Article

Sociological Methods & Research
2018, Vol. 47(3) 430-457

DOI: 10.1177/0049124116630563
journals.sagepub.com/home/smr

Design Effects of Multilevel Estimates From National Probability Samples

Laura M. Stapleton[1] and Yoonjeong Kang[2]

Abstract

This research empirically evaluates data sets from the National Center for Education Statistics (NCES) for design effects of ignoring the sampling design in weighted two-level analyses. Currently, researchers may ignore the sampling design beyond the levels that they model which might result in incorrect inferences regarding hypotheses due to biased standard error estimates; the degree of bias depends on the informativeness of any ignored stratification and clustering in the sampling design. Some multilevel software packages accommodate first-stage sampling design information for two-level models but not all. For five example public release data sets from the NCES, design effects of ignoring the sampling design in unconditional and conditional two-level models are presented for 15 dependent variables selected based on a review of published research using these five data sets. Empirical findings suggest that there are minor effects of ignoring the additional sampling design and no differences in inference would be made had the first-stage sampling design been ignored. Strategically, researchers without access to multilevel software that can accommodate the sampling might consider including stratification variables as independent variables at level 2 of their model.

[1] Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, USA
[2] American Institutes for Research, Washington, DC, USA

Corresponding Author:
Laura M. Stapleton, Department of Human Development and Quantitative Methodology, University of Maryland, 3304 Benjamin Building, College Park, MD 20742, USA.
Email: lstaplet@umd.edu

Keywords

multilevel, sampling, survey, weights, structural modeling

The use of multilevel modeling with data from national probability samples has become more common in educational and behavioral research in recent years (O'Connell and McCoach 2008). These models posit relations within clusters, such as classrooms, schools, or neighborhoods, as well as posit relations among cluster constructs (Raudenbush and Bryk 2002; Sniders and Bosker 2012). Methods used to select samples for many national probability studies are excellent for such analyses, given that clusters of individuals within units are usually approached for response. However, conducting a multilevel analysis does not necessarily address all elements of the sampling design and thus inference may be compromised (Kish 1965; Wolter 1985). In particular, the estimates of standard errors, or sampling variances, may be inappropriate. If one or more stages of sampling is ignored, then standard errors may be underestimated and, conversely, if stratification in the sampling design is ignored, then standard errors may be overestimated (Kish 1965). In this article, we discuss how stratification and multistage selection might be addressed in a multilevel analysis of national probability sample data, specifically in an educational context where students or teachers are nested in schools. First, a short discussion of typical sampling procedures used by the National Center for Education Statistics (NCES) is provided, and the multilevel modeling of such data is then described as well as concerns regarding the impacts of ignoring the stratification and multistage selection present in most sampling designs. Next, we summarize the sampling designs used for five currently popular public release data sets and review the published multilevel analyses that have used these data sets. We then present an empirical evaluation of the effects of running weighted two-level analyses on these data and the possible inference concerns with not fully addressing the sampling design. Specifically, we document the design effects (or misestimation of the standard errors if the stratification or clustering is ignored). This article concludes with steps that applied researchers can take to evaluate the informativeness of primary sampling unit (PSU) clustering and of stratification in the sampling design and therefore the possible design effects they may encounter if the full sampling design is ignored in a multilevel analysis. The estimation of measures of informativeness can be easily calculated within an analysis of variance (ANOVA) framework and thus does not require more sophisticated software.

Background

National education-related surveys generally are not conducted using simple random sampling designs (e.g., Ingels et al. 2005; Tourangeau et al. 2009). Some designs used by the NCES involve three stages of sampling: PSUs of single counties or groups of counties, then schools within those selected counties, and then ultimate sampling units (USUs) of students or teachers within the selected schools. At the first two stages, stratification and probability proportional to size (PPS) sampling[1] might be used and at the final stage, stratification often is used with disproportionate sampling across strata. The first-stage stratification can be complex, with the use of certainty strata and noncertainty strata (Kish 1965). These strata may be defined by various combinations of variables such as Census region, proportion of specific race/ethnicity, size of PSU, and average per capita income. Within noncertainty strata, PPS sampling often is used to select PSUs per stratum. Within PSUs, schools may be stratified by such variables as public/private status, urban/rural location, or grade level and then sampled with implicit stratification[2] by school characteristics. Schools are thus treated as secondary sampling units. Finally, students (or teachers) might be selected from the sampled schools using stratification on individual characteristics, with a given target sample size per school. Some NCES studies use a two-stage, instead of three-stage, stratified sampling approach (see Tourkin and colleagues 2004, as an example). In these designs, the PSUs are the schools, stratified by such variables as level, region, and percent minority, selected with PPS sampling. Within schools, a fixed sample size of individuals may be selected using stratified sampling across characteristics such as race/ethnicity and gender.

