Finite Population Correction for Two-Level Hierarchical Linear Models

Mark H. C. Lai

University of Cincinnati

Oi-man Kwok, Yu-Yu Hsiao, and Qian Cao

Texas A&M University

Author Note

Mark H. C. Lai, School of Education, University of Cincinnati; Oi-man Kwok, Department of Educational Psychology, Texas A&M University; Yu-Yu Hsiao, Department of Educational Psychology, Texas A&M University; Qian Cao, Department of Educational Psychology, Texas A&M University.

Correspondence concerning this article should be addressed to Mark Lai, School of Education, University of Cincinnati, Cincinnati, OH 45221.

Email: mark.lai@uc.edu

An earlier version of this work was presented at the 2015 annual meeting of the American Educational Research Association.

Abstract

The research literature has paid little attention to the issue of finite population at a higher level in hierarchical linear modeling. In this article, we propose a method to obtain finite-population-adjusted standard errors of level-1 and level-2 fixed effects in two-level hierarchical linear models. When the finite population at level 2 is incorrectly assumed as being infinite, the standard errors of the fixed effects are overestimated, resulting in lower statistical power and wider confidence intervals. The impact of ignoring finite population correction is illustrated by using both a real data example and a simulation study with a random intercept model and a random slope model. Simulation results indicated that the bias in the unadjusted fixed-effect standard errors was substantial when the level-2 sample size exceeded 10% of the level-2 population size; the bias increased with a larger intraclass correlation, a larger number of clusters, and a larger average cluster size. We also found that the proposed adjustment produced unbiased standard errors, particularly when the number of clusters was at least 30 and the average cluster size was at least 10. We encourage researchers to consider the characteristics of the target population for their studies and adjust for finite population when appropriate.

Keywords: hierarchical linear model, finite population, sampling

Finite Population Correction for Two-Level Hierarchical Linear Models

In the past decade, hierarchical linear modeling (HLM) has become a popular choice for handling educational and behavioral data with a multilevel structure. However, the theory behind HLM was developed for cases where observations are sampled from a population of infinite size, with little attention given to situations where, at some higher level, the sampled units are a subset of a finite target population.

As discussed in Cochran (1977), for sampling without replacement, when the sample size exceeds the population size by as little as 5%, finite population corrections (FPCs) need to be applied to sample estimates. That is, supplying realistic information about the population yields more accurate inference with higher statistical power and narrower confidence intervals. Because applications of FPC to HLM have not been thoroughly discussed in the literature, in the present study we proposed an adjustment method for the fixed-effect standard errors (*SE*s), illustrated it with a real data set, and evaluated its performance using simulations.

Many data sets in the social sciences are collected using sampling schemes other than simple random sampling. Schemes such as cluster sampling or other multistage sampling methods result in multilevel data in which, for example, students are nested within schools, employees are nested within organizations, and residents are nested within countries. It is usually assumed that the sampled units at a higher level may be considered a random sample of some population of interest; otherwise, any known sampling biases need to be adjusted. Conventional single-level statistical models (e.g., multiple regression), on the other hand, assume that observations are independent, an assumption that is violated for nested data. As a result, use of single-level models that ignores the dependent nature of the data leads to biased (and generally underestimated) *SE*s (e.g., Snijders & Bosker, 2012), and analytical approaches that take into account the dependency among observations, such as HLM, therefore, are needed to correctly model data with complex sampling design.

Finite Populations and Modes of Statistical Inference

The very concept of a finite population is beyond the scope of what many researchers have learned in applied statistics courses. For example, basic statistics courses traditionally present the notion that population parameters are almost always unknowable precisely because the population is innumerable. One possible reason for the lack of references to a finite population beyond survey data relates to the popularity of model-based inference for many areas of social science research, whose validity depends on a correctly specified model, as opposed to design-based inference, whose validity depends on successfully accounting for the sampling design. Another possible reason is that populations for many single-level studies in the social sciences are huge and that the negligence of FPCs, therefore, hardly makes a difference.

Below we briefly discuss the use of FPCs in single-level studies as part of the historical debate over model-based versus design-based inference and position the use of FPCs for multilevel studies in the context of fixed and random group effects. Readers interested in a more in-depth discussion of model-based and design-based inference are referred to Little (2004), Smith (1994), Snijders and Bosker (2012, Chapter 14), and Sterba (2009).

Model-Based Versus Design-Based Inference

Historically, statisticians have been split into two camps with regard to modes of statistical inference: design-based and model-based (Smith, 1994). In design-based inference, the sampling scheme is generally well defined, and the aim is to make inference about a well-defined finite population. For example, if one is interested in the city average of subjective well-being in City A with a population size of *K*, one can draw a simple random sample of size *k* from the list of all citizens in City A and get an unbiased sample mean $\bar{Y}$ for the finite population (FP) mean: $\mu^{\text{FP}} = \sum_{i=1}^{K} Y_i / K$. When sampling without replacement, the sampling variance of $\bar{Y}$ equals the usual expression under an infinite population, S^2 / k (where S^2 is the variance of Y for the finite population), times the finite population correction (FPC) factor, (K – k) / K (Cochran, 1977, p. 24). If, instead, the sample is complex (e.g., clustered by schools), it is not possible to obtain valid inference on the population characteristics without incorporating those design features in the sampling process (e.g., Binder & Roberts, 2003).

On the other hand, the model-based approach assumes an underlying probability model with certain distributional assumptions and parameters that generates the observed sample data, and the goal is to make inference about the parameters of the model. If we consider the previous example of finding the population mean of City A using a model-based framework, one simple and popular model would be to assume that each observation is generated from a normal distribution with mean μ and variance σ^2, and that all observations are independent and identically distributed. (Note that μ and σ^2, the model parameters, are not necessarily the same as μ^{FP} and S^2 under the finite population, unless the assumed model is an accurate depiction of the finite population.) A usual estimator of the model parameter μ would also be the sample mean $\bar{Y}$, with the sampling variance of $\bar{Y}$ equaling σ^2 / K. Using the model-based framework, the validity of the statistical inference depends on how closely the model approximates the real data-generating mechanism. Although random sampling is not explicitly invoked or required in a model-based framework (Sterba, 2009), use of a nonrandom sample can result in biased estimates if the sampling mechanism is informative but not taken into account (Little, 2004).

For single-level studies in the behavioral sciences, use of design-based inference and FPCs has not been the norm. One reason is that the population of interest commonly comprises a huge number of people (e.g., citizens in a country) and that the sample for analysis usually represents a very small fraction of the population, making the need for FPC minimal. Another appeal to model-based inference is the reference to a hypothetical infinite population from which the finite population is drawn. In previous studies, the hypothetical infinite population is commonly denoted as the *superpopulation* (Cochran, 1977; Gelman & Hill, 2006; Hatley & Sielken, 1975; Lohr, 2010). In this paper, we will use the terms *superpopulation* and *infinite population* interchangeably. The conceptualization of a superpopulation was motivated by the fact that researchers may not be interested only in the current fixed and finite population but in a phenomenon that is believed to occur universally. As an example, consider a researcher investigating the relationship between amount of exercise and well-being in a random sample of adults in New York City in 2015. The researcher may not be interested in only generalizing the relationship to adults in New York City or to adults in the United States in the year of 2015, but may presume that such a relationship is relatively stable (at least for a certain period) and holds across countries. In the latter case, the superpopulation is innumerable and close to infinite, and the omission of FPC is usually justifiable even when the sample size is not much smaller than the size of the finite population.

Finally, many of the techniques used in the social sciences such as regression analysis, path analysis, and structural equation modeling are developed primarily within the model-based framework, whereas design-based inference is more limited to survey data where the target of generalization is usually a well-defined finite population. (As a side note, even if a researcher collects data from a whole finite population like a census in a country, it is still possible to use a probability model to obtain *SE*s for the estimates by assuming a superpopulation, whereas with a design-based approach there will be no sampling variability, and the *SE*s will be zero.)

Integration of Design-Based and Model-Based Inference

Despite the discrepancies between the two modes of statistical inference, the recent literature has mainly focused on how the two modes can be “reconciled” (Smith, 1994, p. 5). Model-based approaches are not valid when important sampling features are not incorporated into the model (e.g., Little, 2004), whereas design-based approaches may be less efficient and limited to research with a rigorous sampling scheme, which tends to be the exception in many areas of the social sciences (Sterba, 2009).

Two examples of integrating model-based and design-based inference include (a) the development of multilevel models to analyze data with clustering structure arisen from cluster sampling (cf. Raudenbush & Bryk, 2002); and (b) discussion of how to incorporate other sampling features, such as sampling weights for unequal probability of selection, into model-based methods such as structural equation modeling (e.g., Cai, 2013; Muthén & Satorra, 1995; Pfeffermann, Skinner, Holme, Goldstein, & Rasbash, 1998; Stapleton, 2002). In addition, philosophical discussions have focused on the integration of design-based and model-based approaches. For example, Särndal, Swensson, and Wretman (1992) advanced a model-assisted design-based framework where a model is hypothesized on the finite population but the inference follows the design-based tradition. Little (2004) suggested the use of a model-based regression model that actively takes into account design features. And more recently, Sterba (2009) proposed a hybrid framework that adjusts model-based estimations for sample design features, such as disproportionate selections and stratifications, for both single- and multilevel studies.

Why Finite Populations and HLM?

Target of Inference: Finite Population Versus Superpopulation

Although the omission of FPC in single-level studies may be justified by referencing an extremely large finite population or a hypothetical infinite superpopulation, for two-level studies the population of interest at level 2 is sometimes fixed and finite. Indeed, when FPC is used in single-level studies, the unit of analysis is usually at a level higher than the person level (or level 1, which contains the smallest units in the analysis), such as all households in three communities in Nicaragua (Brune & Bossert, 2009), all households in one community in Canada (Wilkinson, 2007), and all hospitals in Korea (Yoon, Chang, Kang, Bae, & Park, 2012).

Fixed and random group effects. Within the framework of model-based inferences, researchers have distinguished between fixed and random group effects for studies with observations in groups (see Gelman, 2005, and Gelman & Hill, 2006, for an overview and the controversy). Fixed group effects are commonly invoked when (a) the set of all possible level-2 “units” are few and countable; (b) when the sample includes all “units” in the data collection (e.g., Green & Tukey, 1960); and (c) when the group effects are relatively stable (Searle, Casella, & McCulloch, 2006, sections 1.3 and 1.4). For example, some cross-cultural studies may only include two or three countries (e.g., the United States and China), and applied researchers usually are only interested in making inferences about the effects for the two or three countries in the sample. Thus, they almost always treat country as a fixed effect by using procedures such as the traditional fixed-effect analysis of variance (ANOVA), regression, or multiple-group SEM. Note that in this setup, there are only generalizations for people, not for countries. On the other hand, levels such as “classrooms” may represent a huge collection of units to which a researcher would like to generalize, and such levels are usually treated as random; that is, as a random sample of the collection of all possible “classrooms” (or the superpopulation of “classrooms”), and random-effect ANOVA or HLM is used. Therefore, generalization takes place at both level 1 and level 2.

The choice of treating group effects as fixed or random also depends on practical issues. First, parameter estimations with random group effects may be biased when the number of groups is small (e.g., fewer than 10 or 20; see Hox, 2010), and under such circumstances, it has been shown that treating group effects as fixed provide good point and interval estimations for level-1 regression coefficients (McNeish & Stapleton, 2016a). However, treating group effects as fixed makes it difficult to examine level-2 predictors when explaining group differences, as the level-2 predictors will be completely collinear to the group indicators (e.g., Allison, 2009), whereas adding level-2 predictors to the level-2 regression equation is natural for random group effects.

Nevertheless, it is not always clear whether one should treat the group effects as fixed or as random (Searle et al., 2006, section 1.4). Using an example discussed by Gelman and Hill (2006), if one collects data from all 50 states in the United States, one can treat the effect of states (level 2) as fixed such that the target of inference is only the 50 states within the United States, or one can treat them as random such that the results also generalize to existing “states” (or comparable units) in other countries or to hypothetical “states” (in the future or in some imaginary world). In other words, the target of inference is a superpopulation of “states.” The choice is usually based on the researcher’s judgment. In such an example, Gelman and Hill (2006) suggested that it is probably “more meaningful” (p. 461) to generalize to the finite population of the 50 U.S. states; however, other researchers may be more interested in the superpopulation of “states” across countries. The point is that it is important to explicitly consider the target of inference rather than simply treating the grouping variable as fixed (e.g., in multiple regression with dummy coding) or as random (e.g., in HLM) without justification.

In some situations, the conventional way of treating the group effects as either fixed or random is not ideal. As noted by Gelman and Hill (2006), this fixed-versus-random dichotomy "[left] open the question of what to do with a large but not exhaustive sample" (p. 245). Continuing with the U.S. states example, assume that a researcher only collects data from 20 of the 50 states. The researcher would like to generalize the results to the whole nation of the United States but not necessarily to other "states" in other countries. In this case, treating the state effects as fixed means that there is no underlying distribution of interest for the 20 states in the sample and ignores the sampling error if one wants to also generalize to the remaining 30 U.S. states; treating the state effects as random with an infinite level-2 population would overestimate the sampling error if one does not want to go beyond the 50 states. Instead, it is more reasonable to treat the effects of the 20 states as a random sample of a finite population for the effects of the 50 U.S. states and adjust for the level-2 finite population in the same way as in single-level studies. To generalize to the finite population, the sampled states should be considered representative of the 50 U.S. states in the sense of being a random sample (Sterba, 2009); if the sampled states have known differences that affect the outcome variables from the states that are not sampled (e.g., all sampled states are midwest and east coast states), the HLM with FPC model is not valid without adjustment for those differences.

