Examining the Rule of Thumb of Not Using Multilevel Modeling: The "Design Effect Smaller Than Two" Rule

Mark H. C. Lai and Oi-man Kwok

Texas A&M University

Author Note

Mark H. C. Lai, Department of Educational Psychology, Texas A&M University; Oi-man Kwok, Department of Educational Psychology, Texas A&M University.

Mark H. C. Lai is now at University of Southern California (hokchiol@usc.edu)

This is an Accepted Manuscript of an article published by Taylor & Francis in the Journal of Experimental Education on July 14, 2014, available online: https://www.tandfonline.com/10.1080/00220973.2014.907229.

We are grateful to Victor Willson for his invaluable suggestions on this manuscript.

Correspondence concerning this article should be addressed to Mark Lai, Department of Educational Psychology, Texas A&M University, College Station, TX 77840. Email: marklhc@neo.tamu.edu

Abstract

Researchers in education and other applied areas commonly use the rule of thumb of "design effect smaller than 2" as the justification of why they have not adopted the multilevel analyses or related techniques to account for the multilevel or clustered structure in their data. The rule, however, has not yet been systematically studied in previous research. In the present study, we generated data from three different models (with clustering effect only on the outcome, the level-1 predictor, and the relation between them). With a 3 (level of design effect) × 5 (cluster size) × 4 (number of clusters) factorial Monte Carlo simulation study we found that the rule should not be applied when researchers: (a) are interested in the effects of higher-level predictors, or (b) have a smaller cluster size (i.e., less than 10 observations per cluster). Implications of the findings and limitations of the study are discussed.

Keywords: simulation studies; design effects, multilevel, intraclass correlation, clustering

Examining the Rule of Thumb of Not Using Multilevel Modeling: The "Design Effect Smaller Than Two" Rule

Researchers in the field of education and other applied areas commonly refer to the *design effect* (Kish, 1965) when analyzing complex surveys or clustered data. Sometimes they apply the rule of thumb that "[i]f the design effect is smaller than two, using single level analysis on multilevel data does not seem to lead to overly misleading results" (Hox and Maas, 2002, p. 5; see also Muthén & Satorra, 1995). Surprisingly, to our knowledge there have not been any systematic methodological studies pertaining to the performance of this rule. In the present study, we try to fill this research gap using simulated data with the consideration of three design factors, including number of clusters, cluster sizes, and the magnitudes of the design effect.

In education literature we found that the "design effect smaller than two" rule was regularly invoked. Indeed, Peugh (2010), in a pedagogical article on how to apply multilevel modeling to educational data, recommended applied researchers to use multilevel modeling when the design effect was larger than two. In a recent article about on-task and off-task behaviors among 697 students from 35 classrooms (Kilian, Hofer, & Kuhnle, 2010), the authors used single-level analyses given that the design effects for all variables are smaller than 2 (between 1.01 and 1.61). In another study on victimization and bullying among 73 children among 46 classrooms and 18 schools (Bonnet, Goossens, & Schuengel, 2011), the authors reasoned that as the average number of children in a classroom was 1.6, the design effect must be smaller than 2, and they chose the single-level regression analysis rather than the multilevel model. From our literature review in PsycINFO and ERIC (Educational Resources Information Center database) we found many similar studies in the field of education using this rule as a justification for *not* using multilevel models for multilevel/clustered data (e.g., Bouman et al., 2012; De Los et al. 2011; Deng et al., 2006; Hong & You, 2012; Linnenbrink-Garcia, Rogat, & Koskey, 2011; Ly, Zhou, Chu, & Chen, 2012; Wong et al., 2006; von Grünigen, Kochenderfer-Ladd, Perren, & Alsaker, 2012). This rule was also commonly used in other research areas such as psychology (e.g., Corte & Zucker, 2008; Wagner, Christ, Pettigrew, Stellmacher, & Wolf, 2006), business (Qureshi & Fang, 2011), and medical science (Fuentes, Hart-Johnson, & Green, 2007).

When the clustered data structure is ignored and all variables are treated as if they are from one single level, the estimated regression coefficients are usually still unbiased, but the estimated standard errors associated with these coefficients are likely to be negatively biased or underestimated (Hox, 2010; Raudenbush & Bryk, 2002). This happens because single-level analyses assume that all observations are independent, but for clustered data the observations within a cluster are usually correlated (Thomas & Heck, 2001). An underestimated standard error invites a researcher to underestimate the uncertainty in their results (Goldstein & Spiegelhalter, 1996). It also leads to confidence intervals that are too narrow, as well as spurious statistically significant results (i.e., inflated Type I error rates; Hox, 2010; Snijders & Bosker, 2012).

The extent that the estimated standard error is biased depends on the degree to which individuals are correlated within clusters (Hox, 2010). This can be illustrated with an example in Hox (2010) about the popularity (derived by a sociometric procedure) of 2,000 students from 100 classrooms. There is no doubt variation in students' popularity within a classroom. On the other hand, if we represent each classroom by the mean popularity of its student, likely there will also be variation in the classroom means across classrooms, for reasons such as teachers' characteristics and classroom climate. The clustering effect is stronger when the between-classroom variation is relatively large compared to the within-classroom variation, which means that a randomly picked student from one classroom is much more similar to another one from the same classroom than to a third student from a different classroom. The ratio of the between-classroom variance to the sum of the between- and within-classroom variances is then called the intraclass correlation (ICC) which generally ranges from zero to one. In the popularity example the ICC for the popularity measure was .36, so there were quite a lot of variations of popularity at the classroom level.

If individuals within a cluster (e.g., a classroom) are no more similar to each other than to those in a different cluster (i.e., ICC = 0), then the independent observation assumption for the single-level analyses is not violated, and ignoring the clustered structure will not be a problem (Muthén, 1994). Furthermore, if ICC = 0 and both clusters and individuals within a cluster are randomly sampled, then the sample is equivalent to a simple random sample (Thomas & Heck, 2001). Apart from the ICC, cluster size can elevate the negative bias of the estimated standard error (i.e., more substantial negative bias in standard error with larger cluster size). When each cluster includes only one individual and the clusters can be assumed independent, the independent observation assumption again holds. To quantify the degree that a sample deviates from a simple random sample, Kish (1965) defined design effect (*deff*) of a sample statistic (e.g., of the mean of a variable) as “the ratio of the actual variance of a sample to the variance of a simple random sample of the same number of elements” (p. 258). In multilevel modeling, *deff* can be estimated as a function of the intraclass correlation (ICC) and average cluster size (c) such that (Muthén & Satorra, 1995):

$$deff = 1 + (c - 1) \times \mathrm{ICC}. \qquad (1)$$

The relationship between *deff*, c, and ICC is shown in Table 1. For example, a small ICC of .018 with a relatively large cluster size of 50 yields a *deff* of 1 + (50 − 1) × .018 = 1.9, whereas a large

ICC of .90 with a relatively small cluster size of 2 also yields a *deff* of $1 + (2 - 1) \times .90 = 1.9$. In the popularity example with 20 students per classroom, *deff* of the mean of the popularity measure $= 1 + (20 - 1) \times .36 = 7.84$, indicating that the sampling variance of the mean of the popularity measure will be almost eight times larger than if the 2,000 students are drawn as a simple random sample.