Multilevel Models With NCES Data

Software packages, such as HLM (Raudenbush et al. 2011), MLwiN (Rasbash et al. 2012), and MIXED components of Statistical Package for the Social Sciences (2002) and SAS (SAS Institute Inc. 2013), have long been available for the analysis of two-level models with manifest variables. Additionally, estimation methods have been recently implemented into structural equation modeling programs, for example, Mplus (Muthén and Muthén 2011), LISREL (du Toit and du Toit 2008), and the Gllamm package within Stata (Rabe-Hesketh, Skrondal, and Pickles 2004), allowing researchers to model multilevel relations among latent constructs. Unless otherwise specified, the estimation methods implemented within these software programs

operate on the assumption that clusters are a random selection from some finite population and persons within those sampled clusters are also a random selection thus yielding the assumed independent residuals at each level. Most national education-related data sets use sampling procedures that are more complicated in design however. In three-stage sampling designs in education, data usually have some degree of dependence among observations at the school level. This dependence can lead to negatively biased estimates of sampling variances (Kish 1965) of the parameters of interest at level 2. Additionally, when sample designs include stratification, the stratification usually is intended to provide more efficient estimates of population parameters (Kalton 1983; Kish 1965). When modeling with data obtained through a stratified sample, if the stratification is ignored, the resulting estimates of the sampling variances will tend to be positively biased, assuming that the data exhibit some level of homogeneity within strata (Kalton 1983; Kish and Frankel 1974).

Researchers in the social sciences have used multilevel model-based techniques, both manifest and latent, with national data sets thus addressing *some* of the complexity of the sample design, specifically the clustering of USUs in a higher order cluster (e.g., Hox 2002; Kaplan and Elliott 1997; Lee et al. 2006; Palardy 2008). For example, we might suppose a simple two-level bivariate fixed regression of some response variable, y_{ij}, for student i nested in school j as

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_{1j}x_{ij} + r_{ij}, \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}x_j + u_{0j}, \\ \beta_{1j} &= \gamma_{10}, \end{aligned}$$

where x_{ij} is the student-level predictor and x_j is the school mean of the level-1 variable and where r_{ij} is assumed distributed $\sim N(0,\sigma^2)$ and u_{0j} is assumed distributed as $N(0, \tau_{00})$. As one example of a problem with using this model with data collected via complex sampling designs, in this model, the u_{0j} is assumed independent across the level-2 units. This assumption would likely be violated with stratified multistage sampling designs, as those schools within the same strata would share a dependency and, likewise, those schools within the same PSU. This lack of independence has ramifications for the estimates of the standard errors for each of the parameter estimates in the model (γ_{00}, γ_{10}, γ_{01}, σ^2, and τ_{00}). While the entire sampling design can be accommodated using a model-based approach (modeling a level for each of the sampling stages, e.g., level 1 being students, level 2 being schools, and level 3 being geographic areas), typically researchers are interested only in substantive questions about the

two lowest levels such as students within schools and therefore a completely model-based approach would be unnecessarily complex. Furthermore, a strictly model-based approach would require the inclusion of stratification variables, both explicit and implicit (Sterba 2009), and predictor variables would need to be modeled with interactions with the stratification indicators. This approach appears overly complicated, given the focus of most research questions.