Below we review some research reports applying HLM where the target of generalization at level 2 can be considered a finite population. We then situate the use of FPC with HLM between the fixed and random ends of the group effect continuum.

Multilevel Studies in Which the Level-2 Population Is Finite

Multilevel models (including HLMs) are used to address data dependencies when the sampling designs generate clustered data. However, the default model in most multilevel statistical software assumes that the sampled units for each level make up a negligible portion of the target population (Searle et al., 2006), and researchers using HLM seldom explicitly state their target population or justify their use of a superpopulation. As previously discussed, sometimes observations analyzed by HLM come from populations that are finite in size. Perhaps finite population plays the most important role in cross-cultural research involving multiple nations (in 2015 there are fewer than 200 countries in the world by most standards). For instance, Peretz and Fried (2012) studied performance appraisal practices adopted by organizations from 21 countries; in a recent meta-analysis, Rockstuhl, Dulebohn, Ang, and Shore (2012) studied leader-member exchange based on samples from 23 countries. Although these studies used HLM, which assumes an infinite population, careful examination may reveal that, in practice, the target population in these studies only included a limited number of units.

But there are other instances, besides cross-cultural research and national survey studies, where a finite and enumerable population at the higher level could be of interest. For example, Nielsen (2009) studied 165 out of 269 companies listed on the Swiss Stock Exchange for factors affecting top management homogeneity. Mani, Anita, and Rindfleisch (2007) examined the relationship between entry mode and equity level in 4,459 subsidiaries nested within 858 foreign direct investment firms in Japan, emphasizing that the sample represented 40 percent of all the Japanese subsidiaries. Other examples include the largest cosmetics companies in the world (Armonas, Druteikiene, & Marcinskas, 2010), army companies in Haiti (Bliese, Halverson, Schriesheim, 2002), teams in major league baseball (Todd, Crook, & Barilla, 2005), and privatized state-owned enterprises in Taiwan (Wu, Su, & Lee, 2008).

We argue that in similar situations where one has sampled a nontrivial portion (e.g., > 10%) of level-2 units from the target level-2 population, one can obtain a more accurate estimate of the sampling variability by applying FPC to the *SE*s obtained under HLM, and thus the procedure differs from a conventional random-effect model in the sense that the level-2 population is finite and well defined. For a given study, how one treats the group effects should be based on one’s judgment (Hatley & Sielken, 1975) and should be determined before carrying out the analyses, as such a decision results in different degrees of precision of the parameter estimates, as discussed in the next section (see also Gelman, 2005). The HLM with FPC approach is most useful in situations where the population is clearly defined with a known and limited size, such as the examples reviewed above.

Trade-Off Between Using a Finite Population and a Superpopulation

Because the use of FPCs usually results in smaller *SE*s and researchers may abuse this approach by arbitrarily redefining a given population to deflate the uncertainty of their estimate and to obtain statistical significance, it is important to look at the trade-off between smaller *SE*s (by using a finite population) and result generalizability (by using a superpopulation) at level 2 before going into derivations and Monte Carlo simulations.

We summarize the different scenarios discussed in this section in Table 1. The table mainly serves to emphasize the role of researchers’ judgment in deciding the target of generalization and is by no means an extensive account of all possible analytical approaches. Again, the decision to generalize to a finite population or to a superpopulation should be explicitly stated and justified during the design phase of the study and before the data analysis.

Consider a hypothetical study where the sample includes data from participants in 30 countries collected in 2015 (corresponding to case 3 in Table 1). The researchers can choose to treat the group effects as (a) random with a superpopulation, (b) random with a finite population, or (c) fixed. Choosing (a) implies that the researchers want to generalize their findings not only to all existing countries in 2015, but also to countries that may be established in the following decades or centuries, as long as those new “countries” can be considered similar to the existing countries. Using this framework, the presumed population will be big, so inferences using the usual HLM will be justified, and the validity of generalizations will depend on how similar and comparable the new “countries” in the future are to the existing countries in 2015 with regard to the phenomenon of interest. Indeed, many published studies that used HLM on cross-cultural data fall into this category (e.g., Davidov, Dülmer, Schlüter, Schmidt, & Meuleman, 2012; Jäckle & Wenzelburger, 2015; Rudnev, 2014; Smits & Huijts, 2015).

On the other hand, the researchers can choose (b) if they decide that their target of generalization is all the existing countries in 2015 and potentially a few years before and after 2015, if the overall number of countries in the population is assumed to be relatively stable. This population will be smaller, and the use of HLM with FPCs will be more appropriate. In contrast, using HLM without FPCs in these situations result in overestimated *SE*s and confidence intervals that are too wide.

Finally, the researchers can choose (c) and carry out a fixed-effect ANOVA or a multiple-group analysis without multilevel modeling if they simply want to compare only the 30 countries present in the data with no intention to generalize the results to all other existing countries.

Similar to previous studies (e.g., Little, 2004), we considered the use of FPC in HLM as incorporating design features into a model-based approach with random group effects. This is analogous to developing multilevel models by modifying the model assumption of independent observations in a multiple-regression model to allow for dependent observations within clusters;

here we proposed to relax the distributional assumptions in HLM to allow for a finite population that is the target of statistical inferences. To ensure valid inference from the level-2 sample to the level-2 population, either the level-2 sample should be representative of the level-2 population or the selection bias in the sample should be adjusted.

Finite Population Basics

Most standard research methods textbooks treat the issue of finite population only briefly, so basic properties in this area may be unfamiliar to readers. The following discussion provides background information in this area, limited to the properties associated with a two-stage cluster sampling scheme where, at each level, units are randomly sampled without replacement. Both the level-1 and the level-2 sample sizes are assumed fixed, which means that sample sizes are specified before sampling begins, so that the sample size decision is not affected by data acquired during the sampling process. This appears to be a common probability sampling method used in studies that employ HLM. Other sampling plans can give rise to other properties.

We will first discuss single-level samples from a finite population and apply the same concept of finite population correction (FPC) to two-level samples. For a single-level sample in which the sample is selected from a known and countable population, the assumption of an infinite population no longer holds. Consider, for example, a random sample of size k from a known population of size K. As discussed in Thompson (2012), the sample mean $\bar{Y}$ of a variable Y in general is an unbiased estimator of the population mean. However, the sampling variance of $\bar{Y}$ depends on the population size, such that

$$\mathrm{Var}(\bar{Y}) = \left(\frac{K-k}{K}\right)\frac{\sigma^2}{k}, \tag{1}$$

where σ^2 is the population variance, which can be estimated by the sample variance,

$$s^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(y_i - \bar{y}\right)^2 . \tag{2}$$

Because s^2 / k is an unbiased estimator of Var($\bar{Y}$) when the population is assumed infinite, it follows that the unbiased estimator under a finite population is equal to its counterpart under an infinite population (s^2 / k), times a correction factor. This correction factor is usually called the *finite population correction* (FPC) factor (Thompson, 2012), where

$$\text{FPC} = \frac{(K-k)}{K} = 1 - \frac{k}{K} \cdot \tag{3}$$

Previous literature suggested that, for single-level studies, finite population becomes an issue when the sample size exceeds even as little as 5% of the population size (Cochran, 1977; Hair, Bush, & Ortinau 2000). Because the sampling variance (i.e., SE^2) of least square estimators of regression coefficients are functions of s^2, the adjusted sampling variance of the regression coefficients may be derived by multiplying the estimated s^2 value obtained from standard software packages (which assumes an infinite population) by the FPC (Cochran, 1977). The FPC is always between 0 and 1, and the closer k is to K, the smaller the correction factor. This implies that correctly applying the FPC results in more appropriate *SE*s of the regression coefficients, and thus better statistical power and confidence intervals, if the target of generalization is finite.

A typical two-stage cluster sampling procedure involves first sampling J level-2 units (or clusters) from a population of size J_{pop}. Then, for each level-2 unit j (j = 1, . . . , J), one samples n_j level-1 units in the jth cluster. Finally one measures the response variable X for each respondent i in cluster j and denotes the response as x_{ij}. If one separates the response x_{ij} into its cluster mean $\bar{x}_{.j}$ and the individual deviation ($x_{ij} - \bar{x}_{.j}$), and assumes that the level-1 units are distributed similarly across clusters after controlling for the cluster mean, one can see that, at level-2, $\bar{X}_{.j}$ (the mean of X for the jth cluster) is a sample of size J from a population of size

J_{pop}; at level-1, ($X_{ij} - \bar{X}_{.j}$) has a size of $N = \sum_{j=1}^{J} n_j$, and is a sample from a population of size N_{pop}. As discussed in subsequent sections, one way to account for finite populations is to apply FPC to the variance components at each level and then express the sampling variance (or *SE*) of the regression coefficients in terms of the adjusted variance components.

For the two-level HLM, the effect of finite populations on the parameter estimates has not been explicitly discussed. Although FPC procedures have been used in two-stage sampling in survey research (e.g., Chromy & Abeyasekera, 2005; Korn & Graubard, 2003), to our knowledge, no previous research has discussed how to incorporate FPC into models with random slopes or coefficients. As Little (2004) noted, model-based approaches such as the HLM may be inappropriate if the sampling process is ignored, so it is important for researchers to incorporate the sampling mechanism in their models.

Hence, the purpose of the current study was to show how researchers can compute *SE*s adjusted for finite populations and obtain correct inference for the fixed effects from their HLM analysis. As discussed in the previous section, we believe that the use of FPC is more relevant and justified at level 2 than at level 1; therefore, we focused more on applying FPC at level 2 to HLM analyses.[1]

We started with a mathematical derivation for a general two-level model that includes a random intercept and multiple random slopes and obtained a matrix expression of the adjusted *SE*s of the fixed effects (which can be solved using the program we provided in Appendix B). We then presented closed-form expressions of the adjusted *SE*s for the special case with no random slopes and with equal cluster sizes, and illustrated the impact of ignoring the finite population issue on *SE*s using data from the World Value Survey 1990–1993 (World Values

Study Group, 1994). In addition, we used a simulation study to examine the performance of the correction with unbalanced cluster sizes.

General Two-Level Linear Mixed Model

We first consider the more general two-level mixed model assuming homogeneous variance of random effects in both levels and with q + 1 level-2 random effects (e.g., q = 1 with one random slope), in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{\gamma}$ is the vector of fixed effects parameters, **X** is an $N \times (p + 1)$ design matrix with all 1s in the first column for the intercept, and the remaining columns containing purely level-1 predictors, purely level-2 predictors, and level-1 predictors with level-2 variances, $\mathbf{u} = \begin{bmatrix} u_0 & \cdots & u_q \end{bmatrix}'$ is a random column vector containing the q + 1 level-2 random effects, and $\boldsymbol{\varepsilon}$ is an $N \times 1$ matrix of level-1 error term. **Z** is an $N \times (q + 1)$ design matrix for the random effects, with all 1s in the first column and the remaining columns are subset of the level-1 variables in **X** that are hypothesized to have a random slope. The two sources for the variability of **y** are **u** and $\boldsymbol{\varepsilon}$. Assume that at the population level, Cov(**u**, $\boldsymbol{\varepsilon}$) = **0**, E($\boldsymbol{\varepsilon}$) = **0**, Var($\boldsymbol{\varepsilon}$) = $\sigma^2\mathbf{I}_N$, E(**u**) = **0**, and Var(**u**) = **G**. For example, for a model with one random intercept and one random slope, we have

$$\mathrm{Var}(\mathbf{u}) = \mathrm{Var}\begin{pmatrix} u_0 \\ u_1 \end{pmatrix} = \mathbf{G} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}.$$

For any model in the form of equation (4), the population variance of **y** is

$$\mathrm{Var}(\mathbf{y}) = \mathbf{V} = \mathrm{Var}(\mathbf{Z}\mathbf{u}) + \mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \sigma^2\mathbf{I}\,. \tag{5}$$

This suggests that **V** can be separated into the level-2 contribution, $\mathbf{ZGZ}'$, and the level-1 contribution, σ^2**I**.

In the infinite population case, for both the maximum likelihood and the restricted maximum likelihood estimation methods, the fixed-effect parameter estimates may be obtained by the generalized least squares (GLS) estimator after the variance matrix **V** has been estimated (Monahan, 2008; Snijders & Bosker, 1993). The vector of estimated fixed effect coefficients, denoted as $\hat{\boldsymbol{\gamma}}$, is (Raudenbush & Bryk, 2002; Snijders & Bosker, 1993)

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \tag{6}$$

with

$$\mathrm{Var}(\hat{\boldsymbol{\gamma}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}, \tag{7}$$

where $\mathbf{X}'$ is the transpose of **X**. The *SE*s are the square roots of the diagonal elements of Var($\hat{\boldsymbol{\gamma}}$).

As shown in Cochran (1977) and Thompson (2012), in two-stage cluster sampling, the sampling variance of a linear function of **y** (e.g., sample mean and the GLS estimator defined in [6]) adjusted for finite population may be obtained by applying the corresponding FPC to each level (see also Appendix A). Thus, with FPC applied, the GLS estimator has a covariance matrix

$$\mathrm{Var}^{\mathrm{FP}}(\hat{\boldsymbol{\gamma}}) = (\mathbf{X}'\mathbf{V}^{*-1}\mathbf{X})^{-1}, \tag{8}$$

where

$$\mathbf{V}^{*} = \mathrm{FPC}_{2} \times \mathbf{ZG}\,\mathbf{Z}' + \mathrm{FPC}_{1} \times \sigma^{2}\mathbf{I} = \mathbf{ZG}^{*}\mathbf{Z}' + \sigma^{2^{*}}\mathbf{I}\ , \tag{9}$$

$\mathbf{G}^{*}$ = FPC$_2$ × **G** and σ^{2*} = FPC$_1$ × σ^2, and FPC$_2$ and FPC$_1$ are the finite population correction factors at level 2 and at level 1, respectively. With unbalanced cluster sizes, closed-form solutions for $\mathrm{Var}^{\mathrm{FP}}(\hat{\boldsymbol{\gamma}})$ are complex and involve inversion of matrices. Therefore, we provide the R code for obtaining the adjusted *SE*s for the fixed effects in general two-level HLMs in Appendix B, which does not assume equal cluster sizes.