A rule of thumb for using *deff* in applied research is that when it is smaller than two, the degree of bias in the standard error is tolerable. Hox and Maas (2002) traced the source of the rule to Muthén and Satorra's (1995) paper. In their subsequent methodological work, Maas and Hox (2004, 2005) evaluated performances of multilevel modeling using only simulated data with *deff* larger than 2. However, in Muthén and Satorra's original simulation it was the ICC rather than the *deff* that was treated as the manipulated factor. In their study they generated data based on a random intercept model with one predictor that had equal level-1 and level-2 regression coefficients on the outcome, and a random effect for the intercept of the outcome. The only three conditions in their study with *deff* smaller than two were: (1) $c = 7$, ICC = .05 (*deff* = 1.3), (2) $c = 7$, ICC = .10 (*deff* = 1.6), and (3) $c = 15$, ICC = .05 (*deff* = 1.65), and in these three conditions, the percentages of bias in the standard error were: −5%, −10%, and −13%, respectively.

Readers should note that in the original paper Muthén and Satorra (1995) did *not* give any conclusions that clustering can be ignored when *deff* is smaller than 2. Also, the average cluster size in their simulated data was at least 7, which may not be applicable to some of the data sets in educational research where the average cluster size can be small (e.g., Bonnet et al. 2011; Jester et al., 2008; Wong et al., 2006). For example, in Bonnet et al.'s (2011) study there were 73 victimized children from 46 classrooms (i.e., an average cluster size of 1.6); and in Semke et al.'s (2010) study about the role of family involvement on students' disruptive behaviors they had 207 parents whose children are from 82 classrooms (i.e., an average cluster size of 2.5). In the simulation study by Clarke (2008), it was found that single-level analyses underestimated the standard errors of both level-1 and level-2 fixed effects (i.e., regression intercepts and regression coefficients) by 10 to 15% for conditions with cluster size smaller than 5, where *deff* was smaller than 1.4. Nevertheless, in that study *deff* was not directly manipulated and the ICC was fixed to .1, so the effects of *deff* and ICC were unknown.

In addition, we are interested in the effect of number of clusters on the standard errors when the rule of thumb is applied and the clustering is ignored. Equation (1) suggests that *deff* is not a direct function of number of clusters. Muthén and Satorra (1995) found no clear difference in their simulation results when the number of clusters was reduced from 200 to 50. To our knowledge there are no other simulation studies on *deff* that treated number of clusters as a design factor. In our simulations we have further included conditions for 20, 30, 50, and 100 clusters to see whether the effect of number of clusters is still ignorable when a higher level predictor is included and the average cluster size is small.

It is also questionable whether the rule of *deff* smaller than two is applicable to random coefficient models with clustering present on the relationship between the predictor and the outcome variable. This can again be illustrated with the popularity example where the effect of extraversion on popularity is of interest. If in reality the effect of extraversion on popularity were constant across classrooms (i.e., fixed coefficient model or model with only random effect for the intercept but not for other coefficients), a single-level analysis only ignores the clustering of students' extraversion and popularity, resulting in underestimated standard errors of the fixed effect estimates. However, there can be substantial variations in the effect of extraversion on popularity from classroom to classroom that single-level analyses do not take into account, but can be handled by the *random coefficient model* that incorporates the variation in the regression coefficient (i.e., the effect of extraversion on popularity). In Muthén and Satorra (1995) they only considered models where the regression coefficients were fixed, and to our knowledge the efficacy of the *deff* rule of thumb for random coefficient model has not been systematically studied. Therefore, in our Monte Carlo study, we have also evaluated the effectiveness of the rule of thumb under the random coefficient model by generating data with a random coefficient.

To the best of our knowledge, there have not been any published studies directly investigating the performance of *deff* and the validity/effectiveness of the rule of thumb. Yet a number of articles did refer to the rule as a justification of *not* using analyses that could account for the clustered structure in their data even though the sample characteristics in these studies were not comparable to those in Muthén and Satorra's (1995) simulation conditions. Given the limitations of Muthén and Satorra's study and other previous studies, in this study we have directly manipulated different levels of *deff*, varied the number of clusters, and included predictors from different levels.

Method

We used three data generating models for our simulation study. Each of them have one dependent variable (Y), one level-1 predictor (X), and one level-2 predictor (W), as shown in Figure 1. The three models differ in the location where clustering is present. In Model 1 (as shown in Figure 1a), the clustering is present only in Y (but not X, which means that ICC of X = 0); In Model 2 (as shown in Figure 1*b*), the clustering is present in both X and Y (but not the relationship between X and Y, which means that regression coefficient of X is still constant across clusters); In Model 3 (as shown in Figure 1c), the clustering is present in X, Y, and also the within-cluster relationship between X and Y.

Simulation Conditions

For each data-generating model, a 3 (value of *deff*) × 5 (*c*, cluster size) × 4 (*n*, number of clusters) factorial design was employed, totaling in 60 conditions. We selected *deff* = 1.1, 1.5, 1.9 to evaluate the bias of the estimated standard error (*SE*) for conditions when the rule of thumb said that the estimated parameter values and corresponding *SE*s from the single level model should be acceptable. Given that *deff* = 1 can be viewed as the lower boundary of *deff* (i.e., *deff* = 1 when ICC = 0), we chose *deff* = 1.1 and *deff* = 1.9 as the two extreme values and *deff* = 1.5 as the middle point. We chose five values of cluster size (*c*): 2, 3, 5, 10, and 50 per cluster, which are similar to the conditions in Clarke (2008). The numbers of clusters were 20, 30, 50, and 100, which is consistent with Maas and Hox (2004) and smaller than the ones used in Muthén and Satorra (1995).