Appropriate Variance Estimation for Multilevel Model Estimates

A few studies have examined the effect of using two-level modeling while ignoring the first stage of sampling with three-stage samples and while ignoring the first-stage stratification in the sampling design (Asparouhov and Muthén 2006; Grilli and Pratesi 2004; Kovacevic and Rai 2003; Rabe-Hesketh and Skrondal 2006). The latter two studies examined estimation with a single outcome variable (ordinal and dichotomous, respectively) while the former examined estimation issues in a multilevel confirmatory factor analysis framework. While the models examined were different, the findings generalize across models. Of most usefulness given the breadth of its simulation design, assuming continuous and normally distributed measures, Asparouhov and Muthén (2006) found that when conducting a two-level confirmatory factor analysis with data from a stratified three-stage sampling design, sampling variances were underestimated when excluding the third level of the sampling design, as expected. Specifically, when first-stage clustering was ignored, given their simulation conditions, standard error estimates were 7 to 15 percent too small depending on the parameter estimate of interest. When first-stage stratification was ignored, on the other hand, standard errors were overestimated about 5 percent for cluster-level parameters. Additionally, and importantly in a structural equation modeling (SEM) framework, when ignoring both components of sampling, likelihood ratio tests had extremely high model rejection rates (90 percent compared to the expected 5 percent given the correct model specification). Grilli and Pratesi (2004) and Rabe-Hesketh and Skrondal (2006) each evaluated, via simulation, the estimation of single outcome manifest variable multilevel models given a two-stage sample with stratification at the first stage. In both, the authors found that estimation of standard errors was positively biased when the stratification was ignored.

Given these concerns, statisticians have proposed methods of estimating two-level models from multistage stratified sampling designs (Asparouhov and Muthén 2006; Grilli and Pratesi 2004; Rabe-Hesketh and Skrondal

2006). In this type of analysis, some sampling design information is modeled, while some is accounted for in the estimator, and thus Rabe-Hesketh and Skrondal (2006) term this type of modeling a hybrid aggregated–disaggregated approach. When a multilevel analysis does not include all facets of the sampling design within the model, multilevel pseudo-maximum likelihood (MPML) estimation has been developed to obtain unbiased parameter estimates. Sampling variance estimation is accomplished with a sandwich estimator, providing linearized estimates based on the first-stage sampling characteristics. This MPML method was evaluated by Asparouhov and Muthén (2006), Grilli and Pratesi (2004), and Rabe-Hesketh and Skrondal (2006) under conditions of continuous, ordinal, and dichotomous outcome data, respectively. Consider our simple two-level bivariate example and suppose that data were collected using a three-stage sampling design. The response variable, y_{ijk}, is of individual i in school j in PSU k, and we might model with covariate x_{ijk} at the individual level and x_{jk} at the school level. At level 1, we hypothesize a density function of y_{ijk} to be $f(y_{ijk}|x_{ijk}, y_{jk}, \theta_w)$, and at level 2, the density function of the school intercept, γ_{jk}, to be $\phi(y_{jk}|x_{jk}, \theta_b)$ where θ_w and θ_b are parameter sets (of both regression coefficients and variance components) to be estimated at the within- and between levels, respectively. The parameters are solved to maximize a weighted likelihood of the two functions, where a weighted likelihood for the jth cluster (or school) in the kth PSU can be found as

$$l(\theta_w, \theta_b)_{jk} = \int \left(\prod_{i=1}^{n_j} f(y_{ijk}|x_{ijk}, \gamma_{jk}, \theta_w) \right)^{w_{ijk}} \phi(\gamma_{jk}|x_{jk}, \theta_b) d\gamma_{jk}. \quad (1)$$

And the total weighted likelihood across clusters and PSUs is taken as the product

$$l(\theta_w, \theta_b) = \prod_{jk} l(\theta_{w,} \theta_b)_{jk}^{w_{jk}}. \quad (2)$$

See Jenkins (2008) and Pfeffermann et al. (1998) for more detailed explanation of the estimation. The estimations can be altered to include stratification at the first stage of sampling by taking the product in equation (2) across each PSU k within each stratum s.

