggregated CFA, however, limits generalizability because

[3] Under the conditionality principle, two random sets of scores of dependent (endogenous) variables should be independent after controlling for independent (exogenous) variables (Sterba, 2009). In other words, the conditionality principle implies that a model has no omitted variables. If researchers ignore clustering, the conditionality principle is violated.

TABLE 1
The Meanings of Standardized Factor Loadings, Factor Correlations, and Regression Coefficients Across Different Types of Analyses

| Analysis | Type of Construct | Meanings |
|---|---|---|
| | | Standardized factor loadings |
| Micro level | Formative | Represent the relationship between factor and indicators controlling for clusters |
| | Reflective | Represent the relationship between factor and indicators controlling for clusters. Researchers may not use the factor structure if micro-level shared systematic variances are not of interest |
| Macro level | Formative | Reflect the degree of differences between measurement intercepts across clusters |
| | Reflective | Represent the relationship between factor and indicators at the macro level |
| Disaggregated | | Represent the relationship between factor and indicators when clustering is ignored. Should not be used because inference under the model-based approach is incorrect (the conditionality principle is not satisfied) |
| Aggregated | | Have the same meanings as the macro-level standardized factor loadings in both formative and reflective measurement models |
| Design-based | | Represent the relationship between factor and indicators in a target finite population uncontrolled for the influences of clusters |
| | | Factor correlations |
| Micro level | Formative | Represent the relationship between two factors controlling for clusters |
| | Reflective | Represent the relationship between two factors controlling for clusters. Researchers may not use the factor structure if micro-level shared systematic variances are not of interest |
| Macro level | Formative | Reflect the relationship between the cluster averages of two factors (that are targeted to measure properties of micro-level units) |
| | Reflective | Reflect the relationship between two factors (that are targeted to measure properties of macro-level units) |
| Disaggregated | | Represent the relationship between two factors when clustering is ignored. Should not be used because inference under the model-based approach is incorrect (the conditionality principle is not satisfied) |
| Aggregated | | Have the same meanings as the macro-level factor correlations in both formative and reflective measurement models |
| Design-based | | Represent the relationship between two factors in a target finite population uncontrolled for the influences of clusters |
| | | Regression coefficients |
| Micro level | Formative | Represent the directional relationship between two variables controlling for clusters |
| | Reflective | Represent the directional relationship between shared systematic differences of two variables within clusters |
| Macro level | Formative | Reflect the directional relationship between the cluster averages of two variables (that are targeted to measure properties of micro-level units) |
| | Reflective | Reflect the directional relationship between two variables (that are targeted to measure properties of macro-level units) |
| Disaggregated | | Represent the directional relationship between two variables when clustering is ignored. Should not be used because inference under the model-based approach is incorrect (the conditionality principle is not satisfied) |
| Aggregated | | Have the same meanings as the macro-level regression coefficients in both formative and reflective measurement models |
| Design-based | | Represent the directional relationship between two variables in a target finite population uncontrolled for the influences of clusters |

clustering is ignored; thus, the conditionality principle is not fully satisfied.

Table 1 summarizes the meanings of standardized factor loadings and factor correlations from different analytic methods. We also provide the meanings of regression coefficients in multilevel regression in this table highlighting the similarity between factor correlations and regression coefficients.

The different meanings of parameters are analogous to the interpretations of parameters from multiple-group CFA. Micro-level parameters from MCFA are analogous to within-group parameters. Macro-level parameters from MCFA detect the differences in mean-structure parameters (e.g., intercepts or latent means) across groups. Disaggregated CFA is analogous to a single-group analysis of the overall sample without considering groups. Design-based CFA is also analogous to a single-group analysis of overall samples but accounts for the proportions of groups in a target finite population.

Disaggregated analysis does not permit inference to an infinite population in model-based inference (because independence of observations is violated) or a finite population in design-based inference (because sampling weights are not used). This study focuses on model-based inference. That is, we investigate the degree to which micro- and macro-level parameters influence disaggregated parameters. Note that parameters from disaggregated CFA have different meanings than micro- and macro-level parameters from MCFA. Therefore, we do not refer to the differences in estimates between disaggregated CFA parameters and micro-level parameters from MCFA as “bias.”

Aggregated CFA parameters and macro-level parameters from MCFA convey the same meanings thus the differences can be interpreted as "bias." Aggregated analysis is usually inferior to MCFA because the reliability of group means is not accounted for (Lüdtke et al., 2011; Lüdtke et al., 2008). This study investigates the degree to which micro-level parameters influence bias in the estimation of macro-level parameters in aggregated CFA.

## EFFECTS OF IGNORING CLUSTERING

Clustering can be ignored in two ways: disaggregation and aggregation. The estimation of parameters and their *SE*s of disaggregated analysis is different from micro-level parameters from multilevel analysis (Chen et al., 2010; Julian, 2001; Moerbeek, 2004; Noortgate, Opdenakker, & Onghena, 2005; Opdenakker & Van Damme, 2000). The disaggregated variance of a variable is greater than the micro-level variance from multilevel analysis because the macro-level variance is not estimated but added to the disaggregated variance instead. When the micro-level and macro-level regression coefficient estimates are the same, the disaggregated regression coefficient remains the same; however, its *SE* can be lower (for an independent variable with high ICC) or higher (for an independent variable with low ICC). If ICC is larger, the difference in the *SE* is stronger. However, when the effect of an independent variable on a dependent variable differs between levels (i.e., a contextual effect exists), the disaggregated regression coefficient is a weighted average of the effects from both levels, which obscures the effect within each level (Raudenbush & Bryk, 2002).

Aggregation is another way of ignoring clustering. The average scores contain some measurement error derived from within-cluster variability yet are treated as having no measurement error at the macro level (Lüdtke et al., 2008; Marsh et al., 2012). The aggregated variance is greater than the macro-level variance because some part of the micro-level variance is not estimated but added to the macro-level variance. The difference between the aggregated and macro-level variances becomes smaller as cluster size increases. However, even when the micro level is ignored, the parameter estimates and *SE*s of regression coefficients for true macro-level predictors are not biased in a balanced design (i.e., a design in which all clusters are the same size; Moerbeek, 2004).

Regarding CFA, Julian (2001) found that the disaggregation method led to inflated chi-square statistics, increased parameters estimates, and decreased *SE*s for all parameters in a model compared with micro-level counterparts. The parameters included factor loadings, factor variances, factor covariances, and unique factor variances. The difference became larger as ICC increased. In every condition, Julian used equal total sample size (i.e., 500) but varied the balance of micro- and macro-level sample sizes. The difference was more pronounced when the total sample was allocated into a smaller number of clusters (i.e., the condition with 10 clusters and cluster size of 50 had larger bias than the condition with 50 clusters and cluster size of 10). However, the difference due to sample size was relatively small compared with that due to ICC. Finally, if the macro-level factor structure differed from the micro-level factor structure (e.g., two factors in the macro level but four factors in the micro level), the difference was larger.

Julian (2001) examined only unstandardized factor loadings and factor covariances. However, researchers usually report standardized factor loadings and factor correlations for interpreting their results. Moreover, the standardized parameters are, in fact, ratios of unstandardized parameters. For example, a factor correlation is the ratio of a factor covariance (numerator) and factor standard deviations (denominator). When both the numerator and denominator are influenced by ignoring clustering, the resulting difference in the factor correlation is difficult to predict. Therefore, in this study, we examine the differences in standardized parameters and their *SE*s that are due to ignoring clustering. Further, this study extends to the situation where ICC and macro-level communalities are different across indicators of the same factor (leading to differences in macro-level standardized factor loadings).

The amount of difference in estimates of standardized factor loadings and factor correlations can be predicted mathematically. The disaggregated standardized parameters are used to represent the factor structure in both micro-level units (e.g., Cassidy et al., 2005; Ebesutani et al., 2011; Garb et al., 2011; Hatami et al., 2010; Law et al., 2011; Merrell et al., 2011; Nelson et al., 2007; Oliver et al., 2006; Philips et al., 2006; Raspa et al., 2010; C. M. Tucker et al., 2011) and macro-level units (mostly in organizational studies; Babakus et al., 2004; C. J. Collins & Smith, 2006; Glisson & James, 2002; González-Romá et al., 2002; Han et al., 2006; Keller, 2001; Patterson et al., 2005; Riordan et al., 2005; Robert & Wasti, 2002; Salanova et al., 2005; Schaubroeck et al., 2007; Takeuchi et al., 2007; van der Vegt & Bunderson, 2005; Zohar & Tenne-Gazit, 2008). Because micro-level constructs are very unlikely to illuminate macro-level constructs (Klein & Kozlowski, 2000) and disaggregated analysis represents the relations among variables in micro-level units (uncontrolled for macro-level units), we do not focus on comparing the results from disaggregated data with macro-level parameters. Rather, the parameter estimates from the analysis of disaggregated data are compared against micro-level parameters (see Moerbeek, 2004). The parameter estimates from the analysis of aggregated data are compared against macro-level parameters because their parameters convey the same meanings. The procedures used to predict the differences of the disaggregated and aggregated parameter estimates (from the micro- and macro-level models, respectively) are described in Online Appendix A. Readers may access all of the appendices on the following website: http://quantpsy.org

We also derive closed-form formulas for the difference in the disaggregated and aggregated parameters in three special cases. The first case is when (a) ICCs of all variables are equal, (b) factor correlations are equal across levels, and (c) standardized factor loadings are equal across levels. In this case, the disaggregated standardized loadings and factor correlations are not different from the micro-level parameter estimates, and the aggregated standardized loadings and factor correlations are not different from the macro-level parameter estimates. This case is referred to as *standardized factor loading and factor correlation invariance*. The second case is when (a) the micro-level standardized factor loadings are proportional to the macro-level standardized factor loadings and (b) ICCs of all variables are equal. This case is referred to as *standardized metric (weak) cross-level invariance*. The third case is when (a) ICCs of all variables are equal and (b) unstandardized factor loadings are equal across levels. This case is referred to as *true metric (weak) cross-level invariance* (Jak et al., 2013). As shown in Online Appendix A, in both the standardized and true metric invariance cases, the disaggregated standardized loadings and factor correlations can be computed by

$$\lambda^*_{Drs} = \sqrt{\frac{\lambda^*_{Wrs}\left(\lambda^*_{Wrs}\sigma_{Wrr} + \lambda^*_{Brs}\sigma_{Brr}\right)}{\sigma_{Wrr} + \sigma_{Brr}}} \tag{14}$$

and

$$\psi^*_{Dst} = \sqrt{\frac{\psi_{Bss}}{\psi_{Bss} + \psi_{Wss}}}\,\psi^*_{Bst}\sqrt{\frac{\psi_{Btt}}{\psi_{Btt} + \psi_{Wtt}}} + \sqrt{\frac{\psi_{Wss}}{\psi_{Bss} + \psi_{Wss}}}\,\psi^*_{Wst}\sqrt{\frac{\psi_{Wtt}}{\psi_{Btt} + \psi_{Wtt}}}, \tag{15}$$

where $\lambda^*_{Wrs}$ and $\lambda^*_{Drs}$ are the micro-level and disaggregated standardized factor loadings linking indicator $r$ to factor $s$; $\sigma_{Wrr}$ and $\sigma_{Brr}$ are micro- and macro-level observed variances of indicator $r$; $\psi_{Wss}$ and $\psi_{Bss}$ are the micro- and macro-level variances of factor $s$ when using the marker variable approach of scale identification in both levels (and choosing the same marker indicator); and $\psi^*_{Wst}$, $\psi^*_{Bst}$, and $\psi^*_{Dst}$ are the micro-level, macro-level, and disaggregated factor correlation between factor $s$ and factor $t$.

As shown in Online Appendix A, under either standardized metric invariance or true metric invariance, the aggregated standardized loadings and factor correlations can be computed by

$$\lambda^*_{Ars} = \sqrt{\frac{\lambda^*_{Brs}\left(\lambda^*_{Wrs}\frac{\sigma_{Wrr}}{n} + \lambda^*_{Brs}\sigma_{Brr}\right)}{\frac{\sigma_{Wrr}}{n} + \sigma_{Brr}}} \tag{16}$$

and

$$\psi^*_{Ast} = \sqrt{\frac{\psi_{Bss}}{\psi_{Bss} + \frac{\psi_{Wss}}{n}}}\,\psi^*_{Bst}\sqrt{\frac{\psi_{Btt}}{\psi_{Btt} + \frac{\psi_{Wtt}}{n}}} + \sqrt{\frac{\frac{\psi_{Wss}}{n}}{\psi_{Bss} + \frac{\psi_{Wss}}{n}}}\,\frac{\psi^*_{Wst}}{n}\sqrt{\frac{\frac{\psi_{Wtt}}{n}}{\psi_{Btt} + \frac{\psi_{Wtt}}{n}}}, \tag{17}$$

where $\lambda^*_{Brs}$ and $\lambda^*_{Ars}$ are the macro-level and aggregated standardized factor loadings from indicator $r$ to factor $s$, and $\psi^*_{Ast}$ are the aggregated factor correlation between factor $s$ and factor $t$. Other symbols are defined as in Equations 14 and 15.