Model Equations

For all three data-generating models, *X* (the level-1 predictor) was group-mean centered: centered with respect to the cluster mean of $X_{\cdot j}$ for the *j*th group (where $i = 1, \ldots, c$ and $j = 1, \ldots, n$). We chose to do group-mean centering so that the level-1 effect and the level-2 effect of *X* on *Y* are not confounded. An example is given in Kreft and de Leeuw (1988) about gender as a level-1 variable and gender ratio of a classroom, where both are hypothesized to have an effect on reading ability. However, they found that whereas female gender had a positive effect on reading scores within a classroom, classrooms with a higher proportion of female students tended to have lower mean reading scores. If researchers do single-level analyses and only do grand-mean centering, or do not center the predictors at all, the estimated effect is the combined effect of both level-1 and level-2 (Enders & Tofighi, 2007), which is not comparable to the effect estimated using multilevel modeling. Hence, as recommended by Enders and Tofighi (2007), we adopted group-mean centering to generate our data.

Under the traditional multilevel modeling framework (Raudenbush & Bryk, 2002), the level-1 (i.e., within-cluster) model can then be expressed as

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - X_{\cdot j}) + e_{ij}, \tag{2}$$

where Y_{ij} denotes the Y score of the ith individual in the jth cluster, β_{0j} denotes the regression intercept specific to the jth cluster, β_{1j} denotes the within-cluster regression coefficient of X for the jth cluster, and e_{ij} denotes the level-1 residual, which in this study was assumed to be normally distributed. The level-2 (i.e., between-cluster) model can be expressed in the following equations

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + \gamma_{02}X_{\cdot j} + u_{0j}, \tag{3}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \tag{4}$$

where γ_{00} is the level-2 intercept (or the average intercept across all n clusters) of Y; γ_{10} is the mean of level-1 regression coefficient β_{1j} across the n clusters; γ_{01} is the level-2 regression coefficient of Y regressing on W; γ_{02} is the level-2 regression coefficient of Y regressing on the cluster mean of X (i.e., $X_{\cdot j}$). The term u_{0j} is the normally distributed level-2 residual when β_{0j} is regressed on $X_{\cdot j}$ and W, and u_{1j} is the difference between β_{1j} from the mean (i.e., γ_{01}) for the jth cluster. We use σ^2 to denote the level-1 residual variance (i.e., $\text{Var}(e_{ij}) = \sigma^2$), τ_{00} to denote the level-2 intercept residual variance (i.e., $\text{Var}(u_{0j}) = \tau_{00}$), and τ_{11} to denote the level-2 slope variance (i.e., $\text{Var}(u_{1j}) = \tau_{11}$). The three components were generated to be independent (and so τ_{10}, the covariance between u_{0j} and u_{1j}, is zero). Combining the two levels, the model can be expressed as:

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + (\gamma_{10} + u_{1j})(X_{ij} - X_{\cdot j}) + \gamma_{02}X_{\cdot j} + u_{0j} + e_{ij}. \tag{5}$$

Model 1. Both τ_{11} and $\mathrm{Var}(X_{\cdot j})$ were set to zero in Model 1, and the model simplifies to

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}(X_{ij} - X_{\cdot j}) + u_{0j} + e_{ij}. \quad (6)$$

For all conditions in Model 1, we set $\gamma_{00} = 0$ without loss of generality. Both *W* and *X* were mean centered, had a variance of 1.0, and exerted a medium effect on *Y*, such that the variance explained was equal to 10% for both level-1 and level-2. The level-2 residual variance, τ_{00}, was fixed to 1.0, and thus based on partitioning of variance, for Model 1 the regression coefficient of *W* was

$$\gamma_{01} = \{[R^2_{\text{between}} \times \tau_{00} / (1 - R^2_{\text{between}})] / \mathrm{Var}(W)\}^{0.5} = 1 / 3, \quad (7)$$

where R^2_{between} referred to the proportion of explained variance of *Y* in level-2. The total variance of *Y* in level-2 was thus $\tau_{00} / (1 - R^2_{\text{between}}) = 10 / 9$. In level-1, the regression coefficient of *X* predicting Y was

$$\gamma_{10} = \{[R^2_{\text{within}} \times \sigma^2 / (1 - R^2_{\text{within}}) / \mathrm{Var}(X)]\}^{0.5} = \sigma / 3, \quad (8)$$

where R^2_{within} referred to the proportion of explained variance of *Y* in level-1. The total variance of *Y* in level-1 was thus $\sigma^2 / (1 - R^2_{\text{within}}) = 10\sigma^2 / 9$. Because ICC was defined as the ratio of the level-2 variance of *Y* to the sum of level-1 and level-2 variances of *Y*, in this study it was equal to:

$$(10 / 9) / [(10 / 9 + 10\sigma^2 / 9)] = 1 / (1 + \sigma^2), \quad (9)$$

and thus $\sigma^2 = (1 - \text{ICC}) / \text{ICC}$. For example, when $deff = 1.9$, $c = 2$, ICC is equal to .90 and $\sigma^2 = (1 - .90) / .90 = 0.11$. The corresponding specification of this model (under the multilevel structural equation modeling framework) is presented in Figure 1a.

Model 2. For Model 2, we set $\tau_{00} = 0$, $\mathrm{Var}(X_{\cdot j}) > 0$, and $\gamma_{10} = \gamma_{02}$. The model is expressed as:

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}(X_{ij} - X_{\cdot j}) + \gamma_{02}X_{\cdot j} + u_{0j} + e_{ij}. \quad (10)$$

The only difference between Model 1 and Model 2 is that the former has ICC of X being fixed to zero, whereas the latter one has ICC of X constrained to be equal to the ICC of Y. Because we are interested in the impact of the rule of thumb on the *SE*s of the fixed effects, to simplify the data generating model we set $\gamma_{10} = \gamma_{02}$ so that theoretically the estimates of both the single-level estimator and the multilevel estimator are the same. Due to the presence of the effect of $X_{\cdot j}$, γ_{01} was smaller than 1 / 3 and depended on the ICC of X so that the explained variance of Y in level-2 remained 10%. Note that the setup of Model 2 is the same as the regression model in the simulation study by Muthén and Satorra (1995), except that we have included a level-2 predictor W. For Model 2, the level-1 variance of X was kept to 1.0, but the level-2 variance of X, or Var($X_{\cdot j}$), was set to a value such that the ICC of X matched that of Y. For example, when $deff$ = 1.5, c = 5, ICC was equal to 0.125, and Var($X_{\cdot j}$) was equal to ICC / (1 – ICC) = 0.125 / 0.875 = 1 / 7. The corresponding specification of this model is presented in Figure 1b.

Model 3. As compared with model 1, in model 3 both τ_{00} and Var($X_{\cdot j}$) were larger than 0, and the model has the form as equation (5) with γ_{02} set to zero (i.e., X only has a level-1 effect on Y for Model 3, similar to the model in Clarke, 2008). The level-2 variance of X was the same as for Model 2, and consistent with Kwok, West, and Green (2007) the slope variance was set to be half of the value of the intercept variance, that is, $\tau_{11} = \tau_{00} / 2 = 0.5$. The corresponding specification of this model is presented in Figure 1c.