The Special Case for a Random Intercept Model

On the other hand, with equal cluster sizes, we can obtain closed-form solutions for $\mathrm{Var}^{\mathrm{FP}}(\hat{\boldsymbol{\gamma}})$ both at level 1 and at level 2 with the random intercept model, which can give an approximate adjustment for the fixed-effect *SE*s when the raw data are not available, and also provide some insight into what may affect the magnitude of the adjustment.

In a random intercept model with a balanced design such that $n_j = n$ for all j, **G** is reduced to a scalar, as the only random effect at level 2 is the random intercept, with Var(u_0) = τ_{00}. For the design matrix **X**, often clustering is also present in the level-1 predictors so that raw (uncentered) level-1 predictors can be further decomposed into the group means and the level-1 residuals through group-mean centering (see Enders & Tofighi, 2007; Raudenbush & Bryk, 2002). Centering is an important issue for multilevel models (Hofmann & Gavin, 1998), and different centering approaches can result in different parameter estimates and interpretations (Enders & Tofighi, 2007; Kreft, de Leeuw, & Aiken, 1995). The group means can be used as level-2 variables when group-mean centering is applied. Although the adjustment procedure discussed in the previous section can accommodate different centering choices, to simplify the discussion, in this section we assume that the level-1 predictors have been group-mean centered with the group means entered as level-2 predictors and that, as a result, all level-1 predictors have zero means and are independent of the level-2 predictors (Enders & Tofighi, 2007). Enders and Tofighi (2007) showed that when group mean centering is used, the design matrix for level-1 predictors, $\mathbf{X}^{(1)}$, and the design matrix for level-2 predictors, $\mathbf{X}^{(2)}$, are orthogonal. Thus,

$$\mathrm{Var}(\hat{\boldsymbol{\gamma}}) = \left(\begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(2)} \end{bmatrix}' \mathbf{V}^{-1} \begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(2)} \end{bmatrix} \right)^{-1} = \begin{bmatrix} (\mathbf{X}'^{(1)}\mathbf{V}^{-1}\mathbf{X}^{(1)})^{-1} & 0 \\ 0 & (\mathbf{X}'^{(2)}\mathbf{V}^{-1}\mathbf{X}^{(2)})^{-1} \end{bmatrix},$$

implying that $\hat{\boldsymbol{\gamma}}^{(1)}$, the vector of estimated coefficients of level-1 fixed effects, and $\hat{\boldsymbol{\gamma}}^{(2)}$, the vector of estimated coefficients of level-2 fixed effects (including the group means of the level-1 predictors), are uncorrelated.

Under the assumption of homogeneity of both σ^2 and τ$_{00}$ for all observations, Snijders (2005) showed that

$$\mathrm{Var}(\hat{\boldsymbol{\gamma}}^{(1)}) = (\mathbf{X}'^{(1)}\mathbf{X}^{(1)})^{-1}\sigma^2 . \tag{10}$$

Applying the correction factor at level 1, the adjusted covariance matrix of $\hat{\boldsymbol{\gamma}}^{(1)}$ is

$$\mathrm{Var}^{\mathrm{FP}}(\hat{\boldsymbol{\gamma}}^{(1)}) = (\mathbf{X}'^{(1)}\mathbf{X}^{(1)})^{-1}\sigma^{2*} = (\mathbf{X}'^{(1)}\mathbf{X}^{(1)})^{-1}\sigma^2 \times \mathrm{FPC}(\sigma^2) = \mathrm{Var}(\hat{\boldsymbol{\gamma}}^{(1)}) \times \mathrm{FPC}_1 . \tag{11}$$

Thus,

$$SE^{\mathrm{FP}}(\hat{\boldsymbol{\gamma}}^{(1)}) = SE(\hat{\boldsymbol{\gamma}}^{(1)}) \times \sqrt{\mathrm{FPC}_1} , \tag{12}$$

For level-2 predictors, if we denote **w***$_j$* as a vector for the level-2 predictor values for the *j*th cluster, and **1***$_j$* is a vector of all ones of length *n$_j$*, we can write $\mathbf{X}_j^{(2)}$ as $\mathbf{1}_j \mathbf{W}_j'$. Define **W** as the aggregated level-2 predictor matrix with *J* rows, where the *j*th row is $\mathbf{w}'_j$. When all clusters have the same number of observations, Snijders (2005) showed that

$$\mathrm{Var}(\hat{\boldsymbol{\gamma}}^{(2)}) = (\mathbf{W}'\mathbf{W})^{-1}\left(\tau_{00} + \frac{\sigma^2}{n}\right). \tag{13}$$

Using the FPC for τ$_{00}$ at level 2, the adjusted covariance matrix of $\hat{\boldsymbol{\gamma}}^{(2)}$ is

$$\mathrm{Var}^{\mathrm{FP}}(\hat{\boldsymbol{\gamma}}^{(2)}) = (\mathbf{W}'\mathbf{W})^{-1}\left(\tau_{00} \times \mathrm{FPC}_2 + \frac{\sigma^2 \times \mathrm{FPC}_1}{n}\right). \tag{14}$$

With τ$_{00}$ and σ^2 unknown in the model, they can be replaced with the estimated values, yielding

$$\mathrm{Var}^{\mathrm{FP}}(\hat{\boldsymbol{\gamma}}^{(2)}) = (\mathbf{W}'\mathbf{W})^{-1}\left(\hat{\tau}_{00} \times \mathrm{FPC}_2 + \frac{\hat{\sigma}^2 \times \mathrm{FPC}_1}{n}\right) \propto \hat{\tau}_{00}^{*} + \frac{\hat{\sigma}^{2*}}{n} , \tag{15}$$

where $\hat{\tau}^{*}_{00}$ = FPC$_2$ × $\hat{\tau}_{00}$. Therefore, one can estimate the adjusted *SE* as

$$SE^{\mathrm{FP}}(\hat{\gamma}^{(2)}) = SE(\hat{\gamma}^{(2)}) \times \sqrt{\frac{\hat{\tau}^{*}_{00} + \dfrac{\hat{\sigma}^{2*}}{n}}{\hat{\tau}_{00} + \dfrac{\hat{\sigma}^{2}}{n}}}. \tag{16}$$

Two observations are worth mentioning. First, from equation (12), we see that for a purely level-1 predictor, only FPC at level-1 affects the *SE*, whereas from equation (16), for a purely level-2 predictor, both FPC$_1$ and FPC$_2$ affect the *SE*. Second, from equation (16), if we define $\hat{\delta} = \hat{\tau}_{00} / \hat{\sigma}^2$, we can write

$$SE^{\mathrm{FP}}(\hat{\gamma}^{(2)}) = SE(\hat{\gamma}^{(2)}) \times \sqrt{1 - \frac{(1-\mathrm{FPC}_2)\hat{\delta} + (1-\mathrm{FPC}_1)/n}{\hat{\delta} + 1/n}}. \tag{17}$$

From equation (17), we see that when both *n* and $\hat{\delta}$ are small (i.e., $\hat{\tau}_{00}$ is small relative to $\hat{\sigma}^2$), the adjustment is primarily dominated by FPC$_1$, whereas when *n* is large, the adjustment is primarily dominated by FPC$_2$.

Real Data Example

The potential need for FPCs may be better understood through an example using real data. Here we use a subset of the data from the World Values Survey 1990–1993 (World Values Study Group, 1994). There were 43 countries in the original data set; however, because the missing data rates for some countries were high, we included only 51,673 participants from 38 countries for illustrative purposes. The hypothetical research question was whether an individual's life satisfaction (on a 10-point scale with higher scores reflecting higher satisfaction with life) could be predicted by individual-level financial satisfaction (on a 10-point scale with higher scores reflecting higher satisfaction with one's financial situation) and a country-level human rights index. Financial satisfaction was grand-mean centered in the analyses. Human rights indexes for the period 1990–1993 were obtained from Gupta, Jongman, and Schmid (1994), which combined gross human rights violation, political right violation and civil rights violation, with higher scores representing a lower level of human rights in a given country (ranging between 13.39 for the United States and 32.36 for China). As the unit of human rights index was arbitrary, we standardized it before the HLM analysis.

We used the R package `lme4` (Bates, Maechler, Bolker, & Walker, 2014) to analyze a two-level model with a random intercept, along with the Satterthwaite's approximation of the degrees of freedom for the *t* tests for the fixed effect estimates using the R package `lmerTest` (Kuznetsova, Brockhoff, & Christensen, 2016). Using restricted maximum likelihood estimation, the fixed-effect estimate (and the corresponding *SE*s) for financial satisfaction was 0.365 (*SE* = 0.0036), 95% confidence interval [CI] [0.358, 0.372], *t*(47,008.4) = 98.97, *p* < .001, and the estimated effect for human rights was −0.247 (*SE* = 0.089), 95% CI [−0.429, −0.066], *t*(32.12) = −2.68, *p* = .012. Using the conventional significance test with a .05 significance level, it could be concluded that there was evidence of the negative effect of a lack of human rights on life satisfaction, as the 95% symmetric confidence interval did not include zero.

The previous analysis assumed that the 38 countries in the study were sampled from an infinite population. However, it is probably more reasonable to think that the level-2 population, which included all countries, was finite. To be conservative, we used 200 as the population size, so

$$\text{FPC}_2 = \frac{200 - 38}{200} = 0.81 .$$

On the other hand, given that the level-1 sample size was only a small subset of the world's population, it would hardly make any difference on the *SE*s if we applied FPC_1 to σ^2. From the HLM analyses, we obtained $\hat{\tau}_{00}$ = 0.289 and $\hat{\sigma}^2$ = 3.898. Using the R code provided in

Appendix B, we obtained the adjusted *SE* for human rights as 0.080, which corresponds to a 9.9% reduction compared to sampling from an infinite population. The adjusted test statistic was $t(32.12) = -2.97$, $p = .006$, 95% CI [−0.411, −0.084]. Therefore, if the target of generalization was all existing countries, failure to incorporate the population size could lead to an *SE* overestimated by $(1 / 0.901 - 1) \times 100\% = 11\%$, resulting in a confidence interval that was too wide and an inflated Type II error rate. By incorporating realistic information about the target population for generalization, one can obtain more accurate statistical inference for a finite population.

Besides cross-cultural research, finite population sampling can also be an issue in organizational research when the sampled organizations represent a large portion of the target population. For example, Mani et al. (2007) studied entry mode and equity level for 4,459 Japanese subsidiaries nested within 858 firms—the authors suggested that the sample represented 40% of the total number of Japanese subsidiaries. The population size at the firm level was not reported, but for illustration purpose we assume that the 858 firms also represent 40% of the total number of Japanese firms. If we substitute $\text{FPC}_1 = .60$ and $\text{FPC}_2 = .60$ in equations (12) and (17), we see that the *SE*s can be overestimated by 29% for both level-1 and level-2 fixed effects, if the targets of generalization are indeed the finite populations' subsidiaries and firms. As long as the target of generalization at level 2 is finite and enumerable, we can similarly apply the adjustment procedures to educational research (e.g., having a large representative sample of schools in a state or a country, or a representative sample of all accredited higher education programs, as in Hatcher, Wise, & Grus, 2015); health research (e.g., Yoon et al., 2012, studied electronic health records in a sample that accounted for 10% of all hospitals in Korea); and other areas in social sciences.

Monte Carlo Simulations

The aim of the Monte Carlo simulation study is mainly to examine how the proposed correction performs in data sampled from finite populations under conditions with unbalanced cluster sizes (i.e., $n_j \neq \bar{n}$ for at least some *j*). Two data-generating models were used in the present study: a random intercept model and a random slope model.

Data-Generating Models

Random intercept model. As shown in Figure 1a, the first data-generating model had one outcome, *Y*, two level-2 predictors, W_1 and W_2, and one level-1 predictor, *X*, with the intraclass correlation (ICC; i.e., the ratio of level-2 variance component to the sum of level-1 and level-2 variance components) of *X* being 0 in the population. All three predictors had only fixed effects on *Y*, and we had the mixed-model equation

$$y_{ij} = \gamma_{01}W_{1j} + \gamma_{02}W_{2j} + \gamma_{10}X_{ij} + u_{0j} + \varepsilon_{ij}. \tag{18}$$

The grand intercept γ_{00} was set to zero without loss of generality. Both W_1 and W_2 were normally distributed with a mean of 0 and a variance of 1 for simplicity, and the correlation between W_1 and W_2 was set as $r_{W_1W_2}$ = .5.[2] The fixed effects γ_{01} and γ_{02} were set as 0.2 and 0.45 to represent small and medium effects, respectively. Each of the random effects u_{0j} and ε_{ij} followed a normal distribution with a mean of 0. We fixed Var(u_{0j}) = τ_{00} to 1, and Var(ε_{ij}) = σ^2 was determined based on the ICC of *y*, which equals τ_{00} / (τ_{00} + σ^2), as described later. At the population level, the variance of *Y* explained by the two level-2 predictors was $0.2^2 + 0.45^2 + 2(0.2)(0.45)r_{W_1W_2} = 0.3325$; thus, the proportion of explained variance (R^2) at level 2 was 0.3325 / (1 + 0.3325) = 133 / 533 ≈ .250. At level 1, *X* was normally distributed with a mean equal to 2 and a variance equal to 1; we kept the level-1 R^2 equal to that of level 2; therefore, γ_{10} was fixed to $\sqrt{0.3325}\sigma \approx .577\sigma$.