We realize that real-world applications will deviate to some degree from these three simple cases. We use these simple cases to derive some general consequences of ignoring clustering. To account for more complex cases, we also conduct two additional simulation studies. The first simulation study investigates differences due to disaggregation, replicating and extending Julian's (2001) study. The second simulation study examines the bias due to aggregation. In these simulation studies, we compare single-level CFA (which ignores the nested nature of data) against MCFA in terms of model fit, standardized parameter estimates, and their *SE*s.[4]

## HYPOTHESES TO BE TESTED

Our hypotheses are based on only the situations where (a) ICCs are equal across variables and (b) the standardized factor loadings (and thus communalities) are equal across variables within each level (but not equal across levels). From Equation 14, if $\lambda^*_{Wrs} = \lambda^*_{Brs}$, then $\lambda^*_{Drs} = \lambda^*_{Wrs}$. If $\lambda^*_{Wrs} < \lambda^*_{Brs}$, then $\lambda^*_{Drs} > \lambda^*_{Wrs}$ and vice versa. The same relations hold among $\psi^*_{Dst}$, $\psi^*_{Wst}$, and $\psi^*_{Bst}$. Therefore, the disaggregated standardized factor loadings or factor correlations are expected to deviate from the micro-level standardized factor loadings or factor correlations toward the values of the macro-level standardized factor loadings or factor correlations, respectively. The degree of this deviation is expected to be larger if ICC is higher. This statement implies that if the micro- and macro-level standardized loadings (or factor

[4] Based on a reviewer's suggestion, we also analytically prove the consequences of ignoring clustering on scale reliability. We consider coefficient alpha ($\alpha$), coefficient omega ($\omega$), and maximal reliability ($H$) in this proof. In Online Appendix A, we show that disaggregated and aggregated $\alpha$, $\omega$, and $H$ lay between micro- and macro-level $\alpha$, $\omega$, and $H$ assuming true or standardized metric invariance, equal micro- and macro-level standardized loadings across items, and equal micro- and macro-level observed variances across items. Geldhof, Preacher, and Zyphur (2014) discuss some consequences of ignoring the nested data structure on scale reliability and also discuss consequences for other types of reliability estimates.

correlations) are equal (Case 1), the disaggregated standardized parameter estimates are not deviated.

From Equation 16, if $\lambda^*_{Wrs} = \lambda^*_{Brs}$, then $\lambda^*_{Ars} = \lambda^*_{Brs}$. If $\lambda^*_{Wrs} < \lambda^*_{Brs}$, then $\lambda^*_{Ars} < \lambda^*_{Brs}$ and vice versa. The same relations hold among $\psi^*_{Ast}$, $\psi^*_{Wst}$, and $\psi^*_{Bst}$. Therefore, the aggregated standardized factor loadings or factor correlations are anticipated to deviate from the macro-level standardized factor loadings or factor correlations toward the values of the micro-level standardized factor loadings or factor correlations, respectively. The deviation will be larger if ICC is lower or the cluster size is smaller. This statement implies that if the micro- and macro-level standardized loadings (or factor correlations) are equal (Case 1), the aggregated standardized parameter estimates are not biased.

## SIMULATION STUDY 1: DISAGGREGATION

This simulation study examines the case in which clustering is ignored and the data are analyzed by single-level CFA directly. We examine some conditions similar to those investigated by Julian (2001) for the sake of comparability.

The model used in this simulation study has six observed variables. Both micro and macro levels have the same factor structure: two factors with three indicators each. All micro-level standardized factor loadings are fixed to .7. As a consequence, all micro-level uniquenesses are .51. The micro-level factor correlation is fixed to .5. The macro-level factor correlation and standardized loadings are varied across conditions.

### Design Conditions

*Sample size.* There are four sample size conditions. Two conditions, with a total sample size of 500, are 100/5 and 10/50. The first and the second numbers indicate the number of clusters and cluster size, respectively. The other two conditions, with a total sample size of 8,000, are 400/20 and 40/200. We add these much larger total sample size combinations in order to increase the convergence rate in MCFA so the results from the disaggregated single-level CFA and MCFA can be compared. In sum, there are four sample size conditions: 100/5, 10/50, 400/20, and 40/200.

*Intraclass correlation.* There are five ICC conditions in this study: .05, .15, .25, .50, and .75.

*Distribution of intraclass correlation across items within a factor.* ICCs of all indicators within the same factor can have the same or different values. To specify values for the unequal ICC conditions, first, the distances from the ICC and both 0 and 1 are calculated (e.g., .15 − 0 = .15 and 1 − .15 = .85). Next, the shortest distance is chosen (e.g., choose .15). Then, the shortest distance is divided by 2 (e.g., .15/2 = .075), which is referred to as the *margin*. In each factor, three indicators have an ICC value equal to the average ICC, the average ICC minus the margin, and the average ICC plus the margin (e.g., .075, .15, and .225).

*Macro-level communalities.* There are three conditions for the macro-level communality: low, medium, and high. For the medium condition, we set the macro-level communality equal to the micro-level communality: .49 (i.e., standardized factor loadings of .7). The low and high communalities are different from the medium communality condition by .25, that is, .24 and .74, respectively (i.e., standardized factor loadings of .49 and .86).[5]

*Distribution of macro-level communalities across items within a factor.* Similar to ICC, we also specify communalities to be equal or unequal across indicators of a factor. For unequal conditions, the communality of one indicator is less than the average communality by .20, the communality of the second indicator is equal to the average communality, and the communality of the last indicator is greater than the average communality by .20. For example, the communalities in the low, unequal condition are .04, .24, and .44.

When unequal ICCs were combined with unequal communalities, we prevented an extreme match (e.g., high ICC and high communality) on the same indicator, such that the first indicator has the largest communality with the average ICC, the second indicator has the average communality with the lowest ICC, and the third indicator has the lowest communality with the highest ICC.

*Macro-level factor correlation.* The macro-level factor correlation is .2, .5, or .8. Note that the medium level is the same as the value of the micro-level factor correlation.

Therefore, there are 4 × 5 × 2 × 3 × 2 × 3 = 720 conditions in total. Each condition is analyzed in 1,000 replications.

### Data Generation and Analysis Methods

Data were generated by an MCFA model with the parameter values described earlier. Three analysis methods were used: (a) full MCFA, (b) MCFA with a saturated macro level, and (c) single-level CFA that ignores clustering. For the disaggregated single-level CFA, we use the reparameterization

[5] In most MCFA applications with formative constructs (similar to Simulation Study 1), the macro-level communalities are expected to be higher than micro-level communalities. Because the measurement error variances consist of systematic and random error variances, the amount of random error variances in the macro level is usually lower because random errors are averaged out by multiple observations (i.e., the central limit theorem), especially with a large cluster size. The amount of systematic error variance, however, could be different at each level. Although macro-level communalities are typically higher than micro-level communalities, in order to account for all possibilities, we include all scenarios such that communalities could be lower or higher at the macro level.

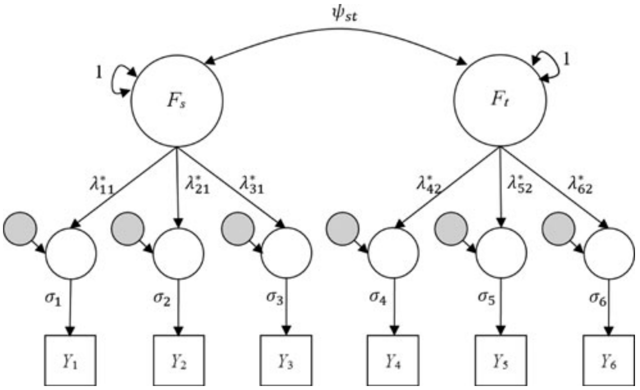

FIGURE 1 The single-level confirmatory factor analysis (CFA) model. The grey circles represent unique factors, variances of which are constrained equal to $1 - \lambda^{*2}_{rs}$. $\sigma_r$ is the standard deviation of indicator $r$.

method from Equations 2–4 to obtain standardized coefficients. Figure 1 shows the single-level CFA analysis model. For the full MCFA model, we use the reparameterization method from Equations 5 and 8–13. Figure 2 shows the full MCFA model. Finally, the partially saturated MCFA model is used to fit a CFA model to the micro level using the reparameterization method (Equations 2–4) and freely estimate all covariances among variables in the macro level (a saturated model). Figure 3 depicts the partially saturated model used in this simulation. We used Mplus 7 (L. K. Muthén & Muthén, 1998–2013) for both data generation and data analysis. Online Appendix B shows example Mplus code.

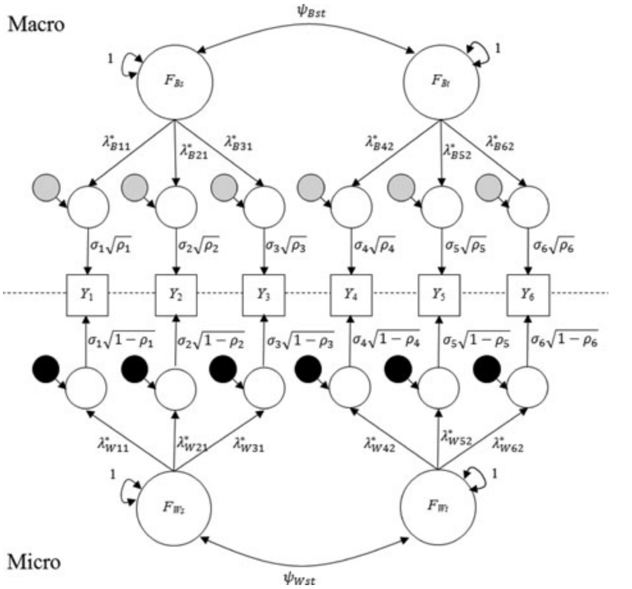

FIGURE 2 The multilevel CFA model. The grey circles represent macro-level unique factors, the variances of which are constrained equal to $1 - \lambda^{*2}_{Brs}$. The black circles represent micro-level unique factors, the variances of which are constrained equal to $1 - \lambda^{*2}_{Wrs}$. $\sigma_r$ is the standard deviation of indicator $r$. $\rho_r$ is the intraclass correlation of indicator $r$.

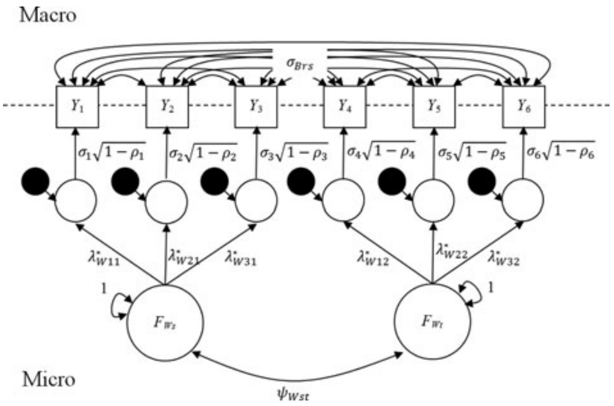

FIGURE 3 The micro-level CFA model with a saturated macro level. The black circles represent micro-level unique factors, the variances of which are constrained equal to $1 - \lambda^{*2}_{Wrs}$. The macro-level CFA model with a saturated micro level is similar to this diagram but has the CFA model at the macro level and estimates all possible covariances at the micro level. $\sigma_r$ is the standard deviation of indicator $r$. $\rho_r$ is the intraclass correlation of indicator $r$.

## The Evaluation of Different Analysis Methods

All three analysis models are investigated in terms of (a) model fit, (b) differences in standardized loadings and their *SE*s, and (c) differences in factor correlations and their *SE*s. Because the data are generated from the full MCFA model, the model fit values for the full MCFA model and the partially saturated MCFA model should indicate good fit. The model fit estimates for disaggregated CFA models should indicate poor fit because clustering is ignored (Julian, 2001; Stapleton, 2006). We use two indices to evaluate model fit: the chi-square goodness-of-fit statistic and RMSEA. First, the rejection rate for the chi-square test based on an alpha level of .05 is used to evaluate each analysis model. The rejection rate from the full MCFA model and partially saturated MCFA model should be approximately equal to the nominal alpha of .05, whereas the rejection rate from the disaggregated single-level CFA should be greater than .05. Second, we transformed the chi-square statistic to RMSEA ($\varepsilon$) by

$$\varepsilon = \sqrt{Max\left(\frac{\chi^2 - df}{df \cdot q}, 0\right)} \tag{18}$$

where $\chi^2$ is the observed chi-square value from each replication, $df$ is the degrees of freedom from each analysis model, and $q$ is the appropriate sample size correction for each analysis model. The sample size corrections for the full MCFA model, the partially saturated MCFA model, and the disaggregated single-level CFA model are $N$ – 1, $N$, and $N$ – 1, respectively, where $N$ is the total sample size (see Equations 8 and 19 in Ryu & West, 2009). The average RMSEA is calculated in each condition. The average RMSEA for the full MCFA and partially saturated MCFA models should be close to 0 to indicate no misfit and that for the single-level CFA model should be systematically greater than 0.