As an aside, note that two sampling weights, one at the individual level in equation (1), w_{ijk}, and one at the cluster (school) level in equation (2), w_{jk}, appear as exponents in this estimation. Researchers have suggested that MPML-based analyses use conditional sampling weights within clusters at level 1, $w_{i|jk}$ (the inverse of the selection probability of the individual given

selection of the cluster), and the inverse of the selection probability of the cluster as the sampling weight, w_{jk}, at level 2 (Asparouhov and Muthén 2006; Rabe-Hesketh and Skrondal 2006). Current versions of Mplus (Muthén and Muthén 2011), HLM (Raudenbush et al. 2011), and Stata software's generalized linear latent and mixed models (*gllamm*; Rabe-Hesketh and Skrondal 2006) package are able to appropriately include these disproportionate sampling rates at both levels of the model, if those weights are provided with the data set. In all analyses reported in this article, appropriate sampling weights, as specified above, are used.

To estimate sampling variances with the MPML estimation, the asymptotic covariance matrix of the $\hat{\theta}_w$ and $\hat{\theta}_b$ parameter sets is a sandwich estimator of

$$\text{cov}(\hat{\theta}) = I^{-1} V I^{-1}, \tag{3}$$

where I is the observed pseudo-Fisher information at the MPML estimates of $\hat{\theta}$ (the second derivative of the log of the total weighted likelihood) and V is a covariance estimate that is calculated based on the sample design. Rabe-Hesketh and Skrondal (2006) present an estimator V based on an assumed with-replacement design at the first stage of sampling, while Asparouhov and Muthén (2006) provide two additional estimates of V for without-replacement designs,[3] incorporating finite population correction factors and unequal probability of selection of PSUs in the design. Although these latter two new estimators are available starting in M*plus* with version 6, the analyst must provide sampling design information not present in the public release data set, and therefore the with-replacement estimator is more typically used and is the focus of this article. Both the Stata *gllamm* package and M*plus* provide the with-replacement assumed MPML estimator. In this estimation, V represents an estimate of the variance in parameter estimates across PSUs within strata, summed across strata. As PSUs within strata become more similar in their estimates, elements of this V matrix decrease in size, yielding adjustment to the sampling variance estimates.

While this robust MPML estimation with the sandwich estimator should be the preferred approach when undertaking multilevel analyses with stratified and/or three-stage sampling designs, it is of interest to determine the effect of ignoring these sampling design elements with NCES data. First, decades of multilevel research using data from national probability samples has been published that have ignored these for the most part and it would be of interest to determine the extent to which those standard error estimates might be biased. Secondly, current multilevel researchers may be limited in

their access to multilevel software that can accommodate the sampling design. Prior simulation research is not completely informative on this matter for the typical analyst using NCES data. First, two of the three simulation studies that compared the MPML estimator with robust sampling variance estimation to an approach of ignoring the sampling design used only two stages of sampling (Grilli and Pratesi 2004; Rabe-Hesketh and Skrondal 2006). These studies, therefore, did not compare the ability of the sandwich estimator to account for the missing level of sampling in the sampling variance estimates to the approach of just ignoring the missing level of sampling. Only Asparouhov and Muthén (2006) investigated two-level estimates under conditions involving three stages of sampling. Second, stratification at the first stage of sampling in these simulations involved only two or three strata and contained many PSUs per stratum (in one study 200 PSUs were in one of the strata). Typical education-related data sets utilize dozens of strata within sampling designs with few PSUs within each stratum. A third problem with these studies is that to create the informativeness of the stratified sampling, the researchers split cluster observations into strata using a cut point based on the generated residuals, resulting in extreme informativeness of the stratification variable. For example, clusters with a negative residual were placed in stratum 1 and those with positive residuals were placed in stratum 2. In the Rabe-Hesketh and Skrondal (2006) study, the strata variable was correlated to the response variable at 0.82 at the first stage of sampling and 0.76 at the second stage of sampling. Levels of strata informativeness could not be ascertained from the other studies (Asparouhov and Muthén 2006; Grilli and Pratesi 2004) but given the described data generation logic, they are expected to be similar to that used by Rabe-Hesketh and Skrondal (2006). In order to understand whether the findings from these simulation studies can be generalized to current applied research, an empirical evaluation of currently available national probability sample data is needed to determine the typical level of informativeness of the stratification.