Random slope model. The random slope model had the same form as the random intercept model except that the effect of *X* on *Y* varied across clusters. Specifically, the mixed-model equation became (see Figure 1b)

$$y_{ij} = \gamma_{01}W_{1j} + \gamma_{02}W_{2j} + (\gamma_{10} + u_{1j})X_{ij} + u_{0j} + \varepsilon_{ij}\ , \qquad (19)$$

where the coefficient for X_{ij} had a group-specific component, $u_{1j,}$ in addition to the mean effect, γ_{10}. In this model, u_{1j} followed a normal distribution with mean equal to 0 and variance of τ_{11} = 0.5. The random effect for the intercept, u_{0j}, and that for the slope, u_{1j}, was bivariate normal with a covariance of τ_{01} = 0.25. All other parameters were identical to the counterparts in the random intercept model.

Simulation Conditions

To better relate to previous studies, we referred to Bauer and Sterba (2011), Bliese (1998), Maas and Hox (2005), and LaHuis, Hartman, Hakoyama, and Clark (2014) in choosing our simulation conditions. Specifically, we manipulated four design factors: sample-population size ratio in level 2 ($P = J / J_{\text{pop}}$), intraclass correlation (ICC), number of clusters (*J*), and average cluster size (*n*). As previous guidelines for single-level studies have suggested that FPC is needed when the sample-population size ratio is equal to or larger than .05 (i.e., $P \geq .05$), we chose *P* = .05, .10, .25, and .50 in our simulation. Based on the review by Hedges and Hedberg (2007), we set the levels of ICC at .05, .20, and .35, which cover the majority of multilevel data structures in the social sciences. The number of clusters was 20, 30, 50, or 100, which covers the ranges used in previous simulation studies. The average cluster size was 5, 10, 25, which was similar to the conditions in Bauer and Sterba and Maas and Hox. Therefore, there were a total of

4 × 3 × 4 × 3 = 144 conditions in the simulation study for each data-generating model (i.e., random intercept or random slope). The simulation conditions are summarized in Table 2. All simulations were conducted on Oakley Cluster at the Ohio Supercomputer Center (Ohio Supercomputer Center, 1987)

Data Generation

All data were generated in R 3.2.3 (R Core Team, 2015) and analyzed using the package `lme4` (Bates et al., 2014). For each condition, we randomly generated 500 finite populations of level-2 variables, each of size $J_{\text{pop}} = J / P$. This ensured that the results did not capitalize on the characteristics of a single population. For the random intercept model, in each finite population we generated W_1, W_2, and u_{0j} from a multivariate normal distribution with a mean of 0 for all three variables and forced the generated population to have the exact covariance matrix

$$\begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

so that the model is valid for each population.

As it is more common and more justifiable to make inference to a finite level-2 population, to simplify the simulation we assumed that the level-1 variables were sampled from an infinite population. There were 500 replications for each of the 500 finite populations. Thus, a total of 500 × 500 = 250,000 data sets were generated for each simulation condition. In each replication, a level-2 sample (W_1, W_2, u_{0j}) of size J was drawn without replacement from the generated finite population of size J_{pop}. To simulate unbalanced cluster sizes, for a given number of clusters, there were five groups of cluster sizes, each with J / 5 clusters; the cluster sizes were $\bar{n}/5$, $3\bar{n}/5$, $\bar{n}$, $7\bar{n}/5$, and $9\bar{n}/5$, respectively. The level-1 predictor X and the level-1 random effect ε_{ij} were generated as previously specified, and the response variable y was computed according to equation (18) for the random intercept model with the specified parameter values for a given condition.

The procedure for generating data for the random slope model was identical except that the level-2 variables were (W_1, W_2, u_{0j}, u_{1j}) with an exact covariance matrix of

$$\begin{bmatrix} 1 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.25 \\ 0 & 0 & 0.25 & 0.5 \end{bmatrix},$$

and y was computed according to equation (19).

Evaluation Criteria

For each sample, we fitted the simulated data with the same data-generating model using `lme4` first with grand-mean centering and then with group-mean centering; for both centering methods, we obtained the estimated fixed effects at level 1, $\hat{\gamma}_{10}$, and at level 2, $\hat{\gamma}_{01}$ and $\hat{\gamma}_{02}$, and the corresponding *SE*s, using restricted maximum likelihood (REML) estimation. To obtain more robust results, we used the `pbkrtest` package (Halekoh & Højsgaard, 2014) in R to obtain the *SE*s with the Kenward-Roger correction (Kenward & Roger, 1997). We then computed the adjusted standard errors, SE^{FP}, for the three $\hat{\boldsymbol{\gamma}}$ s. We denote the unadjusted standard error as SE_0 to distinguish it from the adjusted standard error, SE^{FP}. To evaluate SE_0 and SE^{FP}, for the *j*th finite population (j = 1, . . ., 500), we estimated their relative biases by comparing them with the empirical standard errors, $SD_j(\hat{\boldsymbol{\gamma}})$ of the three $\hat{\boldsymbol{\gamma}}$ s across the 500 replications, such that the

$$\text{Relative Bias}\left[SE_j(\hat{\gamma})\right] = \frac{\sum_i SE_j(\hat{\gamma}_i)/R - SD_j(\hat{\boldsymbol{\gamma}})}{SD_j(\hat{\boldsymbol{\gamma}})},$$

where R = 500 is the number of replications for each population. For each simulation condition, we averaged the relative *SE* biases across the 500 finite populations.

Given the large number of effects that can impact relative biases in this study, including the main effects of and the interactions among P, J, $\bar{n}$, ICC, grand-mean vs. group-mean centering, and W_1 vs. W_2 for level-2 fixed effects, we conducted eight ANOVAs, one for each combination of models (random intercept vs. random slope), levels (level-2 vs. level-1 fixed effects), and type of *SE*s (SE_0 vs. SE^{FP}), using relative *SE* bias as the criterion variable. From the analyses we computed the η^2 effect size. Our summary of the results mainly focuses on the effects associated with the largest η^2s. The ANOVA results may be found in Tables 3 to 10 in the supplemental material. Furthermore, to compare the performance of the *SE*s under an infinite level-2 population and a finite level-2 population, we also obtained the relative *SE* bias for the unadjusted *SE* from 10,000 replications where the data were sampled from an infinite level-2 population. We denote this baseline *SE* as SE^{SP}.

In addition to comparing the relative biases of SE_0 and SE^{FP}, we also looked at the root mean squared error (RMSE) of the two *SE* estimators, which also takes into account the sampling variability. For each population, we first computed the mean squared error (MSE) for SE_0 and SE^{FP}:

$$\text{MSE}\left[SE_j(\hat{\gamma})\right] = \sum_i [SE_j(\hat{\gamma}_i) - SD_j(\hat{\boldsymbol{\gamma}})]^2 / R\,.$$

The RMSE for each simulation condition was obtained by averaging the MSEs for the 500 populations and taking the square root.

Results

Given the large number of simulation conditions, for the interpretation and the graphs in Figures 2 to 5, we aggregated the results over design factors with negligible η^2 effect sizes, as explained below. Detailed tables of the percentage relative biases for each condition may be found in Table 11 to Table 18 in the supplemental materials.

Random intercept model.

Level-2 fixed effects. The results for the level-2 fixed effect in the random intercept model are shown in Figure 2, with boxplots for SE^{FP} and SE_0 and SE^{SP} represented by dashed lines for comparison. As the main and higher-order interaction effects involving centering and the difference between $\hat{\gamma}_{01}$ and $\hat{\gamma}_{02}$ had η^2 smaller than .004 for both SE^{FP} and SE_0, the difference between the relative biases for $SE(\hat{\gamma}_{01})$ and for $SE(\hat{\gamma}_{02})$ and the difference between grand-mean and group-mean centering were negligible in all conditions, so their results were aggregated. The effects most strongly associated with variability in relative bias for SE_0 were P ($\eta^2 = .09$), ICC ($\eta^2 = .01$), and P × ICC interaction ($\eta^2 = .01$).

As shown in Figure 2, with a finite population, the unadjusted SE_0 of the level-2 fixed effects became increasingly more positively biased (i.e., overestimated) when the level-2 sample accounted for a larger portion of the level-2 population (i.e., larger P). The relative biases for SE_0 were all between −2.1% to 1.7% when $P = .05$, and between −1.3% to 4.4% when $P = .10$, so FPC was not needed when $P \leq .10$. However, the bias became non-ignorable and exceeded 10% in some conditions with $P = .25$ and .50. The degree of bias also increased with a larger $\bar{n}$ ($\eta^2 = .007$), a larger J ($\eta^2 = .003$), or a larger ICC ($\eta^2 = .014$). When $P = .25$, the relative biases for SE_0 were beyond 10% for the following conditions: ICC = .20, $\bar{n} = 25$, $J \geq 50$, with SE biases of 10.0% to 10.8%; ICC = .35, $J \geq 30$, and $\bar{n} = 25$, with SE biases of 10.2% to 12.4%; ICC = .35, $J = 100$, $\bar{n} = 10$, with SE bias of 10.6%. On the other hand, when $P = .50$, the relative bias for SE_0 was beyond 10% except for a few conditions with $\bar{n} \leq 10$, and ICC = .05, and with $\bar{n}$ = 5, J = 20, and ICC = .20; otherwise, SE_0 was substantially and positively biased by 10.2% to 32.5%, and the degree of biases increased with a higher ICC, a higher $\bar{n}$, or a higher J.

Using the suggested adjustment helped remove most of the *SE* bias on the level-2 fixed effects. The effect most strongly associated with the variability in relative bias for SE^{FP} was *P* (η^2 = .02). Figure 2 shows that the relative biases of SE^{FP} were close to those of SE^{SP}, the baseline, although there were some underestimations for SE^{SP}, especially with a larger *P*. Although the formulae tended to overcorrect the *SE*s and resulted in slightly negatively biased *SE*s in situations with a small *J*, small *n*, and small ICC, the relative bias for SE^{FP} was within 10% for all conditions and was at most −5.3% with ICC ≥ .20, $J \geq 30$, and $n \geq 10$. The RMSE results also showed that SE^{FP} better estimated the *SE* of level-2 fixed coefficients, with the ratio of RMSE(SE_0) to RMSE(SE^{FP}) between 1.01 and 4.16, meaning that in each condition the SE^{FP} estimates were closer to the empirical *SE*s than the SE_0 estimates.

Level-1 fixed effects. For $SE(\hat{\gamma}_{10})$ of the corresponding level-1 fixed effect, SE_0, SE^{FP}, and SE^{SP} all showed virtually no bias (Figure 3), with relative biases between −0.5% to 0.5% for SE_0 and between −1.2% to 0.4% for SE^{FP}. The ratio of RMSE(SE_0) to RMSE(SE^{FP}) was between 0.98 and 1.00, indicating that the two standard error estimators performed very similarly.

Random slope model.

Level-2 fixed effects. The results for the level-2 fixed effect in the random intercept model are shown in Figure 4. As the main and higher-order interaction effects involving centering and the difference between $\hat{\gamma}_{01}$ and $\hat{\gamma}_{02}$ all had η^2 smaller than .005, the difference between the relative biases for $SE(\hat{\gamma}_{01})$ and for $SE(\hat{\gamma}_{02})$ and the difference between grand-mean and group-mean centering were negligible in all conditions, so their results were aggregated. The effect most strongly associated with variability in relative bias was *P* (η^2 = .04 for SE_0, η^2

= .02 for SE^{FP}). As illustrated in Figure 4, in general, the SE^{FP} showed a similar performance to the baseline, SE^{SP}, although there were some underestimations for SE^{SP}, especially with a large P, a large ICC, and a small J. On the other hand, SE_0 were positively biased when P increased. Similar to the random intercept model, when P = .05 or .10, the need for FPC was minimal, as the relative biases of SE_0 were within the acceptable range—between −3.9% and 2.5%, whereas SE^{FP} had relative biases between −6.0% to 0.05% (compared to −6.2% to 0.8% for SE^{SP}). The relative bias of SE_0 were within acceptable range and between 2.0% to 7.9% when P = .25. When P = .50, SE_0 was substantially and positively biased; the relative biases ranged between 9.3% to 19.9%, and the degree of biases increased with a higher ICC, a higher $\bar{n}$, or a higher J.