To examine differences in parameter estimates and *SE*s, we used the results from only the disaggregated CFA and the full MCFA because the (micro-level) parameter estimates and *SE*s from the full MCFA and the partially-saturated MCFA are not different theoretically. The micro-level standardized loadings and factor correlations are averaged across replications. The MCFA model should have average micro-level standardized parameter estimates close to the population parameter values (standardized loadings = .7 and factor correlation = .5), whereas the disaggregated CFA should provide average parameter estimates that follow the patterns described in the aforementioned hypotheses (i.e., positive differences in standardized factor loadings when the macro-level standardized factor loadings are higher). The average *SE*s of standardized loadings and factor correlations in each condition are also calculated.

There are several ways to define differences (or biases in appropriate circumstances) in parameter estimates and their *SE*s, such as relative difference (Hoogland & Boomsma, 1998) or standardized difference (L. M. Collins, Schafer, & Kam, 2001). Because factor loadings and factor covariances often are considered more meaningful when they are in standardized form (i.e., as standardized loadings and factor correlations), we use absolute bias:

$$\text{Absolute Difference}\,(\theta) = \bar{\theta}_{est} - \theta_{true}, \tag{19}$$

where $\bar{\theta}_{est}$ is the average of parameter estimates in each condition and $\theta_{true}$ is the population value of that parameter. If the difference in standardized factor loadings or factor correlations is less than 0.05, the difference is arguably acceptable because this amount of bias rarely changes the interpretation of factor analysis results (Widaman, 1993).

For the *SE*s, an absolute difference may not be directly interpretable. Therefore, we use the relative difference of the estimated *SE* (Hoogland & Boomsma, 1998):

$$\text{Relative Difference}\,(SE_{\theta}) = \frac{\overline{SE}_{\theta_{est}} - \sigma_{\theta_{est}}}{\sigma_{\theta_{est}}}, \tag{20}$$

where $\overline{SE}_{\theta_{est}}$ is the average *SE* of a parameter estimate and $\sigma_{\theta_{est}}$ is the standard deviation of the parameter estimate across replications. We use the standard deviations of the parameter estimates from the full MCFA model and consider a value of 0.10 to be an acceptable difference for the *SE*s (Hoogland & Boomsma, 1998).

To determine which of the design conditions contributed to the rejection rate, RMSEA, and the differences in parameter estimates and *SE*s, we used ANOVA where design conditions are used as fixed factors. Similar to Lüdtke et al. (2011) and Lüdtke et al. (2008), ANOVA was conducted at the cell mean level where the average of a desired result (e.g., rejection rate) of each cell is used as a dependent variable—one observation for each cell so the highest level (seven-way) interaction could not be separated from the error. We used the proportion of variance explained ($\eta^2$) as a measure of effect size for each of the main effects and interaction effects. We considered the factors with $\eta^2$ greater than .01 (Lüdtke et al., 2011) and .05 (Geldhof, Preacher, & Zyphur, 2014) as target factors. We found that the factors with $\eta^2$ between .01 and .05 did not provide detectable differences in our graphical illustrations shown later. Therefore, we considered the factors with $\eta^2$ greater than .05 only.

## Results

*Overall summary.* Table 2 shows $\eta^2$ for all main and interaction effects of each design condition on target dependent variables: rejection rate from the chi-square test, RMSEA, differences in standardized factor loadings, and differences in factor correlations. All main and interaction effects involving equal/unequal ICC and equal/unequal macro-level communalities have $\eta^2$ less than .05. Therefore, we do not show these conditions in Table 2 and provide them in Online Appendix C.

*Convergence rate.* We examined the convergence rate by checking the output of each replication for inadmissable estimates (e.g., negative variances). The full results of the proportions of converged replications in the disaggregated CFA, full MCFA, and partially saturated MCFA models across different sample sizes and ICCs are shown in Online Appendix C. The convergence rate of the full MCFA model is low in conditions with few clusters (10), small total sample size (500), or low ICC (.05) and unequal ICC across items. Most nonconvergent replications of the partially saturated model are in conditions with small total sample size, especially with low and unequal ICC or few clusters. On the other hand, the convergence rate for the single-level CFA is low when ICC is large, especially when the total sample size is small with few clusters (10).

*The rejection rate for the chi-square test.* We summarize the simulation results with regard to the analysis methods, in which the main effect has $\eta^2$ greater than .05. The single-level CFA model is rejected in almost all replications (97%). The full MCFA model and the partially saturated MCFA model are rejected at a rate close to the nominal level (8.6% and 4.6%, respectively).

*RMSEA.* We summarize the simulation results with regard to the analysis methods and ICC only—the main or interaction effects involving these design conditions had $\eta^2$ greater than .05. The results for the average RMSEA across conditions are shown in Figure 4. The RMSEA of the full MCFA and partially saturated MCFA models, on average, are close to 0, indicating good fit. The RMSEA for the disaggregated CFA indicated bad fit, especially in the large ICC condition.

TABLE 2
Eta-Squared Values for Analysis of Variance Table of the Simulation Conditions for the Results in Simulation Study 1

| Effects | RR | RMSEA | Loading | RD *SE* Loading | CorW | RD *SE* CorW |
|---|---|---|---|---|---|---|
| $N$ | .005 | .005 | .002 | .011 | .004 | .029 |
| ICC | .005 | **.122** | .004 | .006 | .006 | .009 |
| $h^2$ | .000 | .002 | **.302** | **.270** | .000 | **.083** |
| CorB | .000 | .000 | .000 | .002 | **.295** | .025 |
| Method | **.945** | **.619** | .010 | .001 | .004 | .005 |
| $N$ : ICC | .006 | .000 | .001 | .025 | .013 | **.078** |
| $N$ : $h^2$ | .000 | .000 | .001 | .002 | .002 | .007 |
| ICC : $h^2$ | .000 | .000 | **.171** | **.131** | .004 | **.067** |
| $N$ : CorB | .000 | .000 | .000 | .002 | .002 | .003 |
| ICC : CorB | .000 | .000 | .000 | .004 | **.140** | .016 |
| $h^2$ : CorB | .000 | .000 | .000 | .001 | .020 | .001 |
| $N$ : Method | .008 | .014 | .000 | .019 | .003 | .013 |
| ICC : Method | .003 | **.226** | .006 | .009 | .003 | .002 |
| $h^2$ : Method | .000 | .005 | **.301** | **.237** | .000 | **.061** |
| CorB : Method | .000 | .000 | .000 | .005 | **.264** | .020 |
| $N$ : ICC : $h^2$ | .001 | .000 | .001 | .006 | .010 | **.098** |
| $N$ : ICC : CorB | .000 | .000 | .000 | .003 | .004 | .017 |
| $N$ : $h^2$ : CorB | .000 | .000 | .000 | .002 | .001 | .018 |
| ICC : $h^2$ : CorB | .000 | .000 | .000 | .004 | .008 | .009 |
| $N$ : ICC : Method | .013 | .001 | .001 | .006 | .006 | .004 |
| $N$ : $h^2$ : Method | .001 | .000 | .001 | .002 | .000 | .000 |
| ICC : $h^2$ : Method | .001 | .001 | **.168** | **.143** | .001 | .042 |
| $N$ : CorB : Method | .000 | .000 | .000 | .000 | .003 | .000 |
| ICC : CorB : Method | .000 | .000 | .000 | .004 | **.145** | .010 |
| $h^2$ : CorB : Method | .000 | .000 | .000 | .000 | .021 | .002 |

*Note.* All four-way or higher order interactions are not shown because $\eta^2 < .01$. All main effects and interaction effects involving unequal ICC and unequal $h^2$ are not shown because $\eta^2 < .05$. Bold values indicate $\eta^2 > .05$. $N$ = sample size (100/5, 10/50, 400/20, and 40/200, where the first value is the number of clusters and the second value is cluster size). ICC = average intraclass correlations across indicators (.05, .15, .25, .50, and .75); $h^2$ = average macro-level communality (low, medium, and high); CorB = average macro-level factor correlation (.2, .5, and .8); Method = the method of analysis (full multilevel structural equation modeling (MSEM), saturated-macro-level MSEM, and disaggregated structural equation modeling); RD = relative difference; RR = rejection rate based on $\chi^2$ test; RMSEA = root mean square error of approximation; *SE* = standard error; CorW = the estimated micro-level correlation. See the full table in Online Appendix C, http://quantpsy.org.

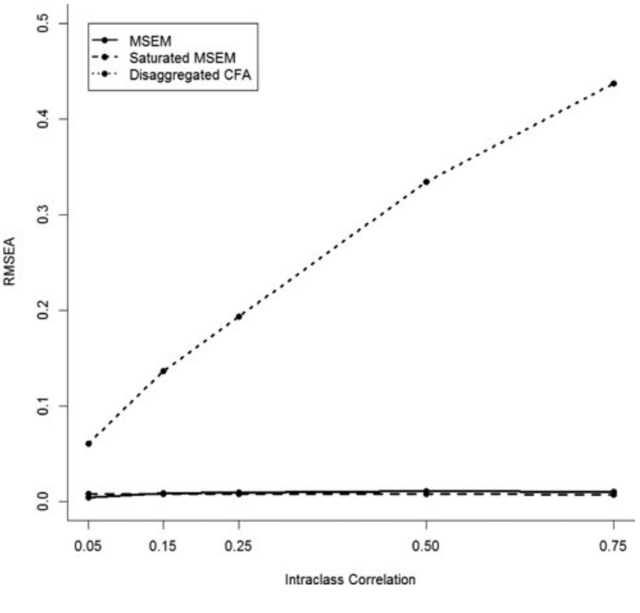

FIGURE 4 The average root mean square error of approximation (RMSEA) of Simulation Study 1. MSEM = Multilevel Structural Equation Modeling; CFA = Confirmatory Factor Analysis.

*Micro-level standardized factor loadings.* As mentioned earlier, only the full MCFA model and the disaggregated single-level CFA model are used. We summarize the simulation results with regard to the analysis methods, ICC, and macro-level communalities only ($\eta^2$s > .05).

The average standardized loadings of all conditions are shown in Figure 5. The results support our hypothesis—the disaggregated standardized loadings are not biased when the micro- and macro-level standardized loadings are equal. When the standardized loadings are not equal across levels, the disaggregated standardized loadings are different from micro-level standardized loadings, especially when ICC is high. The disaggregation method generally results in increases of parameter values when the macro-level communality is high (high macro-level standardized loadings) and decreases when the macro-level communality is low (low macro-level standardized loadings). The absolute difference is within ± .05 when ICC is less than .25 or when the micro- and macro-level standardized loadings are equal.

The simulation results of the relative differences in *SE*s of standardized loadings are summarized for the analysis

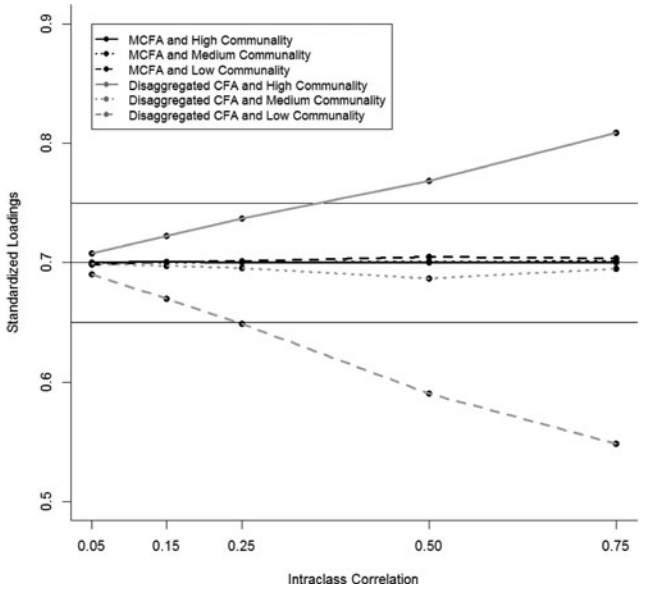

FIGURE 5 The average standardized factor loadings in each condition. The solid horizontal lines denote absolute differences of –.05, 0, and .05 to represent the acceptable range for the average standardized factor loadings. CFA = Confirmatory Factor Analysis; MCFA = Multilevel CFA.

methods, ICC, and macro-level communalities ($\eta^2$s > .05). The relative difference in the *SE* is shown in Figure 6. When the macro-level communality is equal to micro-level communality, the relative difference in the *SE* is not greater than .10 in most conditions. Almost all of the *SE*s in the conditions with low or high macro-level communalities are biased.