Procedures

For each condition, 2,000 data sets were generated using Mplus 7 (L. K. Muthén & Muthén, 1998–2012). For each data set, u_{0j}, u_{1j}, W_j, and $X_{\cdot j}$ (group means of X) were generated from four independent normal distributions with means being zero and the following variances: $\tau_{00} = 1$, $\tau_{11} = 0.5$, Var(W) = 1, and Var($X_{\cdot j}$) was equal to 0 for Model 1 and equal to ICC / (1 –

ICC) for Model 2 and 3 respectively. Then for each cluster j, β_{0j} and β_{1j} were computed from equations (3) and (4). For a cluster j, e_{ij} was generated from a normal distribution with a mean of zero and a variance of 1, and X_{ij} was generated from a normal distribution (independent of e_{ij}) with mean being $X_{\cdot j}$ and variance being 1. Finally Y_{ij} was computed by equation (5). The whole data generating process was automated using the Mplus montecarlo procedure.

Each data set was then analyzed in Mplus using both single-level model (with the assumption of independent observations and maximum likelihood estimation[1]) and multilevel model (with maximum likelihood estimation and robust standard errors, or MLR in Mplus) specifications. The single-level model can be expressed in the following equation

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}(X_{ij} - X_{\cdot j}) + e_{ij} \tag{11}$$

with no random effects (i.e., no u_{0j} and u_{1j}). Again, we chose to do group-mean centering so that both single-level and multilevel analyses estimate the same within-level effect. The multilevel model fitted to the data sets was the same as the data generating model in each of the three scenarios detailed above.

Dependent Variables

The major dependent variables of the present simulation study are the percentages of the relative biases in the *SE*s of the single-level estimators (i.e., when clustering was ignored) for γ_{00}, γ_{01}, and γ_{10}. First, a model ignoring the clustered structure was fitted to the data, and the relative biases for the *SE*s were calculated as

$$[{}^{1}/_{2000} \sum SE(\hat{\theta}_{(i)\mathrm{single}} - SD(\hat{\theta}_{\mathrm{single}})/SD(\hat{\theta}_{\mathrm{single}}), \tag{12}$$

where $\hat{\theta}$ is the estimated value of any parameter of interest (i.e., $\hat{\theta}$ refers to $\hat{\gamma}_{00}$, $\hat{\gamma}_{10}$, or $\hat{\gamma}_{01}$in this study). $SE(\hat{\theta}_{(i)\mathrm{single}})$ is the estimated *SE* of the estimated parameter $\hat{\theta}$ in the misspecified model for the *i*th replication, and $SD(\hat{\theta}_{\mathrm{single}})$ is the standard deviation of the 2,000 $\hat{\theta}$ values across replications (i.e., the empirical *SE* of $\hat{\theta}$) for the misspecified model (i.e., the single-level model as shown in equation (11)). The relative bias can then be interpreted as the average percentage of bias of the *SE* of $\hat{\theta}$ when the clustering of the data is not taken into account. Negative value of the relative bias indicates the percentage of underestimation in the *SE* in the single-level model (compared with the multilevel model).

Results

For all simulation conditions the convergence rates were high: 96.8% for Model 1, 99% for Model 2, and 93.8% for Model 3. Convergence rate was lowest for conditions with cluster size of 2 or 3 and when the design effect was small, which was similar to the results by Clarke (2008). In subsequent sections all results are based on the converged replications. Consistent with the findings from previous simulation studies (e.g., Clarke, 2008), the point estimates for the fixed effects, γ_{00}, γ_{01}, and γ_{10}, were generally unbiased for both the single-level and multilevel models, even with a smaller cluster size of 2 and a small design effect (with percentage of relative bias < 6%). Next, we present the results concerning bias of the *SE*s.

Relative Bias of the Standard Errors for the Multilevel Estimators

Because the present study focused more on the bias when clustering is ignored, the results related to the multilevel estimator are only briefly reported. For all three models, when there were at least 30 clusters (i.e., $n = 30$), the relative bias for the *SE*s of the fixed effect was within the acceptable range of ±10% recommended by Hoogland and Boosma (1998). Only under the $n = 20$ clusters condition, the average relative *SE* bias became slightly out of the ±10% bound. This was consistent with results in previous studies (e.g., Browne & Draper, 2000; Maas and Hox, 2005) that showed that *SE*s of fixed effects were relatively unbiased only when number of clusters was at least 30. Hence, multilevel models for data with number of cluster (n) equal to or fewer than 20 should be used with cautions given the likely bias in the *SE* of the fixed effect estimate. The details of these findings can be obtained from the first author.

Relative Bias of the Standard Errors for the Single-Level Estimators

The percentage of the relative biases for the *SE*s of the fixed effects, γ_{00}, γ_{01}, and γ_{10} (i.e., the grand intercept, the level-2 regression coefficient of W_j, and the average level-1 regression coefficient of X_{ij}), were shown in Tables 2, 3, and 4 respectively. The results below are organized by the three data generating scenarios.

Model 1. Consistent with previous research findings (Clarke, 2008; Muthén & Satorra, 1995), with the clustered data structure ignored and the use of the single-level model, almost all the estimated *SE*s showed a downward bias (or underestimation). On one hand, the estimated *SE*s of γ_{10} did not increase when *deff* increased, and across all conditions the strongest negative bias was −5.7% (or underestimated by 5.7%; Table 2). This supports the rule of thumb based on Muthén and Satorra (1995). On the other hand, there were stronger negative biases for the estimated *SE*s of the level-2 intercept γ_{00} and level-2 coefficient γ_{01} when *deff* increased. For γ_{00} (Table 4), When *deff* = 1.1, the percentage relative biases ranged from −3% to −12% (slightly over the recommended ±10% boundary); When *deff* = 1.5, the percentage relative biases ranged from −17% to −25%; When *deff* = 1.9, the biases ranged from −27% to −34%. *Deff* itself explained the majority of the variances in the relative *SE* biases of the estimated level-2 regression coefficient. In general larger negative biases were found for conditions with smaller cluster size (*c*) and fewer clusters (*n*). In particular with *n* = 20 and *c* = 2, the relative bias of the *SE* of γ_{01} was 11.9% even with a small *deff* of 1.1. The results for γ_{01} (Table 3) were basically in the exact same pattern as those for γ_{00}.

Model 2. With clustering effect on both the predictor *X* and the outcome *Y*, the general pattern of negative bias for the *SE*s of γ_{10}, γ_{00}, and γ_{01} was very similar to those in Model 1.