In this article, we provide a review of empirical data to determine whether these simulation studies provide realistic and generalizable results by examining stratification and clustering informativeness for a broad range of variables from each of five NCES data sets. If research can provide support that ignoring the first-stage sampling design in weighted multilevel analyses of three-stage data or ignoring stratification can provide unbiased sampling variances under the realistic conditions found in most national databases, we can have more confidence in the two-level analysis results that have been published from these data sets. Therefore, in this article, we review

informativeness indices as well as the unconditional and conditional multilevel design effects present in empirical NCES data sets.

A Review of the Empirical Data

In this section, we first briefly describe the sampling structure of the five data sets of interest and highlight the portion of the sampling structure not accommodated by a simple weighted two-level analytic model. Second, we describe the empirical research we reviewed using these data sets and present the most often-used variables included in the published analyses. We conducted a review of the data characteristics of the following five existing public release data sets: Early Childhood Longitudinal Study-Kindergarten of 1998-99 (ECLS-K: Tourangeau et al. 2009), Education Longitudinal Study of 2002 (ELS; Ingels et al. 2005), National Education Longitudinal Study: 1988 (NELS; Spencer et al. 1990), Schools and Staffing Survey of 1999-2000 with Teacher Follow-up Study of 2000-01 (SASS-TFS; Tourkin et al. 2004), and the Trends in International Mathematics and Science Study 1999 (TIMSS; Martin, Gregory, and Stemler 2000).

Summary of documentation of sampling structures. In Table 1, we provide a summary of the sampling structure for each of the five data sets of interest and a more detailed description of the sampling structure for each of the five data sets reviewed is presented in Online Appendix 1. The details provided in the Online Appendix include information about the type of sampling used (e.g., probability proportionate to size), whether disproportionate sampling was utilized, target sample size at each level, and, importantly, the variables used for both explicit and implicit stratification at each level of the sampling plan. In the five data sets that we examined, only three involved some degree of three-stage sampling. TIMSS had the most complicated sampling structure, with selection of PSUs within regional strata, followed by schools within the PSUs and then selection of intact classrooms. For ECLS-K, about one-third of the schools were selected after first selecting a geographic area. For SASS-TFS, private schools were selected after selection of geographic areas. For the remaining data sets and subsets of data sets, the sampling designs suggest that the school-level data should not exhibit dependency due to clustering, given that schools were at the first stage of selection and therefore standard errors from a multilevel analysis would not be expected to be underestimated as is commonly a concern. On the contrary, for many of the analyses from these data sets, the only worry is that the stratification used at the first stage of sampling may not be accommodated in a two-level

Table 1. Summary of Sampling Designs Used With Five Data Sets of Interest.

	Design	First-stage Strata	PSUs	SSU	USU
ECLSK	Stratified three stage	Geo- and demographic grouping of counties	Counties/ groups of counties	Schools	Kindergarten students
ELS	Stratified two stage	Geographic and sector groupings of schools	Schools	—	10th-grade students
NELS	Stratified two stage	Geo- and demographic groupings of schools	Schools	—	8th-grade students
SASS-TFS	Stratified two stage (public) and three stage (private)	Sector and geographic groupings of schools (public) and geographic groupings of counties (private)	Schools (public) and counties (private)	Schools (private)	Teachers
TIMSS	Stratified three stage	Geographic groupings of districts/regions	School districts/ regions	Schools	Classroom (intact)

Note: — indicates that the stage of sampling was not applicable. PSU = primary sampling unit; SSU = secondary sampling unit; USU = ultimate sampling unit; ECLSK = Early Childhood Longitudinal Study-Kindergarten; ELS = Education Longitudinal Study; NELS = National Education Longitudinal Study; SASS-TFS = Schools and Staffing Survey–Teacher Follow-up Study; TIMSS = Trends in International Mathematics and Science Study.

weighted analysis, and therefore there would be a loss in precision, represented by overestimated standard errors. The degree of this overestimation is of interest as we document the empirical data characteristics in the Results section.

Literature search of published articles and description of empirical data. We conducted a literature search of EBSCO, PsychINFO, ERIC, JSTOR, and Google Scholar to examine articles appearing in peer-reviewed journals that utilized the five public release data sets of interest: ECLS-K, ELS, NELS, SASS-TFS, and TIMSS. Search key words included the following: *ECLS-K, NELS 88, SASS-TFS, ELS 2002, TIMSS*. Articles that were not related to applied research, such as letters to the editor or book reviews, and

Table 2. Number of Articles Reviewed and Included in Review for Five Selected Public Release Data Files.