For P =.50, using the suggested adjustment helped remove most of the SE bias on the level-2 fixed effects, but the formulae tended to overcorrect the SEs and resulted in negatively biased SEs in situations with a small J and a high ICC, as shown in Figure 4. However, the underestimation never went below −10% (ranging between −8.1% and −9.4% for J = 20 and ICC = .05), and the overcorrection was partly related to the observation that, under conditions with a small J, a small n, and a high ICC, the usual SE^{SP} also tended to underestimate the true variability of sampling from an infinite level-2 population in the random slope model, as shown in the dashed lines in Figure 4. On the other hand, with $J \geq 30$ and ICC $\leq$ .35, the relative bias for SE^{FP} was between −7.0% and −4.5%. The RMSE results also showed that SE^{FP} better estimated the SE of level-2 fixed coefficients, with the ratio of RMSE(SE_0) to RMSE(SE^{FP}) between 1.01 and 2.43, meaning that in all conditions the SE^{FP} estimates were closer to the empirical SEs than the SE_0 estimates.[3]

Level-1 fixed effects. Unlike in the random intercept model where there was virtually no bias on $SE(\hat{\gamma}_{10})$ and $SE(\hat{\gamma}_{02})$ for SE_0, SE^{FP} and SE^{SP}, there was substantial bias for SE_0 in the random slope model for the level-1 fixed effect when finite population was ignored (see Figure 5), with a pattern similar to that of the level-2 fixed effect in the random slope model but with biases of a bigger magnitude. The effect most strongly associated with the variability in relative bias of SE_0 was P ($\eta^2 = .04$). For the condition with a small sample size ($J = 20$ and $\bar{n} = 5$) and low ICC (ICC = .05), SE_0, SE^{FP} and SE^{SP} all tended to overestimate the empirical SE of the level-1 fixed effect, with relative biases between 9.0% and 10.1% when $P \leq .10$. This positive bias could be related to the Kenward-Roger correction. The need for FPCs was small when (a) $P \leq .10$ (relative biases for SE_0 between 0.6% and 9.9%); (b) $P = .25$, ICC $\leq .20$ or $\bar{n} \leq 10$ (relative biases for SE_0 between 4.2% and 10.3%); and (c) $P = .50$, ICC = .05, $J \geq 30$, and $\bar{n} \leq 10$ (relative biases for SE_0 between 5.5% and 9.5%). Otherwise, the degree of biases in SE_0 increased with a higher ICC and a higher $\bar{n}$, and was between 28.6% and 30.0% when $P = .50$, ICC = .35, and $\bar{n} = 25$.

The adjusted SE, SE^{FP}, performed well for estimating $SE(\hat{\gamma}_{10})$. The effect most strongly associated with variability in relative bias of SE^{FP} was ICC ($\eta^2 = .03$). Aside from the extreme condition with a small sample size and a low ICC where SE_0, SE^{FP} and SE^{SP} all tended to show positive biases, when either $J \geq 30$ or $\bar{n} \geq 10$, the relative biases were between −3.2% and 6.7%, within acceptable ranges for all conditions, as shown in Figure 5. The RMSE results also showed that SE^{FP} better estimated the SE of level-2 fixed coefficients, with the ratio of RMSE(SE_0) to RMSE(SE^{FP}) between 1.02 and 5.02, meaning that in each condition the SE^{FP} estimates were closer to the empirical SEs than the SE_0 estimates.

Discussion

The simulation results indicate that when the assumption of an infinite population at level 2 is violated, the SEs in HLM without the adjustment are generally positively biased. Under conditions specified in our simulation study, the unadjusted *SE*s by the conventional analysis became more positively biased and could be as high as 30% as the ICC increased and as the level-2 sample size approached the level-2 population size (i.e., when *P* increased). For models without random slopes, utilizing the proposed adjustment can largely reduce the bias in the *SE*s of level-2 fixed effects, especially when *P* is large. For models with random slopes, biases were found for both level-1 and level-2 fixed effects. Again, applying the proposed adjustment produced acceptable *SE*s.

Our simulations showed that analyzing multilevel data with regular multilevel modeling software, which in general assumes an infinite population, resulted in positively biased *SE*s for level-2 fixed effects in the random intercept model and for both level-2 and level-1 fixed effects in the random slope model in the majority of the conditions. As expected, application of finite population correction factors did not affect *SE*s for level-1 fixed effects in the random intercept model, as those *SE*s were not functions of the level-2 variance components; however, finite population at level 2 did affect *SE*s at level 1 in the random slope model, as those *SE*s were functions of the random slope variance at level 2 (Snijders, 2005). The unadjusted *SE*s were generally acceptable when the number of level-2 units was less than 10% of the level-2 population size, but were overestimated by more than 10% when the level-2 units corresponded to 25% or more of the population. From equation (17), if one assumes that the average cluster size is large relative to the inverse of the ICC and $\text{FPC}_1 = 1$, the bias of the unadjusted *SE* will be within 10% when $\text{FPC}_2 \geq .826$, and the bias will be within 5% when $\text{FPC}_2 \geq .907$. Therefore, we recommend that researchers pay specific attention *when the level-2 sample size accounts for more than 17% of the level-2 population size*. If more accurate estimation of the *SE*s is needed (i.e., relative bias < 5%), researchers should adjust for finite population sampling *when the level-*

2 sample size accounts for more than 9% of the level-2 population size. Although these recommendations are based on the formula for models using only a random intercept, our simulation results showed that they should be applicable to models with random slopes, too.

In contrast, our results showed that with small cluster size and number of clusters coupled with unbalanced cluster sizes and a large ICC, the *SE*s produced in multilevel models with adjustment were biased downward, albeit the downward bias was also present in the normal use of HLM assuming a superpopulation model, as evidenced in our simulation results. The downward bias was also observed in previous simulation studies (e.g., Maas & Hox, 2004; McNeish & Stapleton, 2016b), which indicates that the asymptotic *SE*s obtained under maximum likelihood or restricted maximum likelihood may not work well for 20 or fewer clusters with unbalanced cluster sizes and a relatively large ICC.

The problem of downward *SE* biases will be alleviated with the use of *t*-tests and *t*-based confidence intervals, as the *SE* was only the scale parameter for the *t* distribution and the true standard deviation of the *t* distribution equals the scale parameter times $\sqrt{\nu/(\nu-2)}$ for $\nu > 2$, where ν is the degrees of freedom for the *t* distribution. Therefore, the *SE* was an underestimate of the true sampling variability, especially when ν is small, even when the test maintained the nominal Type I error rate. To verify this, for each replication in the simulation conditions with the random slope model, $J \leq 30$, $\bar{n} \leq 10$, and ICC = .35, we constructed 95% confidence intervals as $\hat{\gamma} \pm SE \times t_{.975}(\nu)$ with both the adjusted and unadjusted *SE*s, where the degrees of freedom ν was obtained by the Kenward-Roger approximation (Kenward & Roger, 1997). We then obtained the empirical coverage rate as the proportion of replications where the confidence intervals contained the true population value. Whereas the coverage rate for some conditions using the adjusted *SE* was below 95%, the lowest was around 93.6% with $P = .50$, $\bar{n} \leq 10$, and

ICC = .35, which is still acceptable using Bradley's (1978) criterion for liberal test (empirical Type I error rate ≤ .075 corresponding to a coverage rate ≥ 92.5%). All other conditions showed a coverage rate between 94.0% and 96.2% with the adjusted *SE* (see Tables 23 and 24 in the supplemental material, as well as the coverage rates for 80% and 90% confidence in Tables 19 to 22).

To opt for the conservative side, we suggest that our proposed adjustment only be applied when the number of clusters is at least 30 with 10 or more observations in each cluster, on average. For cross-cultural studies, for which we believe FPC is most needed, the sample size is usually large and the bias of the proposed adjustment, therefore, will be negligible. For studies with smaller sample sizes, resampling techniques such as the bootstrap procedure in multilevel settings (Goldstein, 2011; van der Leeden, Meijer, & Busing, 2008) may be modified to accommodate the finite population and provide more robust standard error estimates. Future study is needed to implement the bootstrap procedure under sampling from finite populations and evaluate its performance against our proposed adjustment and other methods for obtaining standard error estimates in regular HLM.

Although the notion of an infinite population greatly simplifies the calculations for statistical inference for many real-life research topics and allow for the greatest generalizability, researchers should be aware that it may not always be the population of interest. Correction for finite population is particularly necessary for areas such as cross-cultural research and organizational research on a *predefined* set of countries or organizations bounded by location and time. In other applications of HLM, the population at level 2 can be assumed to be infinite in both a theoretical and a practical sense. Therefore, the choice between a finite versus an infinite population, which should be made at the very early stage of a study, is not trivial. As a result, we encourage researchers working with multilevel data to carefully consider the nature of their research questions and their population, evaluate the potential trade-off between precision and generalizability, and carefully decide on their population when planning their studies.

Finite population correction should be used with caution. With the introduction of FPC in multilevel studies, researchers may feel that they have the option of redefining their population to obtain a narrower confidence interval and a smaller statistical significance level. Such a practice should be strictly prohibited, as the target population of generalization should be carefully chosen during the planning stage of a research study. Regarding the data as a random sample from a population cannot be justified if the definition of the population can change. Changing the population after conducting the analyses is no different from forming test hypotheses after seeing the data. Neither is acceptable in scientific research practice.

There are several limitations to this research. First, in our simulations we have only studied grand-mean and group-mean centering. Although the matrix version of our proposed adjustment does not depend on whether the predictors are centered or not, it is possible that in finite samples the choice of centering affects the relative biases due to interaction with the inherent biases in HLM when sample sizes are small and the cluster sizes are not equal. Two other commonly used strategies are (a) not centering the level-1 predictors and (b) using group-mean centering for the level-1 predictors with the group means added as level-2 predictors. As discussed in Kreft et al. (1995), a model with no centering is equivalent to a model with grand-mean centering, so we expect that our results also apply to no centering. With regard to group-mean centering with group means added, we do not expect that our results will change regardless of whether the group means are added or not. Because we generated the level-1 predictor X in our simulation to have intraclass correlations of 0 in the population, the variance of the sample group means of *X* is likely to be very small and solely due to random sampling. However, as the group-mean centering with group means added approach was shown to be superior for testing cross-level interactions (Enders & Tofighi, 2007; Hofmann & Gavin, 1998), further research is needed to evaluate how the proposed adjustment performs with the different centering options in the presence of level-2 variance of level-1 predictors and the presence of cross-level interactions.[4]

Second, in the simulation we only studied conditions with average cluster sizes of 5, 10, and 25, as opposed to bigger cluster sizes that are usually observed in national and cross-cultural surveys. We selected smaller cluster sizes to make the conditions comparable to previous simulation studies on multilevel modeling (e.g., Bauer & Sterba, 2011; Cousineau & Laurencelle, 2016). Both our simulation results and those of Maas and Hox (2004) showed that the impact of cluster sizes on the standard error biases of fixed effects were relatively small compared to other factors such as ICC and number of clusters. We also conducted additional simulations for conditions with P = .50, J = 20, ICC = .35 and with a random slope model, which were shown to give the largest downward bias for the adjusted *SE*s, and manipulated the average cluster sizes to 50 or 100. The relative biases for the adjusted *SE* were −9.3% and −8.2%, respectively, for $\bar{n}$ = 50 and $\bar{n}$ = 100 at level 2, and 1.0% and 1.9% at level 1, which is comparable to biases for conditions with $\bar{n}$ = 5, 10, or 25. As a result, we believe that the adjustment will work well for larger cluster sizes too, especially when the number of clusters is 30 or larger. However, again, further research is needed to verify the performance of the adjusted *SE*s with larger cluster sizes.

Third, in this paper we only discussed the adjustment for the fixed-effect standard errors. Adjustment for the standard errors of the variance components seems more challenging, and we encourage researchers to work out the adjustment in future studies.

Fourth, our proposed correction formula only corresponds to one of many possible adjustment procedures for handling the issue of making inferences on a finite population at a higher level. A more appealing approach may be to incorporate FPC into the likelihood functions for finding the maximum likelihood estimates of the fixed and random components and for conducting adjusted likelihood ratio test.

Fifth, this study only demonstrated how to correct the *SE*s in two-level random intercept and random slope models. Although the procedure for deriving FPC for models with more than two levels or with non-nesting structures (Beretvas, 2011) would be similar to that presented in this paper, extra random effects would add complexity to the closed-form approximation of the standard errors. Derivations and evaluations of FPC for more complex multilevel models may be considered in future studies.

Sixth, the discussions in the present study assumed that the level-2 units had equal selection probabilities, which is usually also the assumption of multilevel modeling. In survey research, the estimation can take into account sampling weights (e.g., Korn & Graubard, 2003; Pfeffermann et al., 1998), and future research can extend the methods used in the present study to incorporate sampling weights.

Finally, although we derived the FPC for both level-1 finite population and level-2 finite population, we only demonstrated and evaluated the use of FPC for level 2, as we believed that finite population for level 2 would be seen more often in multilevel studies. However, in some situations, FPC for level 1 might also be necessary, and future research should address when and how such corrections for level-1 finite population should be applied.

References

Allison, P. D. (2009). *Fixed effects regression models.* Thousand Oaks, CA: Sage.

Armonas, R., Druteikiene, G., & Marcinskas, A. (2010). An integrated model of growth strategy in a global industry: Multilevel approach. *Transformations in Business & Economics, 9*, 77–100.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.1-6. Retrieved from http://CRAN.R-project.org/package=lme4

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

Beretvas, S. N. (2011). Cross-classified and multiple-membership models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 313–334). New York, NY: Routledge.

Binder, D. A., & Roberts, G. R. (2003). Design-based and model-based methods for estimating model parameters. In R. L. Chambers & C. J. Skinner (Eds.), *Analysis of survey data* (pp. 29–48). Chichester, England: Wiley.

Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods, 1*, 355–373. http://dx.doi.org/10.1177/109442819814001

Bliese, P. D., Halverson, R. R., & Schriesheim, C. A. (2002). Benchmarking multilevel methods in leadership. *The Leadership Quarterly, 13*, 3–14. http://doi.org/10.1016/S1048-9843(01)00101-1

Bradley, J. V. (1978). Robustness. *British Journal of Mathematical and Statistical Psychology, 31*, 144–152. http://dx.doi.org/10.1111/j.2044-8317.1978.tb00581.x

Brune, N. E., & Bossert, T. (2009). Building social capital in post-conflict communities: Evidence from Nicaragua. *Social Science and Medicine, 68*, 885–893. http://doi.org/10.1016/j.socscimed.2008.12.024

Cai, T. (2013). Investigation of ways to handle sampling weights for multilevel model analyses. *Sociological Methodology*, *43*, 178–219. http://doi.org/10.1177/0081175012460221

Chromy, J., & Abeyasekera, S. (2005). Statistical analysis of survey data. In United Nations (Ed.), *Household sample surveys in developing and transition countries*. New York, NY:Author.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: Wiley.