Model 3. With clustering effect on *X*, *Y*, and their relationship (i.e., the inclusion of u_{1j}), the degree of negative bias for the *SE*s of γ_{00} (Table 4) and γ_{01} (Table 3) was again very similar to the results in Model 1 and Model 2, but the bias for the *SE*s of γ_{10} (i.e., the average level-1 regression coefficient) was stronger. With *deff* = 1.1, the negative *SE* bias of γ_{10} (Table 2) was −9.9% for $c = 2$ and −8.2% for $c = 3$, with $n = 20$; With *deff* = 1.5, the negative bias was larger than 10% for almost all simulation conditions, and ranged from 18% to 23% with $c = 2$ or 3; With *deff* = 1.9 the negative bias was even stronger and ranged from 17% to 29% across conditions. Note that unlike in Model 2 where the negative *SE* bias for γ_{10} was small when $c \geq 10$, in Model 3 even with $c = 50$, the *SE* bias for γ_{10} was still substantial. In summary, there was strong negative bias on the estimated *SE*s of the relationship between *X* and *Y* for all conditions with *deff* ≥ 1.5 when the slope varied across clusters.

Discussion

When Researchers are Only Interested in Level-1 Effects

The present study aimed to evaluate the rule of thumb that *"if the design effect is smaller than two, the effect of clustering can be ignored"* (cf. Hox and Maas, 2002). We have found some support for this rule of thumb when certain conditions hold, including: (a) the cluster size (*c*) is at least 10, (b) the relations between level-1 predictors and the outcome are constant (i.e., no random coefficients allowed), and (c) the predictors are group-mean centered. Under these conditions, the use of the single-level model results in only slightly biased standard errors for the level-1 regression coefficient. This finding can be seen as a successful replication of the results of Muthén and Satorra (1995), as in their simulation conditions the cluster sizes were at least seven and the regression coefficients did not vary across clusters. Therefore, when all the following conditions are satisfied, namely (1) researchers are only interested in the effects of the level-1 predictors, (2) there is evidence that these level-1 predictors do not have any level-2 effects on the outcome, and (3) the effects of the predictors do not vary across clusters (i.e., no random coefficients), the rule of thumb holds reasonably.

On the other hand, it is not uncommon to find variables that have both level-1 and level-2 effects. We have discussed the example given in Kreft and de Leeuw (1988) about gender as a level-1 variable and gender ratio of a classroom. If the single-level analysis is used, researchers will get the unbiased estimate of the level-1 effect and acceptable estimates of the standard errors *only* when they do the *group-mean* centering for all the level-1 predictors, as shown in Table 2 (see also Enders & Tofighi, 2007; Kreft & de Leeuw, 1998). Otherwise if researchers only do grand-mean centering, or do not center the predictors at all, the standard error estimates will be biased with the use of the single-level analyses, especially when the cluster size is small. We did a post-hoc analysis by reanalyzing the data sets for Model 2 with a single-level model along with the un-centered level-1 predictor X. We found that the estimated SE of the regression coefficient of the un-centered level-1 predictor was biased (underestimated) by 10% to 33% when $deff = 1.9$ and $c \leq 5$, compared to 0.1% to 7.2% with centered level-1 predictor X as shown in Table 2. Although data with such a small cluster size may not be common in cross-sectional survey studies, they are the norm in longitudinal design and family related studies (e.g., dyadic data), and researchers handling such data with small cluster sizes should be cautious in basing their choice of analytic techniques on *deff*.

The presence of variations of level-1 coefficients across clusters makes the situation more complicated. It should be noted that in educational research often the effect of a predictor varies across classrooms or across schools. Hox (2010) gave the example where the relation between extraversion and popularity varies across classrooms. As another example, Turner et al. (2002) found that the relation between ethnicity and self-handicapping varies across classrooms. When researchers speculate the existence of such coefficient variations, multilevel techniques are in general more appropriate than single-level analyses. On the other hand, consistent with Maas and Hox (2005) we found that multilevel techniques produced substantial biases with no clear patterns on the fixed effect coefficients under conditions with small number of clusters (e.g., 20 clusters in our simulation study), along with small *deff* (e.g., *deff* = 1.1) and the inclusion of random coefficients in the model. Therefore, single-level analyses may be a more preferable option when there are as few as 20 clusters and when *deff* is small and very close to one.

Unfortunately in real situations, often researchers have little idea whether a predictor has both level-1 and level-2 effects and whether the effect varies across clusters, unless they explicitly test those possibilities using multilevel analyses. Based on our simulation results, we recommend using single-level analyses and ignoring the clustering effect only when *deff* is as small as 1.1. Under other conditions, multilevel analyses produce more accurate standard errors.

When Researchers are Also Interested in Level-2 Effects

The interpretations of the results pertaining to the level-2 fixed effects (i.e., regression coefficients and intercepts) are more straight-forward. Across all three models, we showed that the single-level *SE* estimates of the level-2 regression coefficients were substantially biased when *deff* ≥ 1.5. Therefore, if researchers are interested in the statistical inference of the effects of level-2 predictors, they should use techniques that adequately take into account the complex data structure unless there are too few clusters or *deff* is as small as 1.1. Such techniques include multilevel modeling as discussed in this paper, Taylor series approximation (LaVange, Stearns, Lafata, Koch, & Shah, 1996), and resampling (e.g., jackknife and bootstrap; see Rust & Rao, 1996).

Effect of Number of Clusters

Another question the present research tries to answer is whether the number of clusters, *n*, affects the validity of the rule, because *n* affects the total sample size but *deff* is not a function of *n*. We found that when clustering is ignored, increment in *n* can only slightly reduce the negative bias. A possible reason is that the large cluster size (e.g., 30 clusters or more) already leads to a sufficiently large total sample size (at least with balanced data) and thus the effect of the total sample size reaches a ceiling effect.

Limitations

The present study has several limitations. First, the models we used include only one level-1 and one level-2 predictors. With more predictors and due to potential multicollinearity or other more complex relationships, our findings may not apply. We encourage future research to study the *deff* with models that reflect the complexity of substantive research, which may include the test of mediation effect, or moderation effect, or both simultaneously. Second, we only used a two-level clustering structure, so the results may not generalize to other data structure such as those involving three levels of clustering or those with crossed random effects. Future research addressing questions involving three or more levels of clustering (e.g., students nested within classrooms within schools) or other data structures (e.g., data with crossed random effects, see Beretvas, 2010) is needed. Third, we did not study other potential factors that could influence the results such as the magnitude of the regression coefficients. In addition, we fixed the cluster size to be constant within each simulation condition, that is, all conditions assume a balanced design. This may not hold in real research and previous research has shown that the performance of analyses ignoring the clustered data structure may be worse with an unbalanced design (Clarke, 2008). Despite the limitations, our study shows that the rule of thumb of design effect smaller than two only works in limited situations, and researchers should be with cautious when applying the rule of thumb.