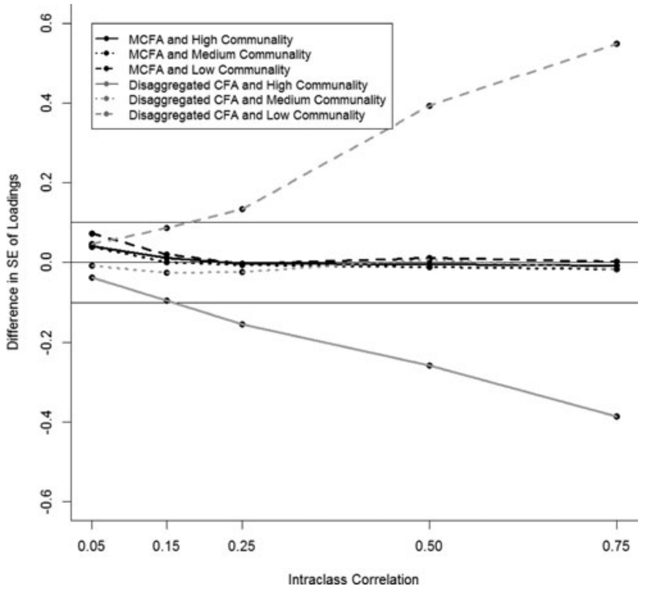

FIGURE 6 The relative differences in standard errors (*SE*s) of standardized factor loadings in each condition. The solid horizontal lines denote relative differences of –0.1, 0, and 0.1 to represent the acceptable range for the relative difference in the *SE*s. CFA = Confirmatory Factor Analysis; MCFA = Multilevel CFA.

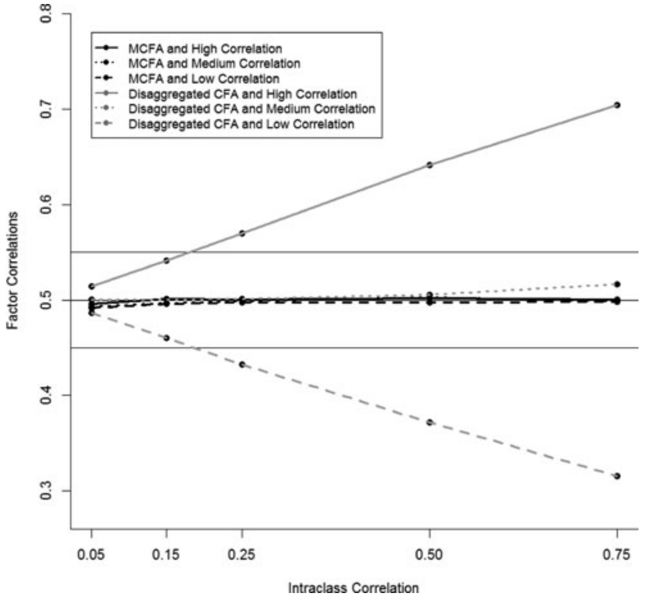

FIGURE 7 The average factor correlation in each condition. The solid horizontal lines denote absolute differences of –.05, 0, and .05 to represent the acceptable range for the average factor correlation. CFA = Confirmatory Factor Analysis; MCFA = Multilevel CFA.

When the macro-level communality is low (low macro-level standardized loadings), the *SE*s tend to be higher, and vice versa. The difference in the *SE* is larger when ICC is higher. In sum, the *SE*s of standardized factor loadings are negatively related to their parameter estimates. When the standardized factor loadings are increased, their *SE*s are decreased. When standardized factor loadings are decreased, their *SE*s are increased.

*Micro-level factor correlation.* The analysis methods and macro-level communality affect the estimates ($\eta^2$s > .05). Therefore, we summarize the simulation results only for ICC and macro-level factor correlations. The average factor correlations are shown in Figure 7. The results for the micro-level factor correlation show patterns similar to those of the micro-level standardized factor loadings, supporting our hypothesis. The absolute bias is within ± .05 when ICC is less than .25 or when the micro- and macro-level factor correlations are equal.

The macro-level communality, rather than macro-level factor correlations, affects the relative differences in *SE* of the estimates so the simulation results are summarized for the analysis methods, sample size, ICC, and macro-level communalities ($\eta^2$s > .05). The relative difference in the *SE*s is shown in Figure 8. The results are similar to the relative differences in *SE* of the standardized loadings. When the macro-level communality is low (low macro-level standardized loadings), the *SE*s tend to be higher, and vice versa. The pattern is clearer for the total sample size of 8,000. In most conditions with the total sample size of 8,000, the magnitude

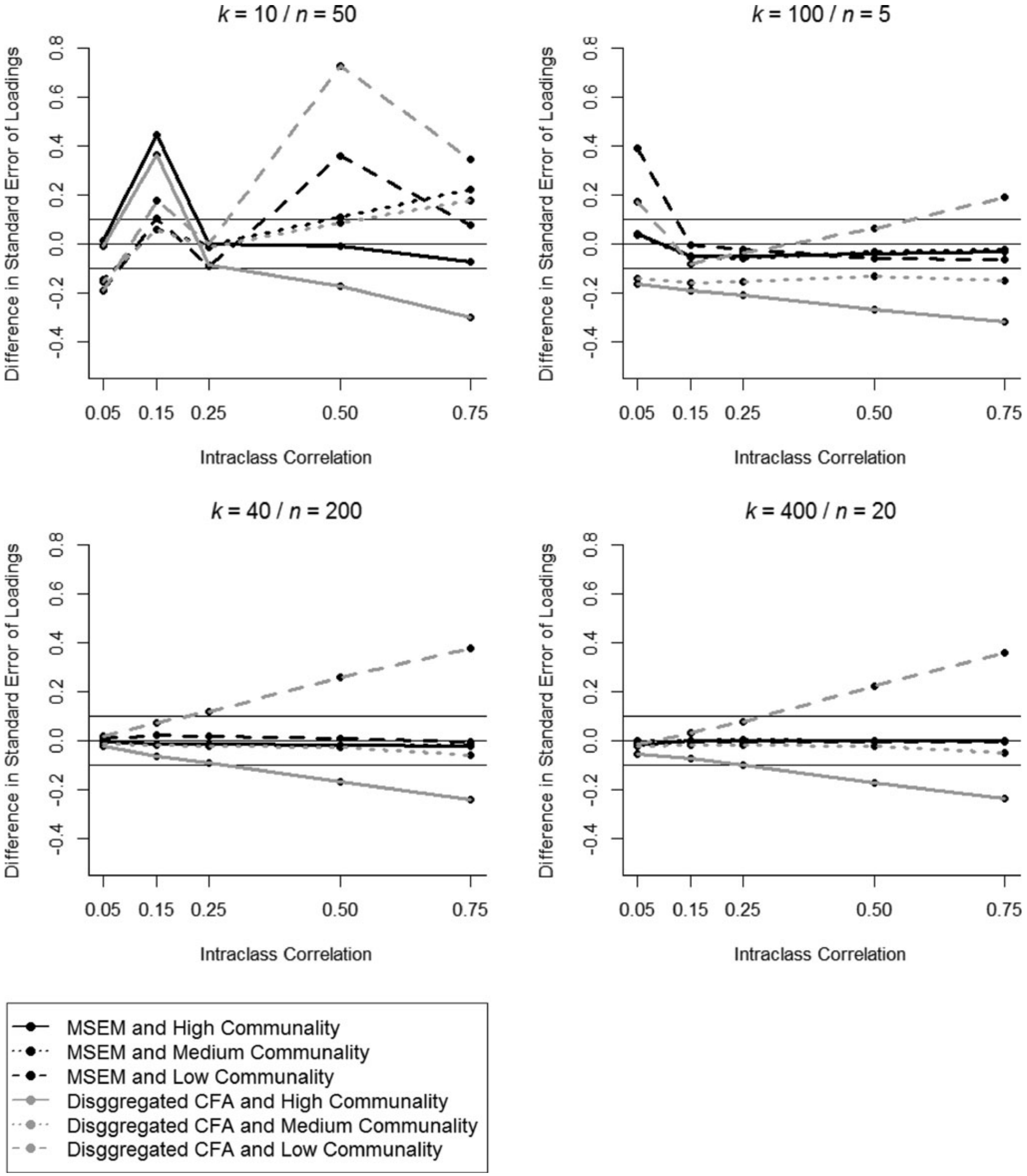

FIGURE 8 The relative difference in standard errors of factor correlation in each condition. *k* is the number of clusters and *n* is the cluster size. The solid horizontal lines in each plot denote relative differences of –0.1, 0, and 0.1 to represent the acceptable range for the relative difference in the standard errors. MSEM = Multilevel Structural Equation Modeling; CFA = Confirmatory Factor Analysis.

of relative difference in the *SE* is less than .10 when ICC is less than .25.

## SIMULATION STUDY 2: AGGREGATION

In this simulation study, we use the same data structure as in the first simulation study—six variables within two factors in both the micro and macro levels—but the micro level is ignored instead. All population macro-level standardized factor loadings are set to 0.7. As a consequence, the population uniqueness of each variable is 0.51. The population macro-level factor correlation is set to 0.5. The micro-level factor correlation and standardized loadings are varied across conditions. Note that we use "bias" in describing the difference between macro-level and aggregated parameters.

### Design Conditions

*Sample size.* Different sample size conditions are used in this study because the sample size in an aggregated analysis (the resulting sample size when ignoring nesting) is the number of clusters, not the total sample size. The number of clusters is 50 or 200. The cluster size is 10 or 40. Thus, there are four conditions for sample size: 50/10, 50/40, 200/10, and 200/40.

*Intraclass correlation.* We examine five ICC conditions: .05, .25, .50, .75, and .95. The ICC of .95 is used rather than .15 in the first simulation because we would like to examine the condition when the micro-level variances are very low, where the aggregation method should have the least impact compared with other ICC conditions.

*Distribution of intraclass correlation across items within a factor.* ICC can be equal or unequal across items within a factor. The values of ICC for each item are determined by the method described in the first simulation study.

*Micro-level communalities.* Similar to Simulation Study 1, the micro-level communalities are specified as low (.24), medium (.49), or high (.74).

*Distribution of micro-level communalities across items within a factor.* Micro-level communalities can be equal or unequal across items within a factor. The values of micro-level communalities for each item are determined by the method described in the first simulation study.

*Micro-level correlation.* Similar to Simulation Study 1, the micro-level correlation is low (.2), medium (.5), or high (.8).

There are $4 \times 5 \times 2 \times 3 \times 2 \times 3 = 720$ conditions in total. Each condition is analyzed in 1,000 replications.

### Data Generation and Analysis Methods

Data were generated by the full MCFA model (Figure 2). Three analysis methods were used: (a) full MCFA, (b) MCFA with a saturated micro level, and (c) single-level CFA that ignores the micro level. The details are similar to those in the first simulation study.

### The Evaluation of Different Analysis Methods

All three analysis models are investigated in terms of (a) model fit (rejection rate by the chi-square test and RMSEA), (b) the bias in standardized loadings and their *SE*s, and (c) the bias in factor correlations and their *SE*s. The full MCFA and partially saturated MCFA models should indicate good fit (rejection rate approximately equals .05; low RMSEA). The aggregated single-level CFA model should still indicate good fit because the analysis does not violate the independence of observations assumption. The MCFA parameter estimates should be unbiased. The parameter estimates from the aggregated single-level CFA should be in the same direction as our hypotheses suggested (e.g., the aggregated parameter estimates are weighted averages of the micro- and macro-level parameter estimates). We also used $\eta^2$ in ANOVA to determine the contributions of each simulation condition.

### Results

*Overall summary.* As in the first simulation study, Table 3 shows $\eta^2$ for all main and interaction effects of each condition on target dependent variables. We do not show all main and interaction effects involving equal/unequal ICC and equal/unequal macro-level communalities ($\eta^2$s $\leq .05$) in Table 3; these are provided in Online Appendix D.

*Convergence rate.* The convergence rate for both the MCFA models is low when the cluster size is small (10), when ICC is low (e.g., .05), or when ICC is not equal across items (see Online Appendix D). The convergence rates for the aggregated single-level model are generally good.