	Total Number of Articles	Total Number of Multilevel	Longitudinal	Contextual[a]
ECLSK (1998)	134	54	35	19
ELS (2002)	47	10	0	10
NELS (1988)	314	52	12	40
SASS-TFS (1999–2000)	15	4	—	4
TIMSS (1999)	7	4	—	4

Note: ECLSK = Early Childhood Longitudinal Study-Kindergarten; ELS = Education Longitudinal Study; NELS = National Education Longitudinal Study; SASS-TFS = Schools and Staffing Survey–Teacher Follow-up Study; TIMSS = Trends in International Mathematics and Science Study.
[a]Only the articles including contextual analyses were included in the analyses in the article.

unpublished manuscripts, such as papers given at conferences, organizational or agency reports, and dissertations, were not included in this review. First, the articles were sorted based on public release data set used and then it was determined whether the procedure used in each article was a form of multilevel modeling. Of those articles that used multilevel modeling, it was established whether the modeling used was longitudinal (e.g., growth curve for individual change) or school and community contextual analysis (e.g., multilevel regression and multilevel structural equation modeling). Table 2 contains the number of articles identified as using multilevel regression or multilevel SEM for each of the five public release data sets.

We then reviewed the contextual multilevel analyses published from each of the five data sets to identify the most often-used measures in the analyses. Sometimes measures were constructed as composites or scales from several items but with too little information to replicate to include in analyses and were therefore eliminated from consideration. Table 3 lists the variables that were most frequently used for each of the data sets (typically, used in a majority of the analyses). Given their use in this empirical literature, we evaluate some data sets for up to six dependent variables (ELS) and some for only one (TIMSS and SASS-TFS). Most dependent variables were interval-scaled exam scores or latent trait scores, however, some dependent variables were binary indicators, such as drop-out status. The breadth of independent variables examined for any data set was a function of their frequency of use in the published literature. In Table 3, we have included details on how variables were coded or transformed, if applicable. Of note, it

Table 3. Most Frequently Used Variables in Published Analyses.

ECLS-K		ELS	NELS	SASS-TFS	TIMSS
Dependent variables					
	Spring Math IRT scale score	Math IRT estimated number correct	Eighth grade math standardized score	Left teaching (1/0)	National Math Rasch score
	Spring Reading IRT scale score	Math standardized score	Eighth grade Reading standardized score		
	Special education status (1/0)	Postsecondary attendance status (1/0)	Engagement (summated score of three items)		
		Postsecondary level (three categories)	Drop out (1/0)		
		Violence victimization (1/0)			
		Property victimization (1/0)			
Independent variables					
Level 1	SES (continuous)	SES (continuous)	SES (continuous)	Salary (four ordinal categories)	Father's education level
	Female	Female	Female	Female	Mother's education level
	Race (four race/ethnic categories)	Race (Four race/ethnic categories)	Race (Four race/ethnic categories)	White (1/0)	Computer in the home (1/0)
	Age (continuous)	Whether English is spoken in home (1/0)	GPA (continuous)	Years in teaching (continuous)	Number of books in home (five categories, treated as continuous)

(continued)

Table 3. (continued)

	ECLS-K	ELS	NELS	SASS-TFS	TIMSS
	Fall Reading IRT scale score		Two parent in home status (1/0)	Master's degree (1/0)	Expectations (sum of two Likert-type items; continuous)
	Fall Math IRT scale score			Assignment type (three categories—full/part/other)	
Level 2	Kindergarten enrollment (continuous)	Enrollment of school (seven categories)	Enrollment of school (continuous)	Level of school (three categories)	Absence (proportion of students absent on typical day)
	Minority status of school (five categories)	School sector (three categories)	Percentage of minority of school (continuous)	Urbanicity (three categories)	
	School sector (four categories)	School urbanicity (three categories)	School sector (three categories)		
			School location (three categories)		

Note: ECLSK = Early Childhood Longitudinal Study-Kindergarten; ELS = Education Longitudinal Study; NELS = National Education Longitudinal Study; SASS-TFS = Schools and Staffing Survey–Teacher Follow-up Study; TIMSS = Trends in International Mathematics and Science Study; IRT = Item Response Theory; SES = socioeconomic status; GPA = grade point average.