References

Beretvas, S. N. (2010). 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.

Bonnet, M., Goossens, F. A., & Schuengel, C. (2011). Parental strategies and trajectories of peer victimization in 4 to 5 year olds. *Journal of School Psychology*, *49*, 385–398. doi:10.1016/j.jsp.2011.04.002

Bouman, T., van der Meulen, M., Goossens, F. A., Olthof, T., Vermande, M. M., & Aleva, E. A. (2012). Peer and self-reports of victimization and bullying: their differential association with internalizing problems and social adjustment. *Journal of School Psychology*, *50,* 759–774. doi:10.1016/j.jsp.2012.08.004

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15,* 391–420. doi:10.1007/s001800000041

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health*, *62*, 752–758. doi:10.1136/jech.2007.060798

Corte, C., & Zucker, R. A. (2008). Self-concept disturbances: Cognitive vulnerability for early drinking and early drunkenness in adolescents at high risk for alcohol problems. *Addictive Behaviors*, *33*, 1282–1290. doi:10.1016/j.addbeh.2008.06.002

De Los Reyes, A., Youngstrom, E. A., Swan, A. J., Youngstrom, J. K., Feeny, N. C., & Findling, R. L. (2011). Informant discrepancies in clinical reports of youths and interviewers' impressions of the reliability of informants. *Journal of Child and Adolescent Psychopharmacology*, *21*, 417–24. doi:10.1089/cap.2011.0011

Deng, S., Lopez, V., Roosa, M. W., Ryu, E., Burrell, G. L., Tein, J.-Y., & Crowder, S. (2006). Family processes mediating the relationship of neighborhood disadvantage to early adolescent internalizing problems. *The Journal of Early Adolescence*, *26*, 206–231. doi:10.1177/0272431605285720

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. doi:10.1037/1082-989X.12.2.121

Fuentes, M., Hart-Johnson, T., & Green, C. R. (2007). The association among neighborhood socioeconomic status, race and chronic pain in Black and White older adults. *Journal of the National Medical Association*, *99*, 1160–1169.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *159*, 385–443. doi:10.2307/2983325

Greene, W. H. (2003). Econometric analysis (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Hong, S., & You, S. (2012). Understanding Latino children’s heterogeneous academic growth trajectories: Latent growth mixture modeling approach. *The Journal of Educational Research*, *105*, 235–244. doi:10.1080/00220671.2011.584921

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

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

Hox, J. J., & Maas, C. J. M. (2002). Sample sizes for multilevel modeling. In J. Blasius, J. J. Hox, E. de Leeuw, & P. Schmidt (Eds.), *Social science methodology in the new millennium: Proceedings of the fifth international conference on logic and methodology* (2nd expanded ed.). Opladen, RG: Leske + Budrich.

Jester, J. M., Nigg, J. T., Buu, A., Puttler, L. I., Glass, J. M., Heitzeg, M. M., … Zucker, R. A. (2008). Trajectories of childhood aggression and inattention/hyperactivity: Differential effects on substance abuse in adolescence. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*, 1158–1165. doi:10.1097/CHI.0b013e3181825a4e

Kilian, B., Hofer, M., & Kuhnle, C. (2010). Value orientations as determinants and outcomes of conflicts between on-task and off-task actions in the classroom. *Learning and Individual Differences*, *20,* 501–506. doi:10.1016/j.lindif.2010.03.003

Kish, L. (1965). *Survey Sampling*. New York, NY: Wiley.

Kreft, G. G., & de Leeuw, E. D. (1988). The See-Saw Effect: A multilevel problem? *Quality and Quantity*, *22*, 127–137. doi:10.1007/bf00223037

Kwok, O., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research*, *42*, 557–592. doi:10.1080/00273170701540537

LaVange, L. M., Stearns, S. C., Lafata, J. E., Koch, G. G., & Shah, B. V. (1996). Innovative strategies using SUDAAN for analysis of health surveys with complex samples. *Statistical Methods in Medical Research, 5*, 311–329. doi:096228029600500306

Linnenbrink-Garcia, L., Rogat, T. K., & Koskey, K. L. K. (2011). Affect and engagement during small group instruction. *Contemporary Educational Psychology*, *36*, 13–24. doi:10.1016/j.cedpsych.2010.09.001

Ly, J., Zhou, Q., Chu, K., & Chen, S. H. (2012). Teacher–child relationship quality and academic achievement of Chinese American children in immigrant families. *Journal of School Psychology*, *50*, 535–553.

Maas, C. J. M., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, *46*, 427–440. doi:10.1016/j.csda.2003.08.006

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

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376–398. doi:10.1177/0049124194022003006

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

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

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*, 85–112. doi:10.1016/j.jsp.2009.09.002

Qureshi, I., & Fang, Y. (2011). Socialization in open source software projects: A growth mixture modeling approach. *Organizational Research Methods*, *14*, 208–238.

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

Rust, K. F., & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, *5*, 283–310. doi:096228029600500305

Semke, C. A., Garbacz, S. A., Kwon, K., Sheridan, S. M., & Woods, K. E. (2010). Family involvement for children with disruptive behaviors: The role of parenting stress and motivational beliefs. *Journal of School Psychology*, *48*, 293–312. doi:10.1016/j.jsp.2010.04.001

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

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage.

Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education*, *42*, 517–540. doi:10.1023/A:1011098109834

Turner, J. C., Midgley, C., Meyer, D. K., Gheen, M., Anderman, E. M., Kang, Y., & Patrick, H. (2002). The classroom environment and students' reports of avoidance strategies in mathematics: A multimethod study. *Journal of Educational Psychology*, *94*, 88–106. doi:10.1037//0022-0663.94.1.88

Von Grunigen, R., Kochenderfer-Ladd, B., Perren, S., & Alsaker, F. D. (2012). Links between local language competence and peer relations among Swiss and immigrant children: The mediating role of social behavior. *Journal of School Psychology*, *50*, 195–213. doi:10.1016/j.jsp.2011.09.005

Wagner, U., Christ, O., Pettigrew, T. F., Stellmacher, J., & Wolf, C. (2006). Prejudice and minority proportion: contact instead of threat effects. *Social Psychology Quarterly*, *69*, 380–390. doi:10.1177/019027250606900406

Wong, M. M., Nigg, J. T., Zucker, R. A., Puttler, L. I., Fitzgerald, H. E., Jester, J. M., … Adams, K. (2006). Behavioral control and resiliency in the onset of alcohol and illicit drug use: A prospective study from preschool to adolescence. *Child Development*, *77*, 1016–1033. doi:10.1111/j.1467-8624.2006.00916.x

Footnotes

[1]As discussed in Greene (2003, chapter 17), in multiple regression analyses when the normality assumption of the errors holds, as is the case for the data generating models of the present study, the ordinary least squares estimates and the maximum likelihood estimates are equivalent.