*The rejection rate from the chi-square test.* As in the first simulation study, we summarize the simulation results with regard to the analysis methods, sample size, and ICC only ($\eta^2$s $> .05$). The results for the rejection rate across conditions are shown in Figure 9. All rejection rates are between .03 and .07 regardless of the analysis methods. As an exception, the full and saturated MCFA models had rejection rates close to 0 if the number of clusters is 200, cluster size is 10, and ICC is .05 (the convergence rate of this condition is only 10%). The rejection rates are slightly greater than the

TABLE 3
Eta-Squared Values for Analysis of Variance Table of the Simulation Conditions for the Results in Simulation Study 2

| Effects | RR | RMSEA | Loading | RB *SE* Loading | CorB | RB *SE* CorB |
|---|---|---|---|---|---|---|
| $N$ | **.215** | **.283** | .007 | .038 | **.062** | **.055** |
| ICC | **.149** | .010 | .008 | .007 | **.088** | .018 |
| $h^2$ | .002 | .000 | **.144** | .013 | .001 | .010 |
| CorW | .001 | .000 | .000 | .000 | **.093** | .002 |
| Method | **.087** | **.559** | .023 | **.179** | .023 | **.144** |
| $N$ : ICC | **.150** | .013 | .033 | **.145** | **.081** | **.199** |
| $N$ : $h^2$ | .003 | .000 | .025 | .006 | .000 | .004 |
| ICC : $h^2$ | .010 | .001 | **.232** | .024 | .002 | .013 |
| $N$ : CorW | .002 | .000 | .000 | .000 | .016 | .002 |
| ICC : CorW | .001 | .000 | .000 | .001 | **.143** | .005 |
| $h^2$ : CorW | .000 | .000 | .000 | .000 | .009 | .000 |
| $N$ : Method | .037 | **.096** | .014 | **.059** | .028 | .049 |
| ICC : Method | **.127** | .015 | .044 | **.347** | **.070** | **.293** |
| $h^2$ : Method | .007 | .000 | **.144** | .001 | .000 | .006 |
| CorW : Method | .001 | .000 | .000 | .000 | **.094** | .001 |
| $N$ : ICC : $h^2$ | .003 | .001 | .029 | .013 | .001 | .008 |
| $N$ : ICC : CorW | .001 | .000 | .000 | .001 | .017 | .004 |
| $N$ : $h^2$ : CorW | .003 | .000 | .000 | .000 | .001 | .000 |
| ICC : $h^2$ : CorW | .000 | .000 | .000 | .001 | .009 | .003 |
| $N$ : ICC : Method | .049 | .016 | .032 | **.120** | **.071** | **.111** |
| $N$ : $h^2$ : Method | .007 | .000 | .022 | .001 | .000 | .000 |
| ICC : $h^2$ : Method | .005 | .000 | **.213** | .001 | .000 | .006 |
| $N$ : CorW : Method | .004 | .000 | .000 | .000 | .015 | .001 |
| ICC : CorW : Method | .003 | .000 | .000 | .000 | **.132** | .003 |
| $h^2$ : CorW : Method | .001 | .000 | .000 | .000 | .008 | .000 |

*Note.* All four-way or higher order interactions are not shown because $\eta^2 < .05$. All main effects and interaction effects involving unequal ICC and unequal $h^2$ are not shown because $\eta^2 < .05$. Bold values indicate $\eta^2 > .05$. $N$ = sample size (50/10, 50/40, 200/10, and 200/40, where the first value is the number of clusters and the second value is cluster size). ICC = average intraclass correlations across indicators (.05, .25, .50, .75, and .95); $h^2$ = average micro-level communality (low, medium, and high); CorW = average micro-level factor correlation (.2, .5, and .8); Method = the method of analysis (full multilevel structural equation modeling (MSEM), saturated-micro-level MSEM, and aggregated structural equation modeling); RB = relative bias; RR = rejection rate based on chi-square test; RMSEA = root mean square error of approximation; *SE* = standard error; CorB = the estimated macro-level correlation. See the full table in Online Appendix D, http://quantpsy.org

nominal level of .05 in most conditions for the aggregated single-level CFA.

*RMSEA.* We summarize the simulation results in terms of the analysis methods and sample size ($\eta^2$s > .05). The results for the average RMSEA across conditions are shown in Figure 10. The RMSEA indicates better fit in the full MCFA model than in the other two analysis models. The average RMSEA values of the partially saturated MCFA model provided slightly better fit than the aggregated single-level CFA model. When the number of clusters increases, the RMSEA indicates better fit.

*Macro-level standardized factor loadings.* We summarize the simulation results of the estimates of standardized loadings in terms of the analysis methods, sample size, ICC, and micro-level communalities only ($\eta^2$s > .05). The average standardized loadings of all conditions are shown in Figure 11. The results support our hypothesis—the aggregated standardized loadings are not biased when the micro- and macro-level standardized loadings are equal. When the standardized loadings are not equal across levels, the aggregated standardized loadings are biased, especially when ICC is low and cluster size is low. Aggregation results in overestimated parameters when the micro-level communality is high (high micro-level standardized loadings) and underestimated parameters when the micro-level communality is low (low micro-level standardized loading). The absolute bias is within ± .05 when ICC is greater than .25 or the micro- and macro-level standardized loadings are equal.

We summarize the simulation results of the relative biases of the *SE* of standardized loadings in terms of the analysis methods, ICC, and sample size only ($\eta^2$s > .05). The relative bias of the *SE*s is shown in Figure 12. In general, the *SE*s from the aggregated single-level CFA are underestimated, especially when ICC is low and cluster size is low. The *SE*s from the MCFA are overestimated when ICC is low and cluster size is low. The relative bias is within ± .1 when ICC is greater than .25 and .5 for the MCFA and the aggregated single-level CFA, respectively.

*Macro-level factor correlation.* Similar to the first simulation study, we summarize the simulation results of the parameter estimates only for the analysis methods, sample

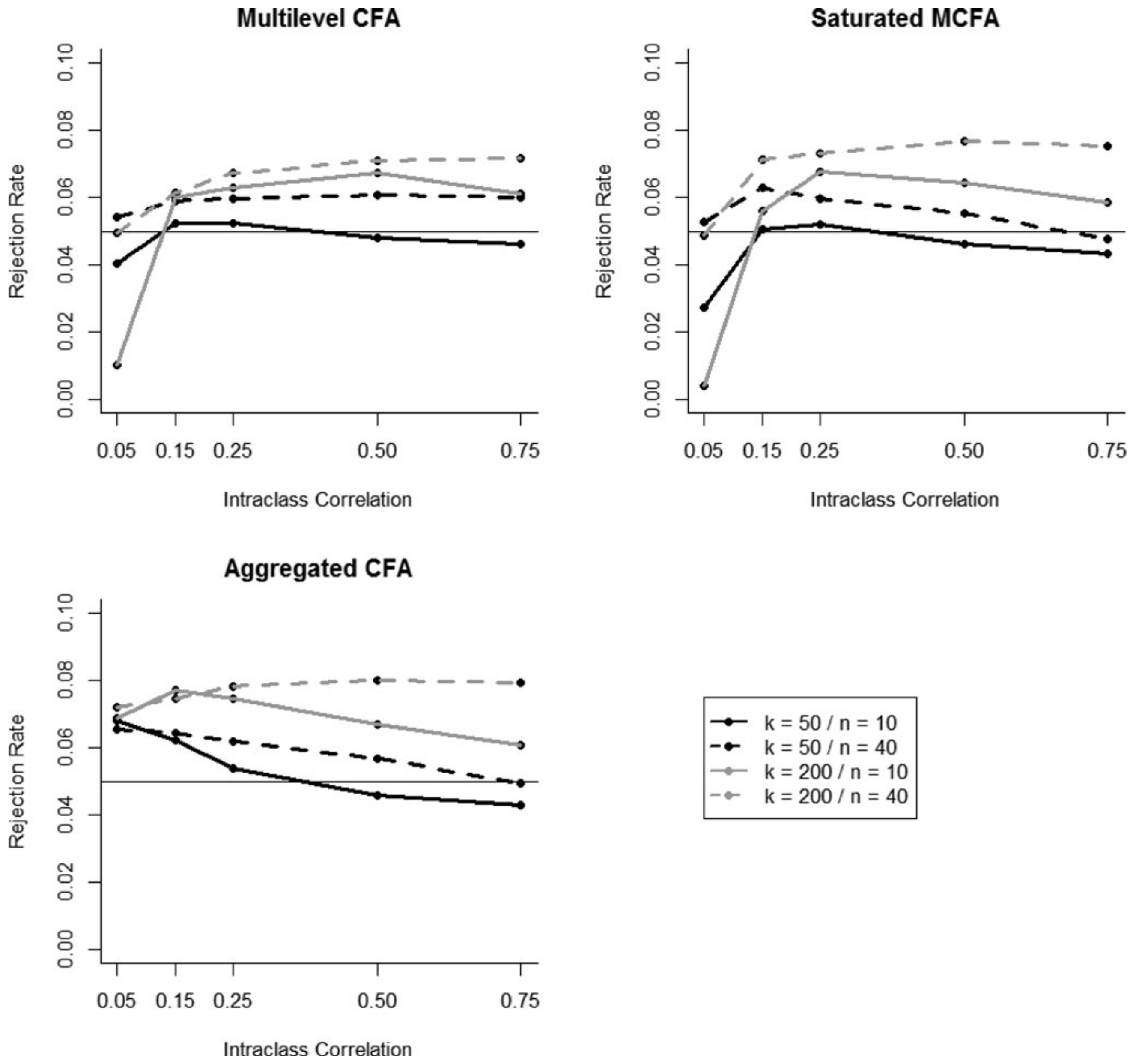

FIGURE 9 Rejection rate based on the chi-square statistics from Simulation Study 2. *k* is the number of clusters and *n* is the cluster size. The solid horizontal lines denote a rejection rate of .05, the nominal alpha. CFA = Confirmatory Factor Analysis; MCFA = Multilevel CFA.

size, ICC, and micro-level factor correlation ($\eta^2$s > .05). The average factor correlations of all conditions are shown in Figure 13. The results of the factor correlation are similar to those observed for the macro-level standardized factor loadings and support our hypothesis—the aggregated factor correlations are not biased when the micro- and macro-level factor correlations are equal, except when ICC is low (.05). When the factor correlations are not equal across levels, the aggregated factor correlations are biased, especially when ICC is low (.05) and cluster size is low (10). Aggregation results in overestimated parameters when the micro-level factor correlation is high and underestimated parameters when the micro-level factor correlation is low. Bias is lower when the cluster size is higher. The absolute bias of the aggregated factor correlation is within ±.1 when ICC is greater than .25.

Similar to the standardized loadings, we summarize the simulation results of the relative bias of *SE* of factor correlation only for the analysis methods, sample size, and ICC. The relative bias of the *SE*s is shown in Figure 14. The pattern is similar to the relative bias of *SE* of standardized loadings. The relative bias is within ± .1 when ICC is greater than .25 and .5 for the MCFA and the aggregated single-level CFA, respectively.

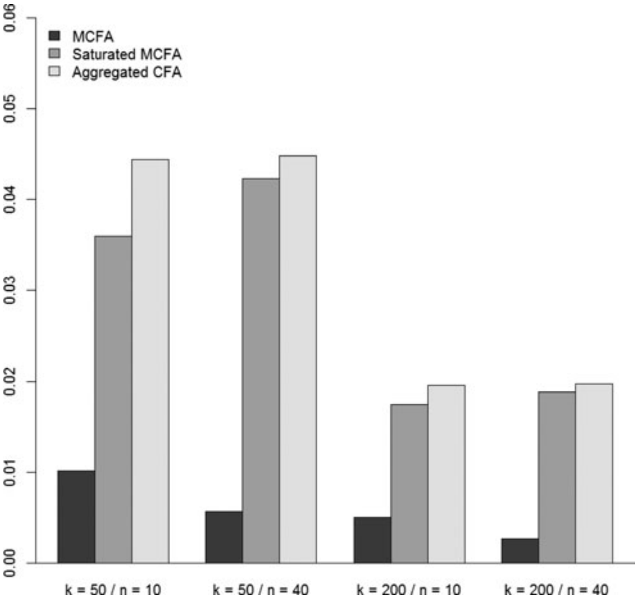

FIGURE 10 The average root mean square error of approximation from Simulation Study 2. *k* is the number of clusters and *n* is the cluster size. CFA = Confirmatory Factor Analysis; MCFA = Multilevel CFA.

## EMPIRICAL ILLUSTRATION

The data include responses from 5,357 American children in 273 schools on eight items of the Interest in Mathematics subscale of the PISA (Organization for Economic Cooperation and Development, 2003, 2005). Table 4 lists these items. These eight items were originally classified equally into two factors: interest in and enjoyment of mathematics (INT) and instrumental motivation in mathematics (MOT). This scale was originally designed for measuring at the student level. Thus, according to Lüdtke and colleagues (2011), these constructs are considered formative, where the construct at the school level can be considered the school average of INT and MOT across students. Researchers may inappropriately use disaggregation in this example so we show the impact of disaggregation here. Aggregated analysis in a formative measurement model is sometimes used in practice (Håvold, 2007). Thus, we investigate the impact of aggregation in this example as well.

Two data-analytic methods were used in confirmatory factor analysis: single-level analysis that ignores nested data structure (disaggregated and aggregated) and multilevel analysis to investigate factor structures at both micro and macro levels simultaneously. Both analyses used the ML estimator implemented in Mplus 7.[6] The chi-square test and RMSEA were used for model fit evaluation. For the multilevel analysis, the RMSEA values at the micro and macro levels are computed using Ryu and West's (2009) method.