Cousineau, D., & Laurencelle, L. (2016). A correction factor for the impact of cluster randomized sampling and its applications. *Psychological Methods*, *21*, 121–135. http://doi.org/10.1037/met0000055

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, *43*, 558–575. http://dx.doi.org/10.1177/0022022112438397

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*, 121–138. http://dx.doi.org/10.1037/1082-989X.12.2.121

Gelman, A. (2005). Analysis of variance—Why it is more important than ever. *The Annals of Statistics*, *33*, 1–53. http://dx.doi.org/10.1214/009053604000001048

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, England: Cambridge University.

Goldstein, H. (2011). Bootstrapping in multilevel models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 163–171). New York, NY: Routledge.

Green, B. F., Jr., & Tukey, J. W. (1960). Complex analyses of variance: General problems. *Psychometrika*, *25*, 127–152. http://dx.doi.org/10.1007/BF02288577

Gupta, D. K., Jongman, A. J., & Schmid, A. P. (1994). Creating a composite index for assessing country performance in the field of human rights: Proposal for a new methodology. *Human Rights Quarterly, 16*, 131–162. http://dx.doi.org/10.2307/762414

Hair, J. F., Jr., Bush, R. P., & Ortinau, D. J. (2000). *Marketing research: A practical approach for the new millennium* (3rd ed.). Boston, MA: McGraw-Hill.

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models—The R package `pbkrtest`. *Journal of Statistical Software, 59*, 1–30. Retrieved from http://www.jstatsoft.org/v59/i09/

Hatcher, R. L., Wise, E. H., & Grus, C. L. (2015). Preparation for practicum in professional psychology: A survey of training directors. *Training and Education in Professional Psychology, 9*, 5–12. http://dx.doi.org/10.1037/tep0000060

Hatley, H. O., & Sielken, R. L., Jr. (1975). A “super-population viewpoint” for finite population sampling. *Biometrics, 31*, 411–422. http://dx.doi.org/10.2307/2529429

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

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24*, 623–641. http://dx.doi.org/10.1177/014920639802400504

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Jäckle, S., & Wenzelburger, G. (2015). Religion, religiosity, and the attitudes toward homosexuality—A multilevel analysis of 79 countries. *Journal of Homosexuality, 62,* 207–241. http://dx.doi.org/10.1080/00918369.2014.969071

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*, 983–997. http://dx.doi.org/10.2307/2533558

Korn, E. L., & Graubard, B. I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society, Series B, 65*, 175–190. http://dx.doi.org/10.1111/1467-9868.00379

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30*, 1–21. http://dx.doi.org/10.1207/s15327906mbr3001_1

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2016). lmerTest: Tests in linear mixed effects models. R package version 2.0-32. Retrieved from https://CRAN.R-project.org/package=lmerTest

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods, 17*, 433–451. http://dx.doi.org/10.1177/1094428114541701

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546–556. http://dx.doi.org/10.1198/016214504000000467

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). Boston, MA: Cengage.

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, 58, 127–137. http://dx.doi.org/10.1046/j.0039-0402.2003.00252.x

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86–92. http://dx.doi.org/10.1027/1614-1881.1.3.86

Mani, S., Anita, K. D., & Rindfleisch, A. (2007). Entry mode and equity level: A multilevel examination of foreign direct investment ownership structure. *Strategic Management Journal, 28*, 857–866. http://dx.doi.org/10.1002/smj.611

McNeish, D., & Stapleton, L. M. (2016a). Modeling clustered data with very few clusters. *Multivariate Behavioral Research, 51*, 495–518. http://dx.doi.org/10.1080/00273171.2016.1167008

McNeish, D. M., & Stapleton, L. M. (2016b). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28*, 295–314. http://doi.org/10.1007/s10648-014-9287-x

Monahan, J. F. (2008). *A primer on linear models*. Boca Raton, FL: CRC Press.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25*, 267–316. http://doi.org/10.2307/271070

Nielsen, S. (2009). Why do top management look the way they do? A multilevel exploration of the antecedents of TMT heterogeneity. *Strategic Organization, 7*, 277–305. http://dx.doi.org/10.1177/1476127009340496

Ohio Supercomputing Center. (1987). *Ohio Supercomputer Center*. Retrieved from http://osc.edu/ark:/19495/f5s1ph73

Peretz, H., & Fried, Y. (2012). National cultures, performance appraisal practices, and organizational absenteeism and turnover: A study across 21 countries. *Journal of Applied Psychology*, 97, 448–459. http://dx.doi.org/10.1037/a0026011

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B, 60*, 23–40. http://dx.doi.org/10.1111/1467-9868.00106

R Core Team. (2015). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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

Rockstuhl, T., Ang, S., Dulebohn, J. H., & Shore, L. M. (2012). Leader-member exchange (LMX) and culture: A meta-analysis of correlates of LMX across 23 countries. *Journal of Applied Psychology, 97*, 1097–1130. http://dx.doi.org/10.1037/a0029978

Rudnev, M. (2014). Value adaptation among intra-European migrants: Role of country of birth and country of residence. *Journal of Cross-Cultural Psychology, 45*, 1626–1642. http://dx.doi.org/10.1177/0022022114548482

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York, NY: Springer.

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance component* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Smits, F., & Huijts, T. (2015). Treatment for depression in 63 countries worldwide: Describing and explaining cross-national differences. *Health & Place*, *31*, 1–9. http://dx.doi.org/10.1016/j.healthplace.2014.10.002

Smith, T. M. F. (1994). Sample surveys 1975-1990; An age of reconciliation? *International Statistical Review/Revue Internationale de Statistique, 62*, 5. http://doi.org/10.2307/1403539

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1570–1573). Chichester, England: Wiley.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, 18, 237–259. http://dx.doi.org/10.3102/10769986018003237

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, England: Sage.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 475–502. http://doi.org/10.1207/S15328007SEM0904_2

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. http://dx.doi.org/10.1080/00273170903333574

Thompson, S. K. (2012). *Sampling* (3rd ed.). Hoboken, NJ: Wiley.

Todd, S. Y., Crook, T. R., & Barilla, A. G. (2005). Hierarchical linear modeling of multilevel data. *Journal of Sport Management*, *19*, 387–403.

van der Leeden, R., Meijer, E., & Busing, F. M. T. A. (2008). Resampling multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel nalysis* (pp. 401–433). New York, NY: Springer.

Wilkinson, D. (2007). The multidimensional nature of social cohesion: Psychological sense of community, attraction, and neighboring. *American Journal of Community Psychology, 40*, 214–229. http://doi.org/10.1007/s10464-007-9140-1

World Values Study Group. (1994). *World values survey, 1981-1994 and 1990-1993* [Computer file, ICPSR version]. Ann Arbor, MI: Institute for Social Research.

Wu, H. L., Su, W. C., & Lee, C. Y. (2008). Employee ownership motivation and individual risk-taking behaviour: A crosslevel analysis of Taiwan's privatized enterprises. *International Journal of Human Resources Management*, *19*, 2311–2331. http://dx.doi.org/10.1080/09585190802479546

Yoon, D., Chang, B. C., Kang, S. W., Bae, H., & Park, R. W. (2012). Adoption of electronic health records in Korean tertiary teaching and general hospitals. *International Journal of Medical Informatics*, *81*, 196–203. http://doi.org/10.1016/j.ijmedinf.2011.12.002

Footnotes

[1] The use of a single FPC at level 1 relies on the assumption that the sampling fraction of each cluster is constant, an assumption usually referred to as equal selection probabilities in design-based methods. Under such an assumption it can be shown that FPC at level 1 is always greater than or equal to FPC at level 2, with equality holding only when one selects all available level-1 units in the sampled level-2 clusters. This can be shown as:

$$f_1 = N / N_{\text{pop}} = \left(\sum\nolimits_{j=1}^{J} n_j\right) \Big/ \left(\sum\nolimits_{j=1}^{J_{\text{pop}}} n_{\text{pop}j}\right) \leq \left(\sum\nolimits_{j=1}^{J} n_j\right) \Big/ \left(\sum\nolimits_{j=1}^{J_{\text{pop}}} n_j\right) = f_2,$$

where f_1 is the sampling fraction at level 1 and f_2 is the sampling fraction at level 2, and thus the need for FPC at level-1 is relatively negligible. When the equal selection probabilities assumption is violated, one cannot apply a single FPC at level 1, but has to incorporate the unequal selection probabilities into the model. See Stapleton (2002) for a relevant discussion.

[2] As the effect of predictor correlations on the estimated standard errors is already accounted for in the analyses before applying FPC, we expected that changing the correlation between the two predictors would have no effect on the simulation results. To confirm, we generated data with correlations equaling .20, .50, and .80. We found virtually no difference in the results, so only results with correlation equaling .50 are presented.

[3] As one anonymous reviewer pointed out, the covariance terms among fixed-effect estimators may also be of interest for testing differences in fixed effects. We obtained simulation results on the three covariance terms for the 12 conditions with J = 20, $\bar{n}$ = 5, and a random slope data generating model, where relative *SE* biases were largest in magnitude. After converting the biases to correlation metrics (by dividing the biases by the corresponding standard errors), we found that biases without applying FPC ranged between −.22 to .01, but after

applying FPC they were between −.04 to .05. As it was obtained using the same procedure as the *SE*, the adjustment should work well for the covariance terms.

[4]We ran further simulations with an additional cross-level interaction effect between *X* and W_1 with a fixed effect of 0.10 for the four conditions with $J = 20$, $\bar{n} = 5$, ICC = .05, and a random slope data-generating model, where relative *SE* biases were largest in magnitude in the original simulation. The relative biases were between 6.6% and 9.0% for SE_0 and between 3.2% and 6.3% for SE^{FP}, somewhere in between the relative biases for purely level-1 predictors and for purely level-2 predictors. This was expected as a cross-level interaction variable can be considered a level-1 variable with non-zero level-2 variance. Therefore, we expect our procedure for finite population adjustment to work for cross-level interaction effects, too.

Table 1

Examples of Different Targets of Generalization for Two-Level Data

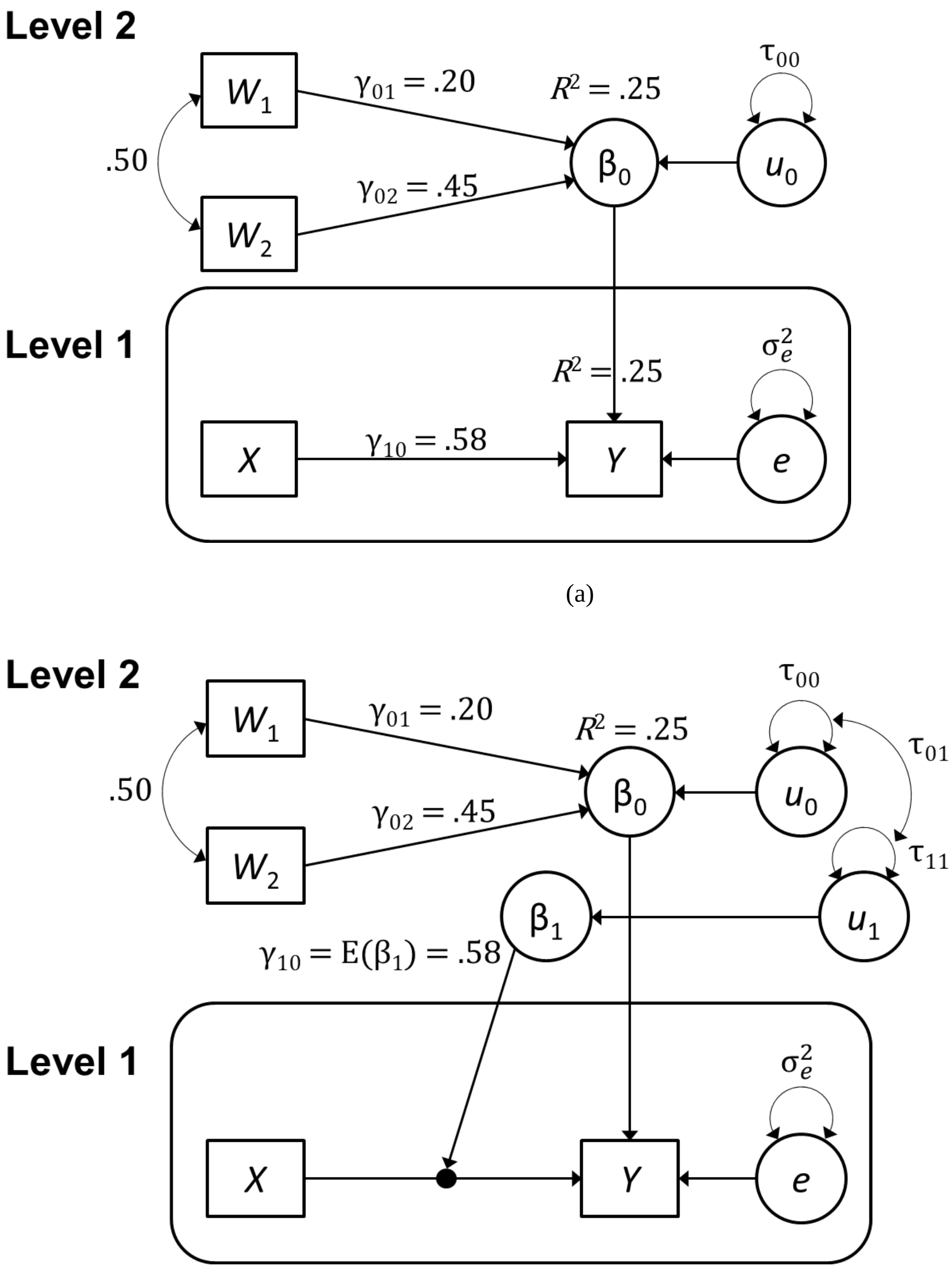
Size of Level-2 Sample	Examples	Targets of Generalization at Level 2	Possible Approaches for Analysis
1. A few level-2 units	Comparing U.S. and Chinese samples	No generalizations	Treating group effects as fixed: Fixed-effect ANOVA/dummy coding/multiple-group analysis
2. 20 or more level-2 units; sample size < 5% of population size	Data from 30 schools in a school district, but generalizing to all schools in the United States	a. No generalizations b. Finite population c. Infinite superpopulation	a. Treating group effects as fixed b & c. Regular HLM as FPC is negligible
3. 20 or more level-2 units; sample size > 5% of population size	Data from 30 countries	a. No generalizations b. Finite population c. Infinite superpopulation	a. Treating group effects as fixed b. HLM with FPC/design-based methods c. HLM
4. All level-2 units in a fixed level-2 population[a]	Data from all 50 states in the United States	a. No generalizations for the level-2 finite population (which is the same as the sample) b. Infinite superpopulation	a. Treating group effects as fixed b. HLM