Table 1

Values of Intraclass Correlation Given the Average Cluster Size and the Design Effect

	deff		
c	**1.1**	**1.5**	**1.9**
2	.100	.500	.900
3	.050	.250	.450
5	.025	.125	.225
10	.011	.056	.100
50	.002	.010	.018

Note. *c* = cluster size. *deff* = design effect.

Table 2

Percentage Relative Bias of the Estimated Standard Errors for the Level-1 Regression Coefficient (γ_{10}) Ignoring Clustered Structures.

		Model 1			Model 2			Model 3		
		deff			*deff*			*deff*		
n	*c*	**1.1**	**1.5**	**1.9**	**1.1**	**1.5**	**1.9**	**1.1**	**1.5**	**1.9**
20	**2**	−5.1	−5.7	−5.3	−6.1	−7.0	−7.2	−9.9	−22.2	−29.0
	3	−3.5	−3.3	−2.8	−5.3	−5.2	−5.0	−8.2	−19.4	−26.4
	5	−0.6	−0.9	−1.1	−2.4	−2.3	−2.4	−3.2	−12.8	−20.0
	10	−0.3	−0.6	−0.8	0.8	0.9	1.0	−3.8	−12.2	−18.8
	50	−0.9	−0.9	−0.9	−0.1	−0.3	−0.5	−2.9	−11.9	−18.5
30	**2**	−4.7	−3.9	−2.9	−2.6	−3.7	−5.6	−9.5	−21.9	−27.9
	3	−1.0	−1.6	−1.8	−3.6	−3.9	−4.0	−5.1	−16.9	−24.2
	5	−0.4	−0.6	−0.6	−0.4	−0.9	−1.3	−5.8	−15.9	−22.9
	10	−0.6	−0.9	−1.1	0.3	0.2	0.3	−2.9	−11.9	−18.7
	50	0.4	0.4	0.5	−1.1	−1.4	−1.5	−1.5	−10.5	−17.4
50	**2**	−2.7	−2.8	−2.0	−2.6	−2.9	−4.8	−8.7	−22.6	−29.0
	3	0.8	2.0	2.5	−3.4	−4.0	−4.2	−4.3	−16.7	−24.2
	5	−0.8	−0.3	0.0	−1.1	−1.5	−1.8	−4.2	−15.7	−23.2
	10	1.6	1.7	1.8	−0.2	−0.3	−0.3	−1.9	−10.8	−17.6
	50	−1.2	−1.3	−1.3	−2.2	−2.2	−2.3	−2.8	−11.3	−17.8
100	**2**	−1.2	−1.5	−2.1	−0.2	−2.4	−3.8	−8.7	−22.5	−28.4
	3	1.7	1.6	1.4	−1.8	−2.3	−2.7	−4.2	−17.1	−24.7
	5	−0.5	−0.9	−1.1	0.3	0.2	−0.1	−3.9	−15.3	−22.7
	10	−1.3	−1.4	−1.5	−0.3	0.0	0.1	−2.8	−12.4	−19.4
	50	−3.5	−3.6	−3.7	−2.0	−2.0	−2.0	−2.2	−10.8	−17.4

Note. *n* = number of clusters. *c* = cluster size. *deff* = design effect. Model 1: random intercept model with ICC of $X = 0$; Model 2: ICC of $X > 0$ and X have both within- and between-level effects on Y; Mode 3: random coefficient model with ICC of $X > 0$ and the regression coefficients of X on Y varying across clusters. Negative value indicates the percentage of underestimation.

Table 3

Percentage Relative Bias of the Estimated Standard Errors for the Level-2 Regression Coefficient (γ_{01}) Ignoring Clustered Structures.

		Model 1			Model 2			Model 3		
		deff			*deff*			*deff*		
n	*c*	**1.1**	**1.5**	**1.9**	**1.1**	**1.5**	**1.9**	**1.1**	**1.5**	**1.9**
20	**2**	−11.9	−24.6	−33.6	−11.5	−25.3	−33.7	−8.7	−19.8	−26.6
	3	−9.1	−22.6	−31.5	−9.3	−23.1	−32.3	−10.8	−23.3	−30.3
	5	−5.2	−20.1	−29.8	−7.4	−22.3	−31.7	−9.4	−21.4	−29.3
	10	−7.7	−20.6	−29.4	−7.0	−21.4	−30.8	−6.0	−19.8	−28.7
	50	−8.9	−22.0	−30.6	−5.3	−19.4	−28.9	−6.8	−20.6	−29.5
30	**2**	−7.4	−21.6	−31.0	−10.1	−23.7	−32.3	−7.2	−19.7	−26.6
	3	−9.8	−24.6	−33.4	−8.0	−21.0	−29.9	−9.9	−21.6	−28.4
	5	−2.9	−17.3	−26.9	−6.4	−21.2	−30.7	−7.2	−19.5	−27.6
	10	−6.3	−19.9	−28.9	−7.7	−21.7	−30.8	−6.7	−20.6	−29.3
	50	−7.3	−21.0	−29.9	−2.4	−17.6	−27.5	−7.7	−20.4	−28.8
50	**2**	−6.8	−19.3	−28.2	−6.0	−20.0	−29.5	−5.6	−18.0	−24.8
	3	−8.2	−23.1	−31.9	−4.2	−18.0	−27.4	−10.4	−21.3	−27.6
	5	−4.7	−18.9	−28.3	−6.7	−21.1	−30.3	−6.8	−18.6	−26.4
	10	−5.7	−18.6	−27.3	−6.1	−20.6	−30.0	−4.8	−19.2	−28.1
	50	−5.8	−18.9	−27.7	−4.6	−19.3	−28.8	−5.8	−19.3	−28.1
100	**2**	−4.4	−17.5	−26.8	−1.7	−16.4	−26.5	−6.1	−17.0	−23.3
	3	−7.3	−20.8	−29.2	−4.2	−18.7	−28.1	−7.2	−19.6	−26.3
	5	−5.4	−19.3	−28.4	−9.0	−22.0	−30.4	−3.7	−17.2	−25.6
	10	−7.0	−19.9	−28.4	−5.6	−19.7	−29.0	−3.8	−18.0	−26.9
	50	−4.9	−19.4	−28.8	−5.1	−19.6	−29.1	−7.9	−20.6	−29.1

Note. *n* = number of clusters. *c* = cluster size. *deff* = design effect. Model 1: random intercept model with ICC of $X = 0$; Model 2: ICC of $X > 0$ and X have both within- and between-level effects on Y; Mode 3: random coefficient model with ICC of $X > 0$ and the regression coefficients of X on Y varying across clusters. Negative value indicates the percentage of underestimation.