In the disaggregated analysis, some researchers may reject the two-factor model because of high RMSEA, $\chi^2(19) = 521.04$, RMSEA = .070. In the aggregated analysis, most researchers will reject the two-factor model because of high RMSEA, $\chi^2(19) = 74.503$, RMSEA = .103. The full multilevel analysis revealed that the two-factor model fits well in the macro level, $\chi^2(19) = 9.97$, RMSEA = 0,[7] but the model fit in the micro level is not as good as in the macro level, $\chi^2(19) = 503.72$, RMSEA = .069. Thus, some researchers may inadvertently reject the two-factor model at both levels when they ignore the nested data structure in spite of the fact that the two-factor solution fits well in the macro-level model.

The standardized factor loadings and their *SE*s from the disaggregated, aggregated, and multilevel analyses are reported in Table 4. The factor correlations from disaggregated and aggregated analyses were .664 (*SE* = .010) and .751 (*SE* = .031), respectively, whereas the factor correlations from the multilevel analysis were .658 (*SE* = .010) in the micro level and .896 (*SE* = .066) in the macro level. The differences in the standardized factor loadings, factor correlation, and their *SE*s between the disaggregated analysis and the micro level of the multilevel analysis were trivial because of low ICCs (.017–.056). If ICCs were higher, the difference between two analyses would be higher. However, the differences in standardized factor loadings, factor correlation, and their *SE*s between the aggregated analysis and the macro level of the multilevel analysis were nontrivial. Standardized factor loadings and factor correlations were underestimated in the aggregated analysis.

[6] Items are measured on a 4-point response scale including (1) *Strongly Agree,* (2) *Agree,* (3) *Disagree,* and (4) *Strongly Disagree.* We used the ML estimator here to show the impact of ignoring nesting as an empirical illustration. We also used the diagonally weighted least squares estimator in Mplus (WLSMV) to account for the categorical nature of the indicators (Flora & Curran, 2004). We found that the impact of ignoring nesting was similar for WLSMV and ML.

[7] Based on a reviewer's suggestion, we found that the chi-square goodness-of-fit test for detecting misspecification at the macro level severely deflates Type I error and has low power. That is, we simulated 1,000 data sets from the obtained parameter estimates from MCFA with a saturated micro level and fitted those simulated data with the data-generating model. Given a cluster size of 20, Type I error rates were less than 1% when the numbers of clusters were 50, 150, 250, 350, and 450—Type I error was calculated from the proportion of obtained chi-square values greater than 30.144, which is the critical value with $\alpha = .05$ for the chi-square distribution with $df = 19$. We also fit the simulated data with a severely misspecified model such that Items 4 and 6 loaded on MOT (instead of INT) and Items 2 and 5 loaded on INT (instead of MOT). The power of the chi-square test for detecting this misspecification was less than 4% for the number of clusters of 50, 150, 250, 350, and 450. The power to reject a severely misspecified model was extremely low even though the number of clusters was 450. Based on the result of this brief simulation, future research is needed to evaluate the performance of the chi-square test and provide an alternative approach for detecting misfit at the macro level.

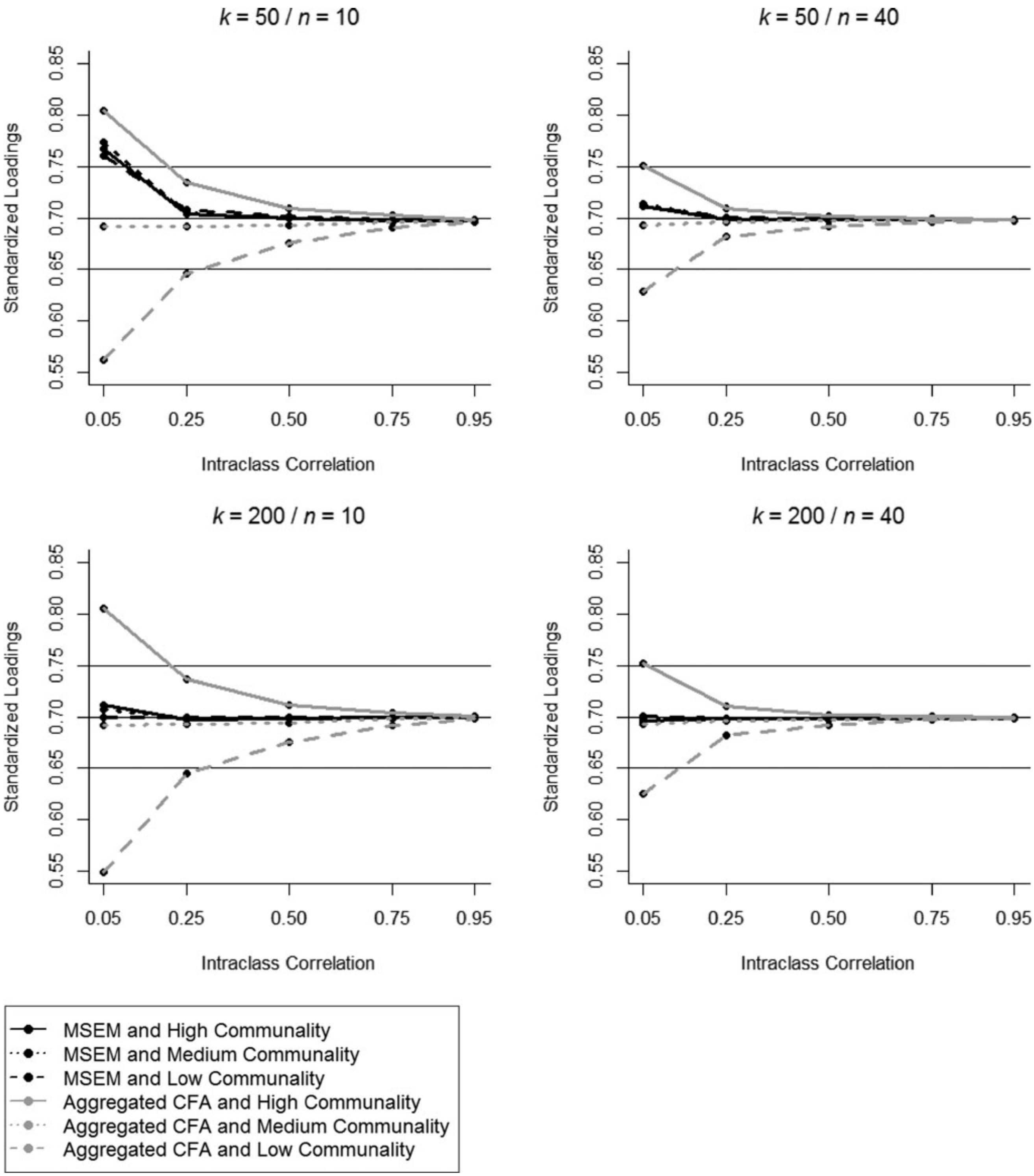

FIGURE 11 The average standardized factor loadings in each condition. *k* is the number of clusters and *n* is the cluster size. The solid horizontal lines in each plot denote absolute biases of –.05, 0, and .05 to represent the acceptable range of bias for the average standardized factor loadings. MSEM = Multilevel Structural Equation Modeling; CFA = Confirmatory Factor Analysis.

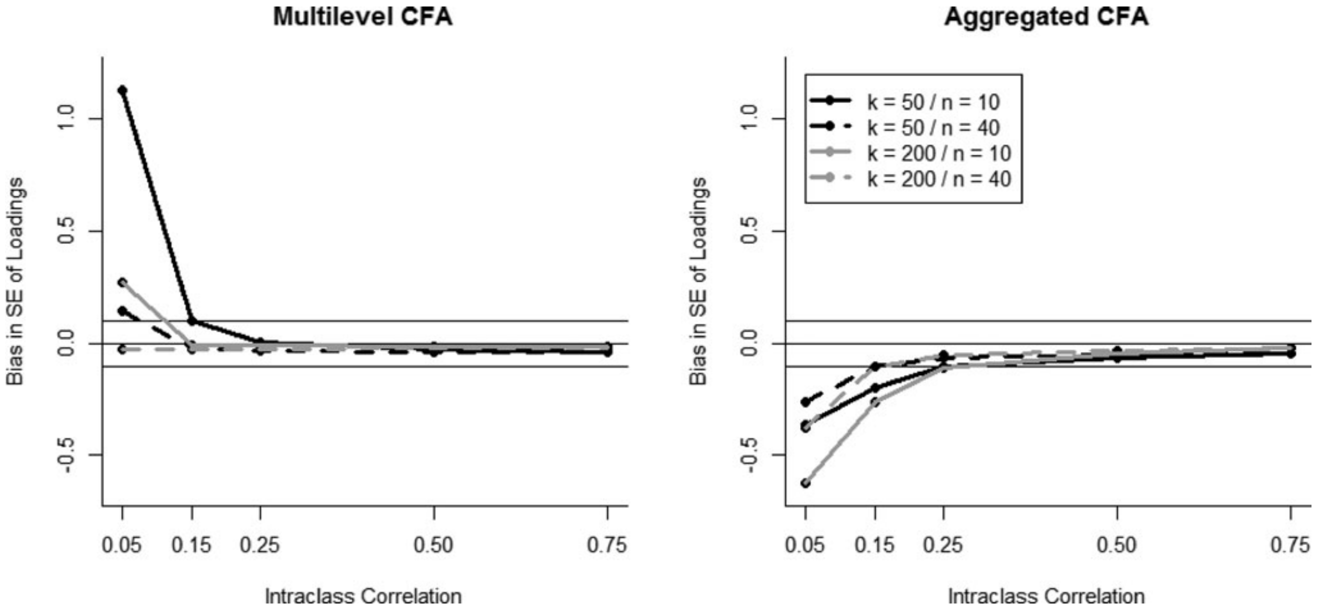

FIGURE 12 The relative bias in standard errors (*SE*s) of standardized factor loadings in each condition. *k* is the number of clusters and *n* is the cluster size. The solid horizontal lines in each plot denote relative biases of –0.1, 0, and 0.1 to represent the acceptable range for the relative bias in the *SE*s. CFA = Confirmatory Factor Analysis.

## DISCUSSION AND CONCLUSION

The purpose of this study was to examine the effects of ignoring clustering by conducting single-level CFA when MCFA is a theoretically more appropriate option. We investigated the effects of ignoring clustering, either by using disaggregation or aggregation, on model fit indices, standardized factor loadings, and factor correlations. We examined standardized parameters rather than unstandardized parameters (e.g., Julian, 2001) because standardized parameters are usually used in interpreting factor analysis results.

When the macro level is ignored, model fit indices (the chi-square statistic and RMSEA) from the disaggregated

TABLE 4
Item Wording for the Interests in Mathematics Scale, Their Intraclass Correlations (ICC), and Their Resulting Standardized Factor Loadings From the Disaggregated and Aggregated Single Level Confirmatory Factor Analysis (CFA) and Full Multilevel CFA (MCFA)

| | | Disaggregated CFA | | Aggregated CFA | | MCFA (Micro) | | MCFA (Macro) | |
|---|---|---|---|---|---|---|---|---|---|
| Items | ICC | INT | MOT | INT | MOT | INT | MOT | INT | MOT |
| 1. I enjoy reading about mathematics. | .032 | .796 (.006) | | .883 (.015) | | .789 (.006) | | .988 (.031) | |
| 3. I look forward to my mathematics lessons. | .056 | .881 (.004) | | .932 (.011) | | .876 (.004) | | .984 (.019) | |
| 4. I do mathematics because I enjoy it. | .029 | .885 (.004) | | .912 (.012) | | .883 (.004) | | .994 (.025) | |
| 6. I am interested in the things I learn in mathematics. | .039 | .850 (.005) | | .884 (.015) | | .844 (.005) | | .988 (.023) | |
| 2. Making an effort in mathematics is worth it because it will help me in the work that I want to do later on. | .020 | | .881 (.006) | | .875 (.017) | | .807 (.006) | | .996 (.056) |
| 5. Learning mathematics is worthwhile for me because it will improve my career (prospects, chances). | .017 | | .832 (.005) | | .854 (.019) | | .830 (.006) | | .984 (.059) |
| 7. Mathematics is an important subject for me because I need it for what I want to study later on. | .018 | | .835 (.005) | | .895 (.015) | | .833 (.006) | | .993 (.053) |
| 8. I will learn many things in mathematics that will help me get a job. | .025 | | .801 (.006) | | .873 (.017) | | .796 (.006) | | .992 (.048) |

*Note*. The values in parentheses are standard errors (*SE*s) of standardized factor loadings. INT = interest in and enjoyment of mathematics; MOT = instrumental motivation in mathematics.