Table 4. Structure of Data Used for Empirical Analyses.

		Strata	PSUs	Schools	Students or Teachers[a]
ECLSK	Frequency	88	427	787	12,678
	Range within unit above		1–64	1–12	1–25
ELS	Frequency	361	—	751	15,244
	Range within unit above		—	2-3	2–50
NELS	Frequency	28	—	1,011	16,489
	Range within unit above		—	1–149	1–49
SASS-TFS	Frequency	NA	—	7,959	37,974
	Range within unit above		—	NA	1–19
TIMSS	Frequency	53	106	221	9,072
	Range within unit above		2–2	1–8	3–78

Note: NA indicates stratum information for the SASS-TFS data is not on the public release data file. — indicates that schools were directly sampled within strata and not sampled within PSUs. ECLSK = Early Childhood Longitudinal Study-Kindergarten; ELS = Education Longitudinal Study; NELS = National Education Longitudinal Study; SASS-TFS = Schools and Staffing Survey–Teacher Follow-up Study; TIMSS = Trends in International Mathematics and Science Study.
[a]For SASS-TFS, level-1 units are teachers; for all other data sets, level-1 units are students.

was not uncommon for level-2 variables to represent some of the explicit stratification variables. Inclusion of these variables in a model would reflect a partial model-based approach of accommodating that aspect of the sampling design. It is these often-used variables for multilevel models with these selected data sets that we examine in the analyses in this article. Specifically, we extracted 91 variables from the databases for review (24, 23, 25, 9, and 10 variables, respectively, for ECLS-K, ELS, NELS, SAS, and TIMSS) as listed in Table 4. Some of the variables extracted were dummy coded versions of nominally scaled data, and some were school means of level-1 variables. Fifteen of the 91 variables are dependent variables and are the focus of the analysis in this article; the remaining variables are used as predictors to evaluate design effects in conditional models.

In summary, our review of the five public release data sets suggests that, for three of the data sets (NELS, ELS, and the public school sample for SASS), the only concern in running a contextual two-level analysis of students within schools is that the first-stage selection of schools was stratified. If the strata variables are not accommodated in an analysis, the sampling variances may be overestimated. For the remaining data sets, a researcher may need to consider whether a two-level contextual analysis of students/teachers within schools should address the effects of both first-stage selection of geographic areas as well as stratification. Our review also provided several

candidate variables to include in our analyses to determine empirically the effect of addressing stratification and first-stage selection. Based on this review, in this article, we examine these empirical data to determine whether the sampling design is informative for these variables and to evaluate possible effects of ignoring the sampling design in weighted two-level analyses on estimates of sampling variance.

Method

For each of the five data sets, we conduct several analyses to determine the effect of ignoring the stratification and first-stage components of the sampling design. Subsets from the original databases were created in some cases. For example, for ECLS-K, because most published analyses included the categorical race/ethnicity variable and excluded students who affiliated with "Other" or multiracial categories that same procedure was used in our analyses. Furthermore, missing data on some variables were accommodated using listwise deletion; although this is not a wise approach for empirical analyses, our interest was in estimating the effect on the standard errors of modeling sampling information inappropriately, and therefore our interest is not in the point estimates of the parameters themselves. Therefore, a single data set using listwise deletion was created for each of the five databases and all analyses used these final data sets and were thus based on the same number of observations for a given survey program.[4] For context, Table 4 includes the counts of the observations used at each level of the analysis for the five data sets as well as information about the variance estimation strata and the PSUs. The SASS-TFS analysis was limited to public schools and therefore did not involve PSUs in the sampling structure.