Table 4

Percentage Relative Bias of the Estimated Standard Errors for the Level-2 Intercepts (γ_{00}) Ignoring Clustered Structures.

		Model 1			Model 2			Model 3		
		deff			*deff*			*deff*		
n	*c*	**1.1**	**1.5**	**1.9**	**1.1**	**1.5**	**1.9**	**1.1**	**1.5**	**1.9**
20	**2**	−9.1	−21.7	−30.8	−6.8	−20.4	−30.7	−7.1	−19.6	−26.4
	3	−8.6	−22.7	−31.9	−5.9	−20.5	−30.1	−10.5	−22.7	−29.9
	5	−5.9	−20.0	−29.4	−3.9	−19.0	−28.8	−5.2	−18.1	−26.4
	10	−6.8	−20.5	−29.4	−5.5	−20.3	−29.7	−7.8	−19.4	−27.4
	50	−7.5	−20.5	−29.2	−3.2	−18.4	−28.3	−5.2	−18.7	−27.6
30	**2**	−8.2	−21.8	−30.7	−5.8	−19.8	−29.8	−7.3	−20.5	−27.2
	3	−6.5	−21.1	−30.5	−6.3	−20.3	−29.4	−6.2	−18.6	−26.2
	5	−6.6	−20.7	−29.9	−5.6	−19.8	−29.4	−3.4	−17.2	−25.8
	10	−7.4	−20.5	−29.2	−7.1	−21.3	−30.4	−6.2	−18.4	−26.7
	50	−6.6	−19.7	−28.6	−3.5	−18.3	−27.9	−5.2	−18.7	−27.6
50	**2**	−7.2	−19.2	−27.1	−5.3	−18.8	−28.9	−6.1	−17.1	−23.1
	3	−6.7	−19.9	−28.9	−5.8	−20.0	−29.1	−4.8	−16.6	−24.5
	5	−8.2	−20.5	−29.1	−4.5	−18.5	−28.1	−3.2	−16.2	−24.7
	10	−6.6	−20.4	−29.5	−4.4	−19.7	−29.4	−6.5	−18.6	−26.9
	50	−6.8	−19.8	−28.5	−5.0	−19.6	−29.2	−4.8	−18.2	−27.2
100	**2**	−8.8	−20.1	−27.7	−4.4	−18.1	−28.1	−7.1	−17.4	−23.3
	3	−6.4	−19.4	−28.3	−4.6	−18.0	−27.0	−5.2	−17.6	−25.2
	5	−7.9	−20.5	−29.2	−2.7	−16.4	−26.0	−4.6	−17.6	−25.9
	10	−5.9	−19.9	−28.9	−4.3	−18.8	−28.3	−4.8	−18.5	−27.1
	50	−5.7	−19.2	−28.0	−4.6	−18.4	−27.8	−4.1	−17.4	−26.3

Note. *n* = number of clusters. *c* = cluster size. *deff* = design effect. Model 1: random intercept model with ICC of $X = 0$; Model 2: ICC of $X > 0$ and X have both within- and between-level effects on Y; Mode 3: random coefficient model with ICC of $X > 0$ and the regression coefficients of X on Y varying across clusters. Negative value indicates the percentage of underestimation.

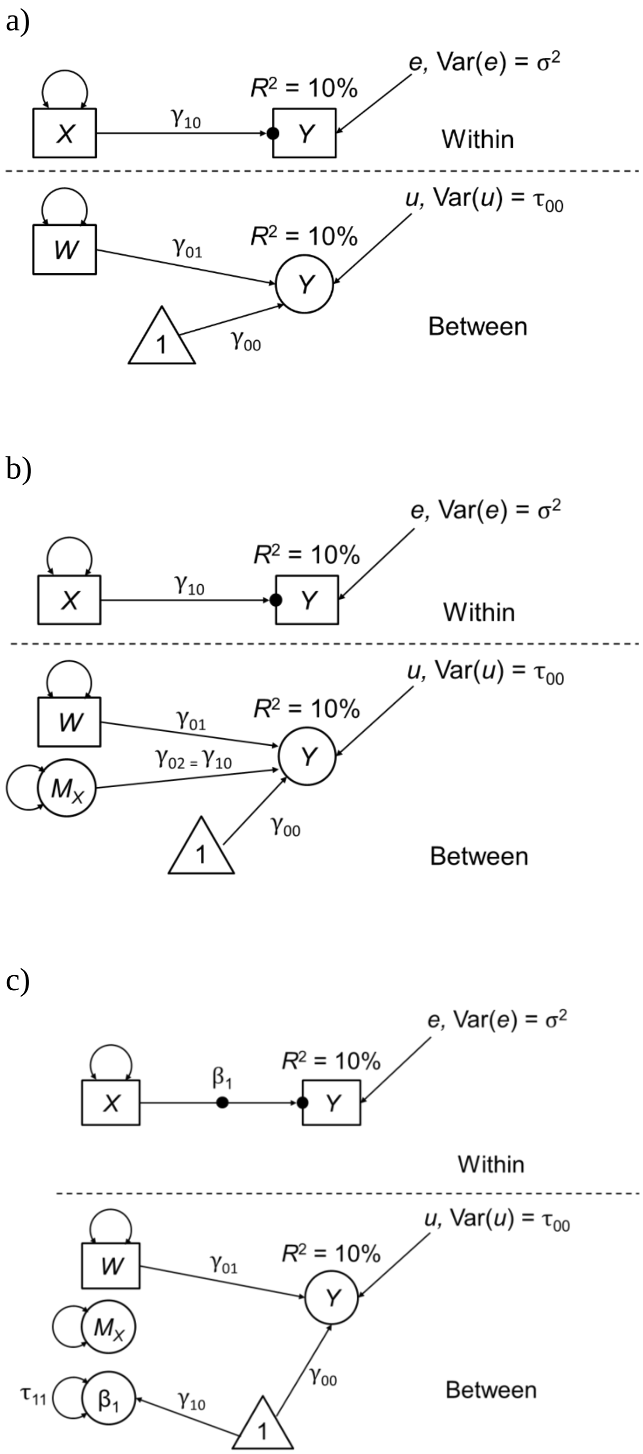

Figure 1. Multilevel path diagram for the data generation of (a) Model 1: random intercept model with ICC of X = 0; (b) Model 2: ICC of $X > 0$ and X have both within- and between-level variances; and (c) Mode 3: random coefficient model with ICC of $X > 0$ and the regression coefficients of X on Y varying across clusters (i.e., τ_{11}).