Note. HLM = hierarchical linear modeling. FPC = finite population correction.
[a]Although treating group effects as fixed is probably more common in practice, for some analyses there are advantages to treating the group effects as a sample from an infinite superpopulation and using HLM. See, for example, chapters 21 and 22 of Gelman and Hill (2006) for the relevant issues.

Table 2

Summary of Design Factors for the Simulation

Design Factors	Manipulated Levels
Sample-population ratio (*P*)	.05, .10, .25, .50
Number of clusters in the sample (*J*)	20, 30, 50, 100
Average cluster size ($\bar{n}$)	5, 10, 25
Intraclass correlation (ICC)	.05, .20, .35
Model	Random intercept model, random slope model



(a)

(b)

Figure 1. Path diagram for the data-generating (a) random intercept model (equation [18], p. 24) and (b) random slope model (equation [19], p. 24) in the simulation study. The data-generating model with random slopes has the same form except that the effect of *X* on *Y* varies across clusters.

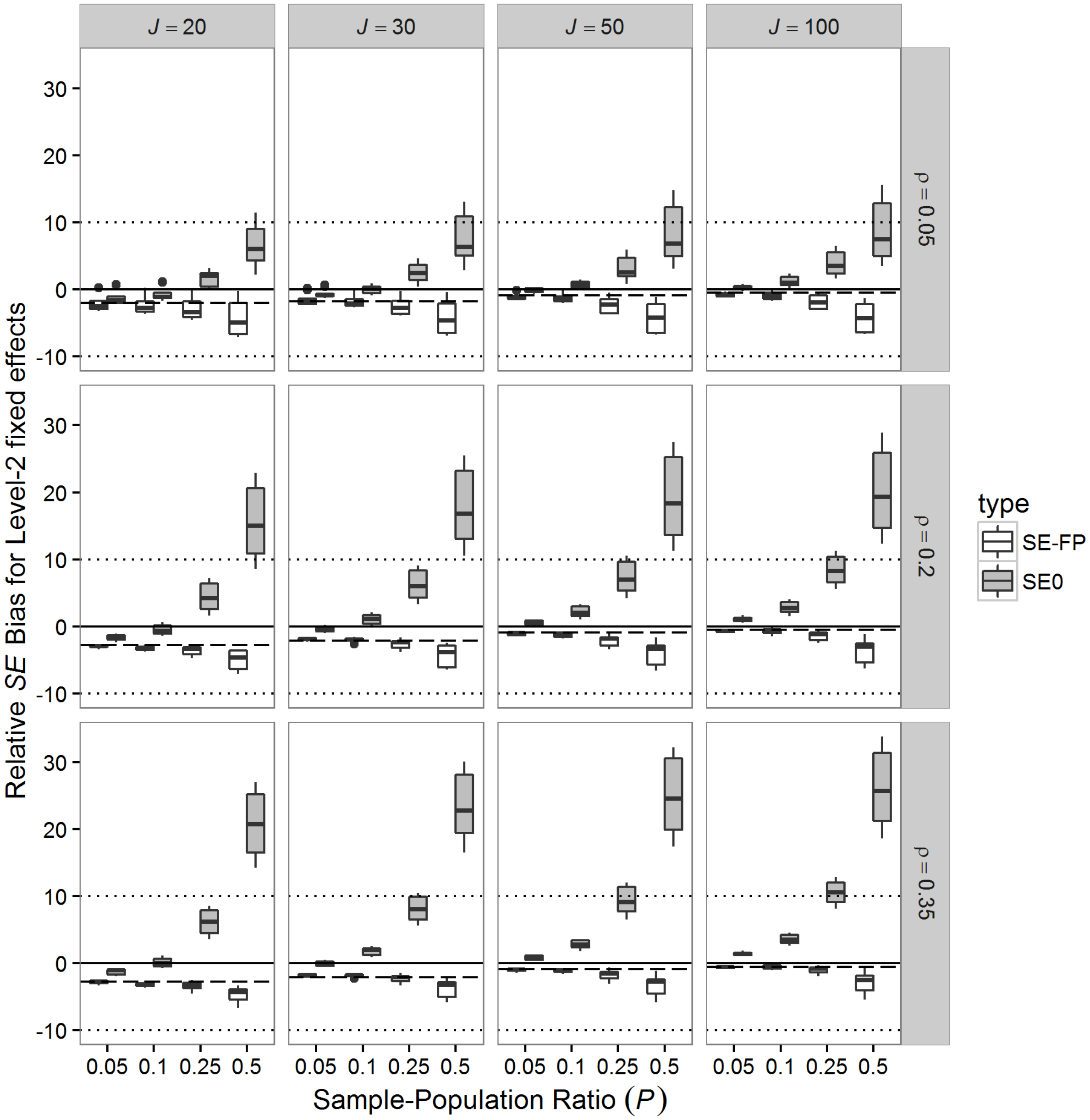


Figure 2. Percentage relative bias in standard errors for level-2 fixed effects in the random intercept model. The dashed line corresponds to the average percentage relative bias for the unadjusted standard errors for data with a level-2 superpopulation (i.e., infinite); that is, SE^{SP}. J = number of clusters. SE-FP = finite population adjusted standard errors for data generated with a finite level-2 population. SE0 = unadjusted standard errors for data generated with a finite level-2 population.

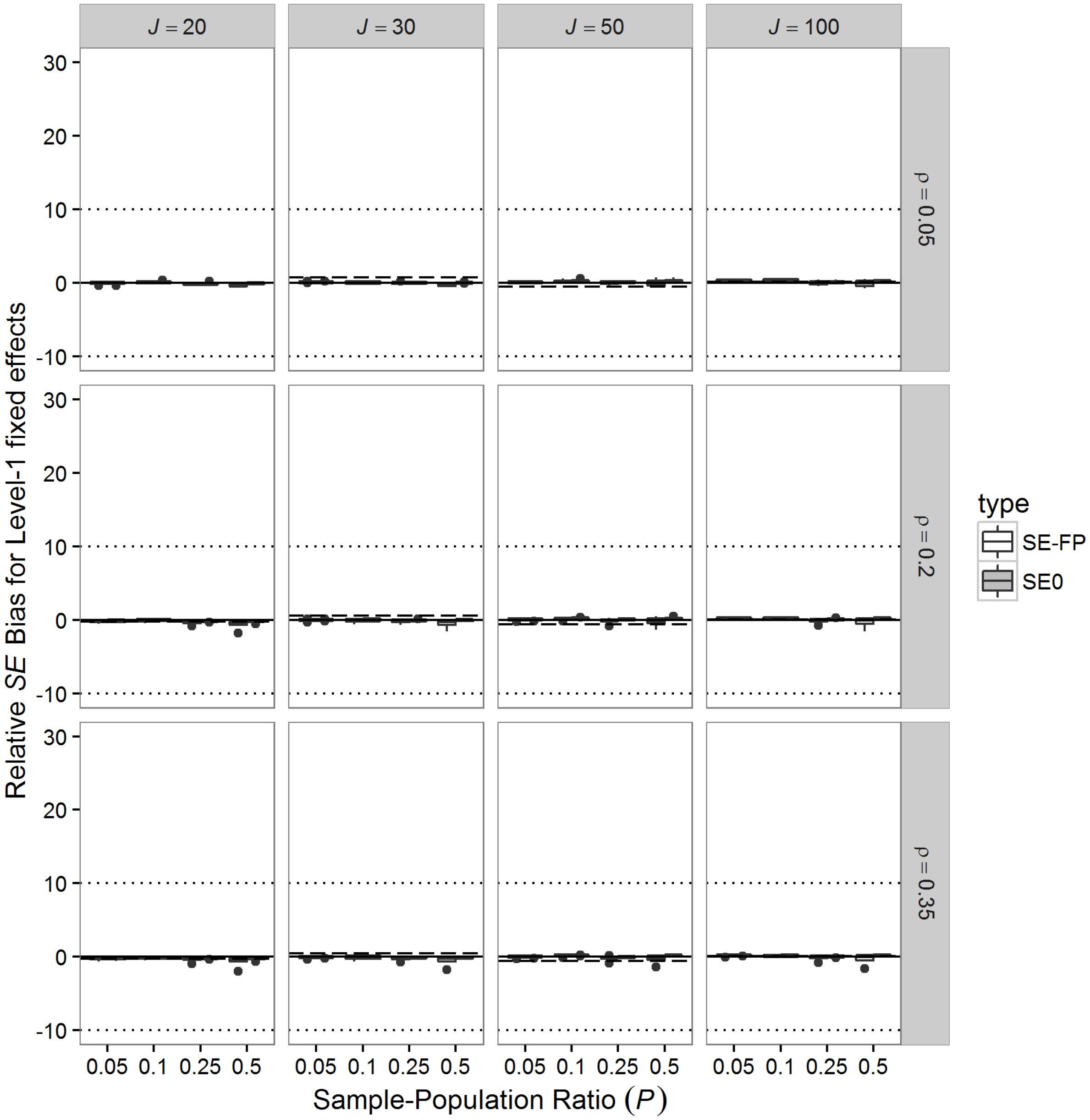


Figure 3. Percentage relative bias in standard errors for level-1 fixed effects in the random intercept model. The dashed line corresponds to the average percentage relative bias for the unadjusted standard errors for data with a level-2 superpopulation (i.e., infinite); that is, SE^{SP}. J = number of clusters. SE-FP = finite population adjusted standard errors for data generated with a finite level-2 population. SE0 = unadjusted standard errors for data generated with a finite level-2 population.

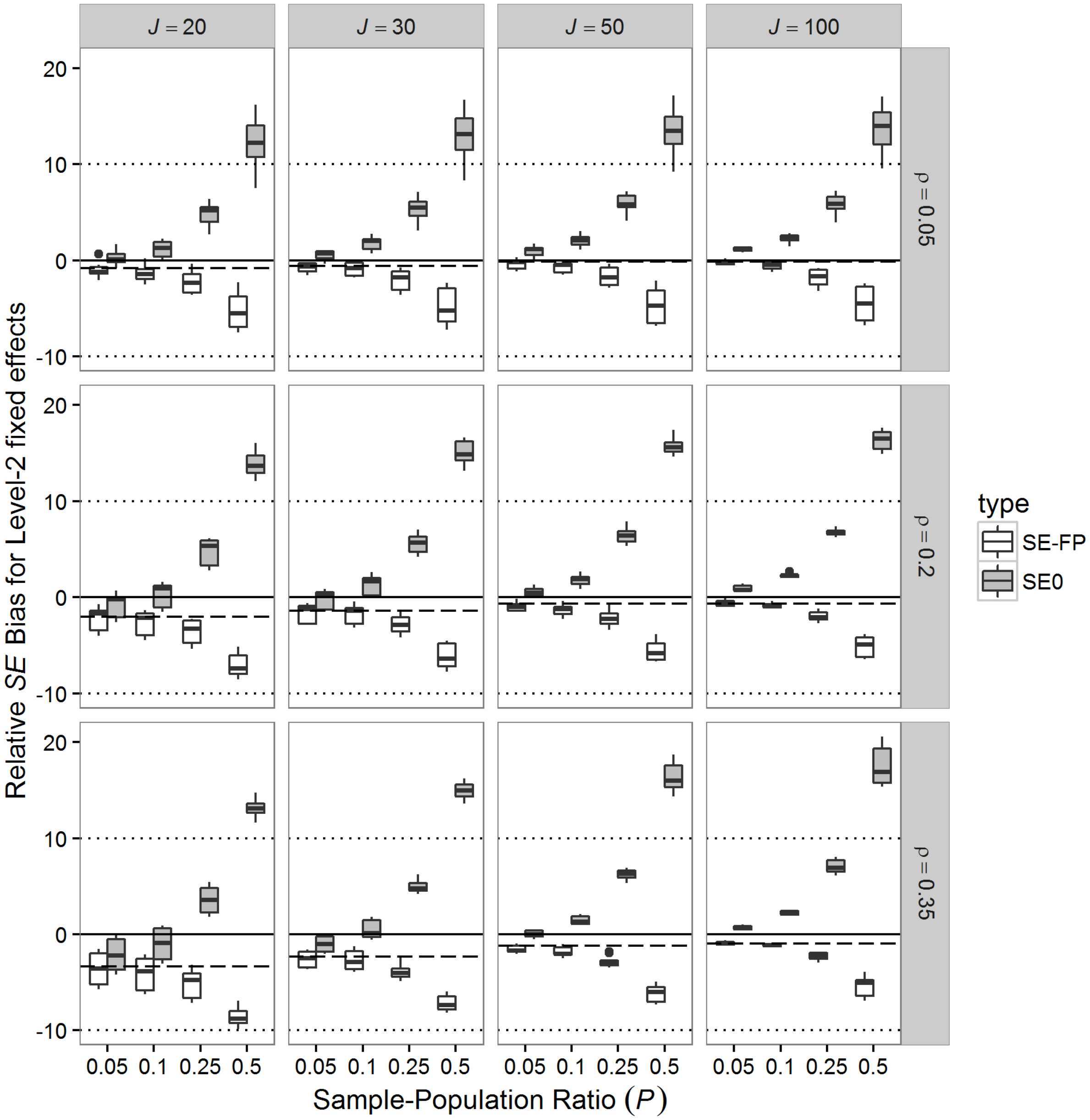


Figure 4. Percentage relative bias in standard errors for level-2 fixed effects in the random slope model. The dashed line corresponds to the average percentage relative bias for the unadjusted standard errors for data coming a level-2 superpopulation (i.e., infinite); that is, SE^{SP}. J = number of clusters. SE-FP = finite population adjusted standard errors for data generated with a finite level-2 population. SE0 = unadjusted standard errors for data generated with a finite level-2 population.