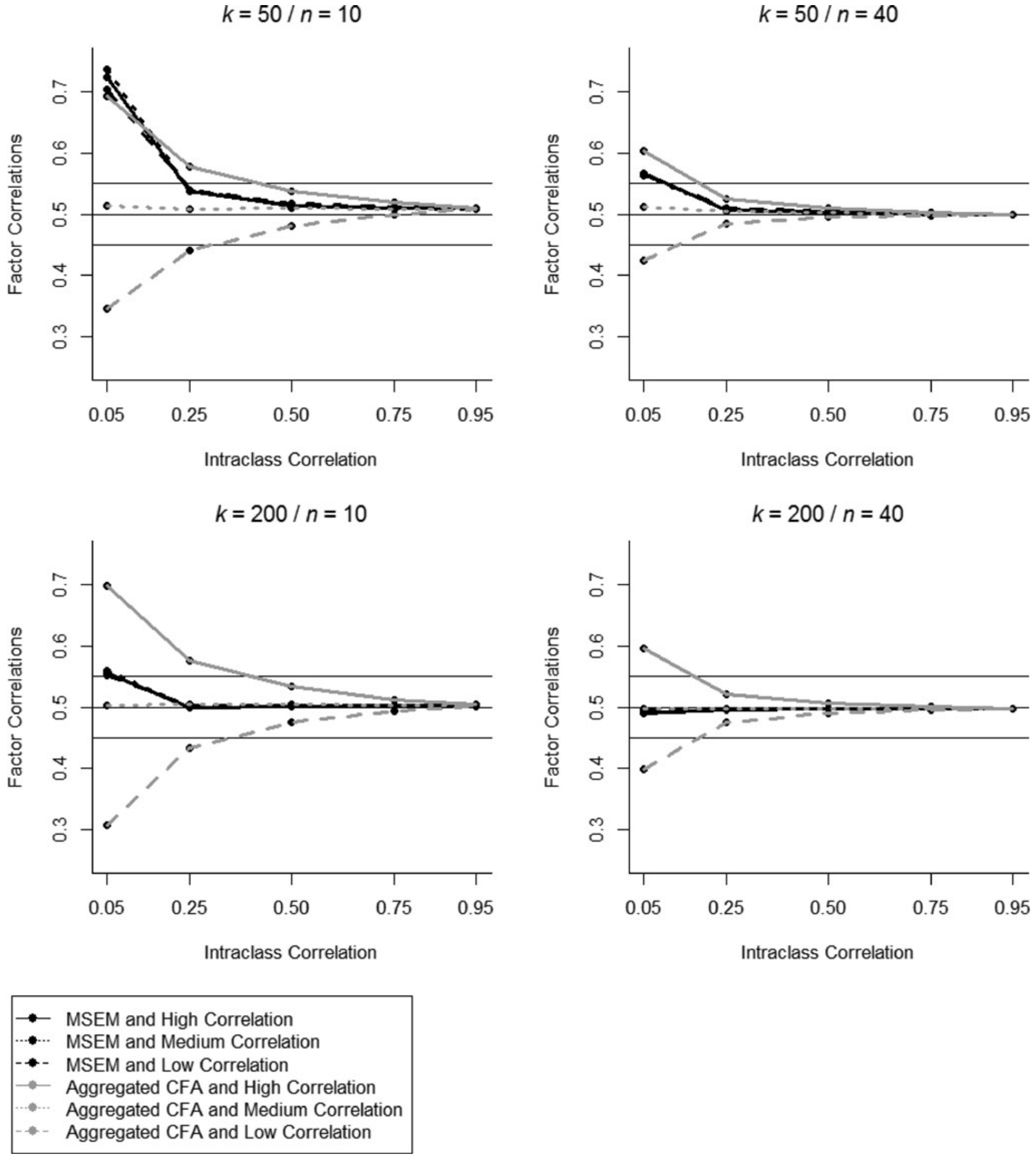

FIGURE 13 The average factor correlation in each condition. *k* is the number of clusters and *n* is the cluster size. The solid horizontal lines in each plot denote absolute biases of –.05, 0, and .05 to represent the acceptable range of bias for the average factor correlation. MSEM = Multilevel Structural Equation Modeling; CFA = Confirmatory Factor Analysis.

single-level CFA indicate poor fit, especially when ICC is large. The model fit indices from the full MCFA and the partially saturated MCFA show good fit. These results show that the model fit of the disaggregated single-level CFA deteriorates when the nested data structure is ignored, even when ICC is very low (.05).

Regarding the standardized parameter estimates, if the micro- and macro-level standardized parameters have the

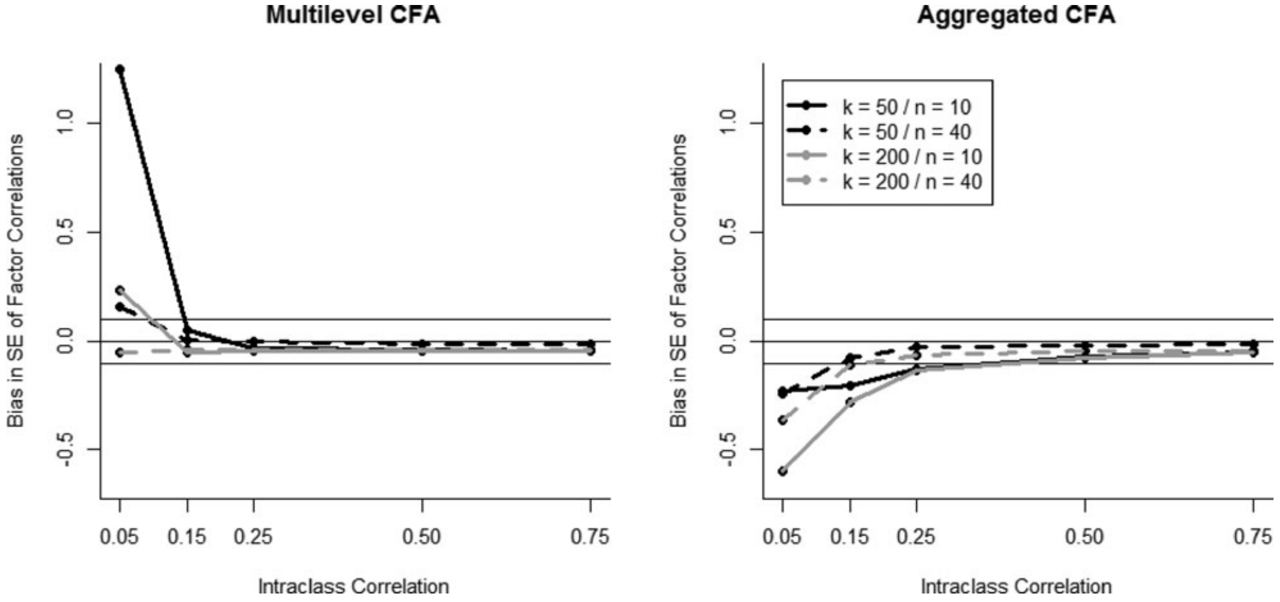

FIGURE 14 The relative bias in standard errors (*SE*s) of factor correlation (right column) in each condition. *k* is the number of clusters and *n* is the cluster size. The solid horizontal lines in each plot denote relative biases of –0.1, 0, and 0.1 to represent the acceptable range of bias for the relative bias in the *SE*s. CFA = Confirmatory Factor Analysis.

same values, the disaggregated estimates are equal to those parameters. If the micro- and macro-level standardized parameters have different values, the disaggregated estimates will be deviated toward the macro-level standardized parameter values. The difference is large when ICC is high (see Equations 14–15 and Online Appendix A). The difference in parameter estimates is ignorable when ICC is less than .15 or the standardized parameters in both levels are equal.

Regarding the *SE*s of standardized parameter estimates, when the macro-level communalities are large, the disaggregated *SE*s are underestimated. When the macro-level communalities are small, the disaggregated *SE*s are overestimated. The degree of under- and overestimation is greater when ICC is higher. The relative difference in *SE*s is ignorable when ICC is .05 or lower.

These results are similar to Julian's (2001) results in that, when ICC is higher, the differences in parameter estimates and their *SE*s are greater. Because this study and Julian's targeted different parameters, two important results are different. First, although Julian found that unstandardized factor loading and factor covariance are always overestimated when ICC is large, our results reveal that standardized parameter estimates are not necessarily different. If standardized coefficients are equal across levels, disaggregated standardized estimates are not biased. On the other hand, disaggregated unstandardized estimates are inflated although unstandardized coefficients are equal across levels (Julian, 2001). If standardized coefficients are not equal, disaggregated standardized parameters are not always overestimated like unstandardized ones (Julian, 2001). If the macro-level standardized coefficients are lower than the micro-level standardized coefficients, the disaggregated standardized estimates will be underestimated.

Second, although Julian (2001) found that the *SE*s of unstandardized estimates are always underestimated when ICC is high, the *SE*s of standardized estimates may be overestimated, underestimated, or unbiased. In the low macro-level communality conditions, the *SE*s of standardized estimates (i.e., standardized factor loadings and factor correlations) are overestimated, whereas in the high macro-level communality conditions, they are underestimated.

In addition, we show in an empirical example that researchers may inadvertently reject a hypothesized (single-level) model even when it fits well in the macro level. Some interesting findings from the macro level can be overlooked if this level is ignored. Based on these findings, we recommend, if possible, using the full MCFA or the partially saturated MCFA. If the number of clusters is small, the fixed effect approach (i.e., multiple-group CFA) can be an option. If MCFA is not a viable option (e.g., if MCFA does not converge), we highly encourage analysts to use the segregation approach for MCFA (Yuan & Bentler, 2007), which provides accurate model fit statistics, parameter estimates, and *SE*s. The sampling-design-based MCFA (Stapleton, 2002, 2006, 2008) can be an option if researchers know how their samples are randomly drawn from their target finite population and wish to interpret parameters uncontrolled for clustering. If researchers adopt the disaggregated single-level CFA, this approach can provide accurate estimates and *SE*s of standardized estimates when, and only when, ICC is very small (.05). However, researchers will still encounter inflated model misfit because of the violation of the independence-of-observations assumption. It is likely that the inflation in model misfit will oblige researchers to reject models that are actually acceptable or modify the models in order to reduce the model misfit, which likely would reduce model parsimony and introduce unnecessary bias.

When the micro level is ignored, the model fit from the aggregated single-level CFA is similar to that from the partially saturated MCFA, especially for high ICC. This finding indicates that both the aggregated CFA and the partially saturated MCFA detect model misfit in the macro level (Ryu & West, 2009). The full MCFA, however, detected the overall model misfit (both micro and macro levels). The aggregated CFA provided the same values of chi-square statistics and RMSEA as partially saturated MCFA.

Regarding the parameter estimates, if the micro- and macro-level standardized coefficients are the same, the aggregated parameter estimates are not biased. If the micro- and macro-level standardized coefficients differ, the aggregated standardized parameter estimates will be biased toward the micro-level standardized estimates. Bias is large when ICC is low or the cluster size is low (See Equations 16–17 and Online Appendix A). The bias in parameter estimates is ignorable when ICC is greater than .25 or the standardized coefficients are equal across levels. Regarding the *SE*s of standardized parameters, the aggregated single-level CFA provides lower *SE*s than the full MCFA, especially when ICC is low. The relative bias is ignorable when ICC is greater than .75.

These findings differ from those of Moerbeek (2004) in that the parameter estimates and *SE*s are unbiased when the micro level is ignored. However, the target parameters in this study are standardized parameters in CFA, and their estimates are biased by aggregation. A potential reason for the biased standardized parameter estimates is that the standardized estimates are functions not only of unstandardized coefficients (i.e., unstandardized factor loadings or factor covariances) but also of indicator or factor variances. If the unstandardized coefficients were not biased, the indicator and factor variances would be overestimated by aggregation (Moerbeek, 2004), thereby biasing the resulting standardized coefficients.

Based on this finding, we still recommend, if possible, using the MCFA approach, except in the case of a formative measurement model with sampling ratio close to 1. If MCFA is not a viable option (e.g., if MCFA does not converge), we highly encourage analysts to use the segregation approach for MCFA (Yuan & Bentler, 2007). If analysts adopt the aggregated single-level CFA, they can obtain accurate parameter estimates and *SE*s of standardized coefficients when ICC is greater than .75, which is quite high and relatively rare in practice.

One limitation of this study is that we do not examine differences when the factor structure differs across level. For example, the numbers of factors in the micro and macro levels were four and five, respectively, in Julian's (2001) simulation study and the disparity in factor structure exacerbated the differences in model fit, unstandardized parameter estimates, and their *SE*s. The differences in parameter estimates, however, may be predicted by the procedures we provided in Online Appendix A. Also, this study does not account for the possibility of random coefficients, such as micro-level standardized factor loadings or standardized factor correlations that differ across clusters. The effects of ignoring the hierarchical structure of nested data when the micro-level coefficients are random would be an interesting topic for future investigation. In this study we investigated the effects of ignoring the nested data structure only when indicators are continuous. Future research is needed to investigate the accuracy of different types of estimator when indicators are not continuous (e.g., ordered categorical) as well as the biases in parameter estimates and standard errors when the nested data structure is ignored.

## REFERENCES

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Ansari, A., Jedidi, K., & Dube, L. (2002). Heterogeneous factor analysis models: A Bayesian approach. *Psychometrika*, *67*, 49–78.

Ansari, A., Jedidi, K., & Jagpal, S. (2000). A hierarchical Bayesian methodology for treating heterogeneity in structural equation models. *Marketing Science*, *19*, 328–347.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling*, *12*, 411–434.

Asparouhov, T., & Muthén, B. (2005). Multivariate statistical modeling with survey data. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*. Retrieved from http://www.fcsm.sites.usa.gov/files/2014/05/2005FCSM_Asparouhov_Muthen_IIA.pdf

Babakus, E., Bienstock, C. C., & Van Scotter, J. R. (2004). Linking perceived quality and customer satisfaction to store traffic and revenue growth. *Decision Sciences*, *35*, 713–737.

Bell, S. J., Mengüç, B., & Stefani, S. L. (2004). When customers disappoint: A model of relational internal marketing and customer complaints. *Journal of the Academy of Marketing Science*, *32*, 112–126.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

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

Breevaart, K., Bakker, A. B., Demerouti, E., & Hetland, J. (2012). The measurement of state work engagement: A multilevel factor analytic study. *European Journal of Psychological Assessment*, *28*, 305–312.

Cassidy, D. J., Hestenes, L. L., Hegde, A., Hestenes, S., & Mims, S. (2005). Measurement of quality in preschool child care classrooms: An exploratory and confirmatory factor analysis of the early childhood environment rating scale-revised. *Early Childhood Research Quarterly*, *20*, 345–360.

Chen, Q., Kwok, O. M., Luo, W., & Willson, V. L. (2010). The impact of ignoring a level of nesting structure in multilevel growth mixture models: A Monte Carlo study. *Structural Equation Modeling*, *17*, 570–589.

Cheung, M. W.-L., & Au, K. (2005). Applications of multilevel structural equation modeling to cross-cultural research. *Structural Equation Modeling*, *12*, 598–619.

Cheung, M. W.-L., Leung, K., & Au, K. (2006). Evaluating multilevel models in cross-cultural research: An illustration with social axioms. *Journal of Cross-Cultural Psychology*, *37*, 522–541.

Chou, C., Bentler, P. M., & Pentz, M. A. (2000). A two-stage approach to multilevel structural equation models: Application to longitudinal data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches and specific examples* (pp. 33–50). Mahwah, NJ: Erlbaum.

Collins, C. J., & Smith, K. G. (2006). Knowledge exchange and combination: The role of human resource practices in the performance of high-technology firms. *Academy of Management Journal*, *49*, 544–560.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351.

Detert, J. R., Schroeder, R. G., & Cudeck, R. (2003). The measurement of quality management culture in schools: Development and validation of the SQMCS. *Journal of Operations Management*, *21*, 307–328.

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *Leadership Quarterly*, *16*, 149–167.

Ebesutani, C., Okamura, K., Higa-McMillan, C., & Chorpita, B. F. (2011). A psychometric analysis of the positive and negative affect schedule for children-parent version in a school sample. *Psychological Assessment*, *23*, 406–416.

Ehrhart, M. G. (2004). Leadership and procedural justice climate as antecedents of unit-level organizational citizenship behavior. *Personnel Psychology*, *57*, 61–94.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*, 286–299.

Garb, H. N., Wood, J. M., & Fiedler, E. R. (2011). A comparison of three strategies for scale construction to predict a specific behavioral outcome. *Assessment*, *18*, 399–411.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*, 72–91.

Gibson, C. B., & Birkinshaw, J. (2004). The antecedents, consequences, and mediating role of organizational ambidexterity. *Academy of Management Journal*, *47*, 209–226.

Glisson, C., & James, L. R. (2002). The cross-level effects of culture and climate in human service teams. *Journal of Organizational Behavior*, *23*, 767–794.

Gong, Y., Chang, S., & Cheung, S.-Y. (2010). High performance work system and collective OCB: A collective social exchange perspective. *Human Resource Management Journal*, *20*, 119–137.

González-Romá, V., Peiró, J. M., & Tordera, N. (2002). An examination of the antecedents and moderator influences of climate strength. *Journal of Applied Psychology*, *87*, 465–473.

Hägglund, G. (1982). Factor analysis by instrumental variables methods. *Psychometrika*, *47*, 209–222.

Han, J., Chou, P., Chao, M., & Wright, P. M. (2006). The HR competencies-HR effectiveness link: A study in Taiwanese high-tech companies. *Human Resource Management*, *45*, 391–406.

Hatami, G., Motamed, N., & Ashrafzadeh, M. (2010). Confirmatory factor analysis of Persian adaptation of multidimensional students' life satisfaction scale (MSLSS). *Social Indicators Research*, *98*, 265–271.

Håvold, J. I. (2007). National cultures and safety orientation: A study of seafarers working for Norwegian shipping companies. *Work and Stress*, *21*, 173–195.

Hoegl, M., & Gemuenden, H. G. (2001). Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. *Organization Science*, *12*, 435–449.

Hoegl, M., Parboteeah, K. P., & Munson, C. L. (2003). Team-level antecedents of individuals' knowledge networks. *Decision Sciences*, *34*, 741–770.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods & Research*, *26*, 329–367.

Jackson, C. J., Levine, S. Z., & Furnham, A. (2003). Gray's model of personality and aggregate level factor analysis. *European Journal of Personality*, *17*, 397–411.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling*, *20*, 265–282.

Jedidi, K., & Ansari, A. (2001). Bayesian structural equation models for multilevel data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 129–157). Mahwah, NJ: Erlbaum.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods*: Chicago, IL: National Educational Resources Chicago.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling*, *8*, 325–352.

Kaplan, D., & Ferguson, A. J. (1999). On the utilization of sample weights in latent variable models. *Structural Equation Modeling*, *6*, 305–321.

Keller, R. T. (2001). Cross-functional project groups in research and new product development: Diversity, communications, job stress, and outcomes. *Academy of Management Journal*, *44*, 547–555.

Kim, E. S., Kwok, O., & Yoon, M. (2012). Testing factorial invariance in multilevel data: A Monte Carlo study. *Structural Equation Modeling*, *19*, 250–267.

King, J. E., & Figueredo, A. J. (1997). The five-factor model plus dominance in chimpanzee personality. *Journal of Research in Personality*, *31*, 257–271.

Klein, K. J., & Kozlowski, S. W. J. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. *Organizational Research Methods*, *3*, 211–236.

Law, B. M. F., Shek, D. T. L., & Ma, C. M. S. (2011). Exploration of the factorial structure of the revised personal functions of the volunteerism scale for Chinese adolescents. *Social Indicators Research*, *100*, 517–537.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual model: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods*, *13*, 203–229.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203–229.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*, 106–124.

Mathisen, G. E., Einarsen, S., Jørstad, K., & Brønnick, K. S. (2004). Climate for work group creativity and innovation: Norwegian validation of the team climate inventory (TCI). *Scadinavian Journal of Psychology*, *45*, 383–392.

Mathisen, G. E., Torsheim, T., & Einarsen, S. (2006). The team-level model of climate for innovation: A two-level confirmatory factor analysis. *Journal of Occupational and Organizational Psychology*, *79*, 23–35.

Merrell, K. W., Felver-Gant, J. C., & Tom, K. M. (2011). Development and validation of a parent report measure for assessing socio-emotional competencies of children and adolescents. *Journal of Child and Family Studies*, *20*, 529–540.

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, *39*, 129–149.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Muthén, B. O. (1990). *Mean and covariance structure analysis of hierarchical data*. Los Angeles, CA: UCLA Statistics Series, #62.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*, 376–398.

Muthén, B. O., & Asparouhov, T. (2009). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143–165). Boca Raton, FL: Chapman & Hall/CRC.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 267–316). Washington, DC: American Sociological Association.

Muthén, L. K., & Muthén, B. O. (1998–2013). *Mplus user's guide: Statistical analysis with latent variables* (7th ed.). Los Angeles, CA: Author.

Nelson, J. M., Canivez, G. L., Lindstrom, W., & Hatt, C. V. (2007). Higher-order exploratory factor analysis of the Reynolds Intellectual Assessment Scales with a referred sample. *Journal of School Psychology*, *45*, 439–456.

Noortgate, W. V., Opdenakker, M. C., & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement*, *16*, 281–303.

Oliver, J. E., Jose, P. E., & Brough, P. (2006). Confirmatory factor analysis of the work locus of control scale. *Educational and Psychological Measurement*, *66*, 835–851.

Opdenakker, M. C., & Van Damme, J. (2000). Effects of schools, teaching staff and classes on achievement and well-being in secondary education: Similarities and differences between school outcomes. *School Effectiveness and School Improvement*, *11*, 165–196.

Organization for Economic Cooperation and Development. (2003). *PISA 2003 technical report*. Paris, France: Author.

Organization for Economic Cooperation and Development. (2005). *PISA 2003 data analysis manual*. Paris, France: Author.

Patterson, M. G., West, M. A., Shackleton, V. J., Dawson, J. F., Lawthom, R., Maitlis, S., ... Wallace, A. M. (2005). Validating the organizational climate measure: Links to managerial practices, productivity and innovation. *Journal of Organizational Behavior*, *26*, 379–408.

Philips, G. A., Shadish, W. R., Murray, D. M., Kubik, M., Lytle, L. A., & Birnbaum, A. S. (2006). The center for epidemiologic studies depression scale with a young adolescent population: A confirmatory factor analysis. *Multivariate Behavioral Research*, *41*, 147–163.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190.

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2007). Multilevel structural equation modeling. In S. Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 209–227). Amsterdam, The Netherlands: Elsevier.

Raspa, M., Bailey, D. B., Jr., Olmsted, M. G., Nelson, R., Robinson, N., Simpson, M. E., ... Houts, R. (2010). Measuring family outcomes in early intervention: Findings from a large-scale assessment. *Exceptional Children*, *76*, 496–510.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Sampson, R. (1999). Assessing direct and indirect effects in multilevel designs with latent variables. *Sociological Methods and Research*, *28*, 123–153.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*, 173–184.

Raykov, T. (2004). Estimation of maximal reliability: A note on a covariance structure modelling approach. *British Journal of Mathematical and Statistical Psychology*, *57*, 21–27.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling*, *9*, 195–212.

Riordan, C. M., Vandenberg, R. J., & Richardson, H. A. (2005). Employee involvement climate and organizational effectiveness. *Human Resource Management*, *44*, 471–488.

Robert, C., & Wasti, S. A. (2002). Organizatinal individualism and collectivism: Theoretical development and an empirical test of a measure. *Journal of Management*, *28*, 544–566.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, *16*, 583–601.

Salanova, M., Agut, S., & Peiró, J. M. (2005). Linking organizational resources and work engagement to employee performance and customer loyalty: The mediation of service climate. *Journal of Applied Psychology*, *90*, 1217–1227.

Schaubroeck, J., Lam, S. S. K., & Cha, S. E. (2007). Embracing transformational leadership: Team values and the impact of leader behavior on team performance. *Journal of Applied Psychology*, *92*, 1020–1030.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling*, *9*, 475–502.

Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling*, *13*, 28–58.

Stapleton, L. M. (2008). Variance estimation using replication methods in structural equation modeling with complex sample data. *Structural Equation Modeling*, *15*, 183–210.

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

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research*, *44*, 711–740.

Subramony, M., Krause, N., Norton, J., & Burns, G. N. (2008). The relationship between human resource investments and organizational performance: A firm-level examination of equilibrium theory. *Journal of Applied Psychology*, *93*, 778–788.

Takeuchi, R., Lepak, D. P., Wang, H., & Takeuchi, K. (2007). An empirical examination of the mechanisms mediating between high-performance work systems and the performance of Japanese organizations. *Journal of Applied Psychology*, *92*, 1069–1083.

Tucker, C. M., Rice, K. G., Hou, W., Kaye, L. B., Nolan, S. E. M., Grandoit, D. J., ... Desmond, F. F. (2011). Development of the motivators of and barriers to health-smart behaviors inventory. *Psychological Assessment*, *23*, 487–503.

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

van der Vegt, G. S., & Bunderson, J. S. (2005). Learning and performance in multidisciplinary teams: The importance of collective team identification. *Academy of Management Journal*, *48*, 532–547.

van Mierlo, H., Vermunt, J. K., & Rutte, C. G. (2009). Composing group-level constructs from individual-level survey data. *Organizational Research Methods*, *12*, 368–392.

Wang, M.- T., Willett, J. B., & Eccles, J. S. (2011). The assessment of school engagement: Examining dimensionality and measurement invariance by gender and race/ethnicity. *Journal of School Psychology*, *49*, 465–480.

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, *28*, 263–311.

Wilkinson, L., & The Task Force of Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist*, *54*, 594–604.

Yuan, K.-H., & Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociological Methodology*, *37*, 53–82.

Zhou, K. Z., Gao, G. Y., Yang, Z., & Zhou, N. (2005). Developing strategic orientation in China: Antecedents and consequences of market and innovation orientations. *Journal of Business Research*, *58*, 1049–1058.

Zohar, D., & Tenne-Gazit, O. (2008). Transformational leadership and group interaction as climate antecedents: A social network analysis. *Journal of Applied Psychology*, *93*, 744–757.

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice*, *12*, 127–140.