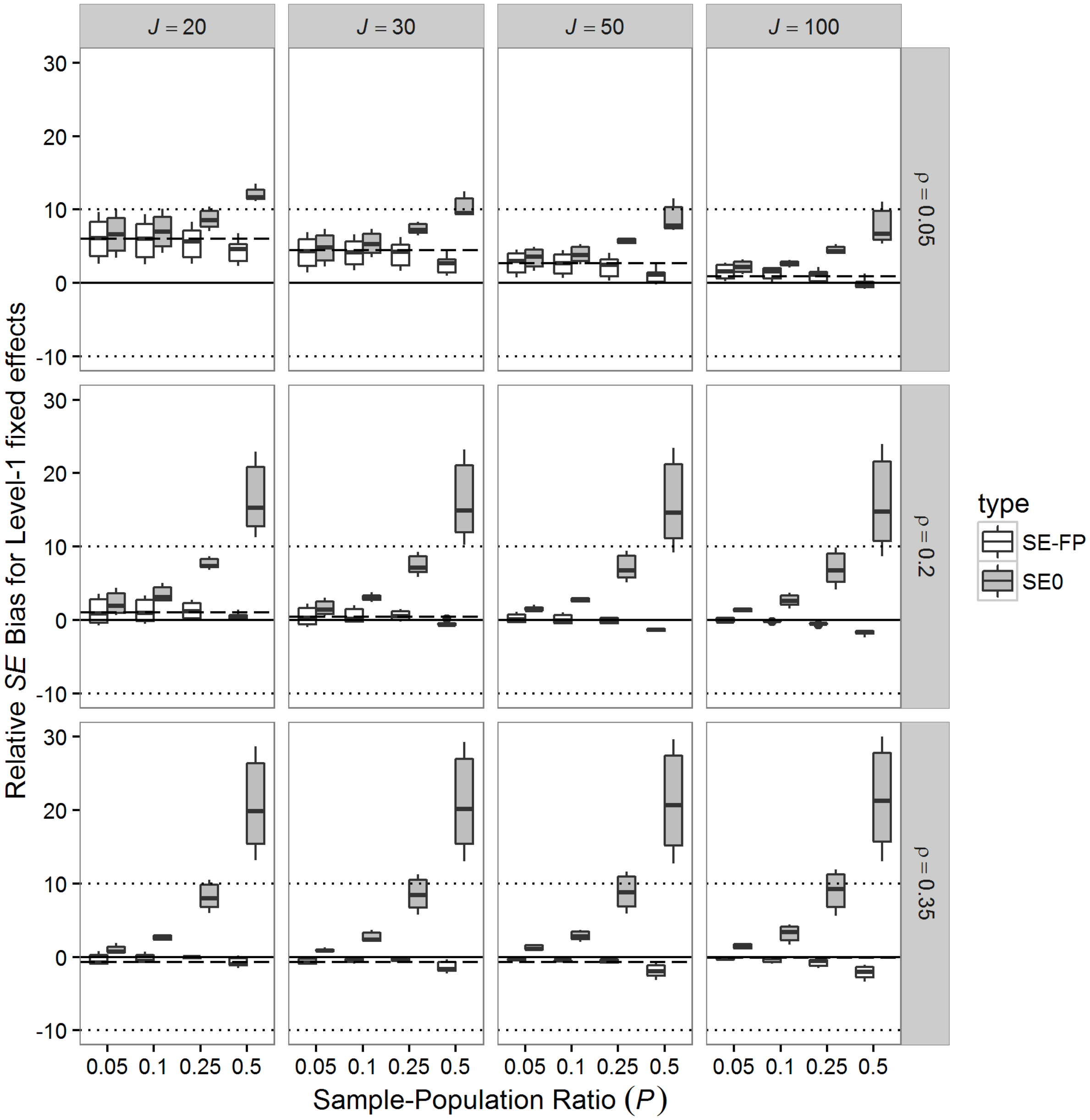


Figure 5. Percentage relative bias in standard errors for level-1 fixed effects in the random slope model. The dashed line corresponds to the average percentage relative bias for the unadjusted standard errors for data with a level-2 superpopulation (i.e., infinite); that is, SE^{SP}. *J* = number of clusters. SE-FP = finite population adjusted standard errors for data generated with a finite level-2 population. SE0 = unadjusted standard errors for data generated with a finite level-2 population.

Appendix A

Derivation of Finite Population Correction for Models With a Random Intercept

To obtain the adjusted *SE* of the fixed effect estimator, we first recognize that from equation (6), $\hat{\boldsymbol{\gamma}}$ is a linear transformation of **y** so that one can write $\hat{\boldsymbol{\gamma}} = \mathbf{Ay}$, where $\mathbf{A} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$, which is assumed known and fixed. Assuming that there are *p* predictors in **X** and writing **A** as a matrix of *p* row vectors such that $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_p \end{bmatrix}'$, it becomes clear that the fixed effect for the *i*th predictor can be written as $\mathbf{a}_i'\mathbf{y}$, where $\mathbf{a}_i$ is a vector of known coefficients $\mathbf{a}_i' = \begin{bmatrix} a_{i11} & a_{i12} & \cdots & a_{iJn_J} \end{bmatrix}$ of the same length as **y**. For simplicity, we dropped the subscript *i* in $\mathbf{a}_i$. Let **D** be a diagonal matrix such that $\mathbf{D} = \text{diag}\begin{bmatrix} a_{11} & \cdots & a_{Jn_J} \end{bmatrix}$ and $\mathbf{a}' = \mathbf{1}'\mathbf{D}$, and by writing $\tilde{\mathbf{y}} = \mathbf{Dy}$ as the transformed data, it follows that

$$\begin{aligned} \text{Var}(\gamma_i) &= \text{Var}(\mathbf{a}'\mathbf{y}) = \text{Var}\left(\mathbf{1}'\mathbf{Dy}\right) = \text{Var}\left(\mathbf{1}'\tilde{\mathbf{y}}\right) = N^2\text{Var}\left(\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\tilde{y}_{ij}\right), \\ &= N^2\text{Var}(\bar{\tilde{y}}_{..}) = N^2\left(\text{FPC}_2\frac{\tilde{\tau}_{00}}{J} + \text{FPC}_1\frac{\tilde{\sigma}^2}{N}\right) \end{aligned} \tag{A1}$$

where $\tilde{\tau}_{00} = \text{E}[\text{Cov}(\tilde{y}_{ij}, \tilde{y}_{i'j})] = \text{E}[\text{Cov}(a_{ij}y_{ij}, a_{i'j}y_{i'j})] = \text{E}\{\text{E}[a_{ij}a_{i'j}\text{Cov}(y_{ij}, y_{i'j}) \mid a_{ij}, a_{i'j}]\} = \text{E}(a_{ij}a_{i'j})\tau_{00}$ for $i \neq i'$ and $\tilde{\sigma}^2 = \text{E}[\text{Var}(\tilde{y}_{ij} \mid j)] = \text{E}[\text{Var}(a_{ij}y_{ij} \mid j)] = \text{E}\{\text{E}[a_{ij}\text{Var}(y_{ij} \mid j) \mid a_{i'j}]\} = \text{E}(a_{ij})\sigma^2$. As the expected values of the level-1 and level-2 variance components for the transformed data $\tilde{\mathbf{y}}$ are proportional to τ_{00} and σ^2, it follows that fixed effect sampling variance under sampling from a finite population may be obtained by multiplying the variance components by the corresponding FPCs in the variance expression when an infinite population is assumed.

Appendix B

R Function Implementing Finite Population Corrections of the Standard Errors

Equations (8) and (9) suggested that it is possible to obtain a consistent estimate of the covariance matrix of the fixed effect coefficients by replacing $\mathbf{G}^{*}$ by its maximum likelihood estimate $\hat{\mathbf{G}}^{*} = \text{FPC}_2 \times \hat{\mathbf{G}}^{*}$ and $\hat{\sigma}^{2*} = \text{FPC}_1 \times \hat{\sigma}^{2}$, but it is desirable to implement the correction with multilevel software packages. It would also be desirable if the correction could automatically take as input the cluster sizes information from a fitted model. A simple implementation is as follows:

```
vcovFPC <- function(object, popsize2 = NULL,
                    popsize1 = NULL, KR = FALSE) {
  # Obtained finite-population-adjusted standard errors for fixed effect
  # estimates for a fitted multilevel model
  #
  # Args:
  #   object: an R object of class merMod as resulting from lmer()
  #   popsize2: population size at level-2; if NULL, an infinite level-2
  #             population is assumed
  #   popsize1: population size at level-1; if NULL, an infinite level-1
  #             population is assumed
  #   KR: Whether Kenward-Roger approximation of standard errors should be used,
  #       which is recommended for small number of clusters and average cluster size.
  #       Default to FALSE.
  #
  # Returns:
  #   The variance-covariance matrix of the fixed effect estimates, as
  #   returned by vcov()
  if (!inherits(object, "merMod")) {
    stop("Wrong input: Not a fitted model from lmer() with class merMod")
  }
  if (length(object@flist) != 1) {
    stop("Wrong input: Only models with two levels are supported")
  }
  if (is.null(popsize1) & is.null(popsize2)) {
    message("No FPC specified; return results from lme4::vcov.merMod()")
    return(vcov(object))
  }
  PR <- object@pp
  N <- unname(object@devcomp$dims["n"])
  nclus <- unname(ngrps(object))
  if (isTRUE(popsize2 > nclus)) fpc2 <- 1 - nclus / popsize2
  else {
    fpc2 <- 1
    message("No FPC needed at level-2")
  }
  if (isTRUE(popsize1 > N)) fpc1 <- 1 - N / popsize1
  else {
    fpc1 <- 1
    message("No FPC needed at level-1")
  }
  if (fpc1 == 1 & fpc2 ==1) {
    message("Return results from lme4::vcov.merMod()")
    return(vcov(object))
  }
  A <- PR$Lambdat %*% PR$Zt
  Astar <- A * sqrt(fpc2)
  X <- PR$X
  Astar_X <- Astar %*% X
  D <- Matrix::Diagonal(nrow(Astar), fpc1) + tcrossprod(Astar)
  Fisher_I <- (crossprod(X) - crossprod(solve(t(chol(D)), Astar_X))) / fpc1
  Phi <- solve(Fisher_I) * sigma(object)^2
  Phi <- as(Phi, "dpoMatrix")
  nmsX <- colnames(X)
  dimnames(Phi) <- list(nmsX, nmsX)
  if (!KR) {
    return(Phi)
  } else {
    if (!require("pbkrtest")) {
      stop("Please install the `pbkrtest` package for the use of Kenward-Roger correction!")
    } else {
      SigmaG <- pbkrtest::get_SigmaG(object, details = 0)
      vcov_kr <- pbkrtest:::vcovAdj16_internal(Phi, SigmaG, X, details = 0)
      vcov_kr <- as(Phi, "dpoMatrix")
      return(vcov_kr)
    }
  }
}
```

The above function takes as input a fitted model object, with optional arguments `popsize2` and `popsize1` as the population sizes of level 2 and level 1. The optional argument `KR`, when set to `TRUE`, yields standard errors based on the Kenward-Roger approximation (Kenward & Roger, 1997) as implemented in the R package `pbkrtest` (). The output of the above function will output the variance-covariance matrix for the fixed effects adjusted for finite population, with the same structure as the output to the function `vcov()` in the `lme4` package. After defining the `vcovFPC()` function in R, one can, for example, type

```
vcovFPC(model1, popsize2 = 200)
```

if one has a fitted model `model1` and the population size for level 2 is 200. If one would like to get the corresponding *SE*s instead of the covariance matrix, this can be accomplished by taking the square roots of the diagonal elements using

```
sqrt(diag(vcovFPC(model1, popsize2 = 200)))
```

.