The combined effect of ignoring the clustering and stratification in the sampling design on all estimates in a multilevel model was evaluated empirically in two ways, using unconditional and conditional models. Specifically, we calculated design effects for parameter estimates from both univariate and multivariate analyses for each of 15 variables of interest across the five data sets; these 15 were chosen as they were typically used as dependent variables in the published multilevel model results reviewed. The design effect is used as a measure of the over- or underestimation of the sampling variance of a specific parameter estimate. It is the ratio of the actual variance of an estimate to the variance of that estimate given a simple random sample of the same number of elements (Kish 1965). Typically, the design effect of the *mean* is reported in database user guides for a variety of variables. The square root of the design effect, named *root design effect* or referred to

simply as *deft*, can be used as a multiplicative standard error adjustment therefore deft values of 1.0 suggest no standard error adjustment is needed while values above 1.0 suggest that the standard error might be underestimated and those below 1.0 suggest that the standard error might be overestimated (Kish 1965).

In order to estimate multilevel design effects found for typical education-related data, we undertook empirical analyses for all 15 dependent variables, all conducted in M*plus* version 6.0 using maximum likelihood estimation. First, we determined a *univariate* multilevel design effect and root design effect of ignoring the additional sampling design in weighted two-level analyses by estimating a null (unconditional) model, shown in equation (4), once with the MPML estimator[5] with explicit first-stage strata and, if applicable, PSU (level 3) clustering identified and once with the traditional maximum likelihood estimator, ignoring stratification and any level-3 clustering:

$$y_{ij} = \gamma_{00} + u_{0j} + r_{ij}. \tag{4}$$

For continuous outcomes, a hierarchical linear model was used, as in equation (4). From these results, we obtained design effects for the estimate of the intercept, the within-school residual variance and the between-school variance. For dichotomous and polytomous outcomes, the model was run using maximum likelihood estimation with a logit link and design effects were obtained for the intercept and between-school variance. All estimation used appropriate sampling weights at each level of the analysis, except the analyses using the NELS data for which a school-level weight is not provided on the public release data file. For the NELS data, the level-2 analyses were unweighted, and the overall unconditional sampling weight was used at level 1.

The multilevel design effect of the intercept can be estimated as a ratio of the estimate of the sampling variance of the intercept from the two estimations, where the prime indicates the estimate is from the properly specified MPML estimation:

$$\widehat{deff} = \frac{\hat{\sigma}^{2'}_{\gamma 00}}{\hat{\sigma}^{2}_{\gamma 00}}. \tag{5}$$

We also calculated design effects for the two other parameters in the unconditional model—within-school variance (σ^2) and between-school variance (τ_{00})—by taking the ratio of the two sampling variance estimates from the MPML and the ML estimations. The root design effect, *deft*, was calculated as the square root of the *deff* estimate in equation (5).

Next, we determined conditional design effects for all estimates from a fixed effects regression model with the 15 selected dependent variables regressed on all level-1 and level-2 selected predictors from Table 3 as shown in equation (6) for the continuous dependent variable case

$$y_{ij} = \gamma_{00} + \sum_{p=1}^{P} \gamma_{p0} X_{ijp} + \sum_{z=1}^{Z} \gamma_{0z} W_{jz} + u_{0j} + r_{ij}, \quad (6)$$

where P is the number of predictors at level 1 and Z is the number of predictors at level 2. As with the null model, we estimated design effects for the intercept and the two variance components; additionally, the design effect for each regression coefficient was calculated and the average and range of these root design effects are reported for each data set.

Results

In Table 5, we report the square root of the unconditional model multilevel design effects (*deft*) for the 15 dependent variables from the five data sets, displayed by parameter. These design effects were obtained based on running a multilevel model assuming simple random sampling at each level of the analysis as compared to a multilevel model that accommodates first-stage stratification and clustering. These root design effects reflect the needed inflation (or deflation) of the standard error estimated while ignoring the first stage sampling design to represent the appropriate sampling variability. Root design effects less than 1.0 indicate that the simple random sampling (SRS)-assumed standard error is overestimated and, conversely, root design effects greater than 1.0 indicate that the SRS-assumed standard error is underestimated.

The two-level models assuming SRS for the ELS, NELS, and SAS data sets ignored the informativeness in the stratification in the sampling design but because there was not a third stage of sampling, the two-level model appropriately accommodated the multistage sampling. Therefore, as expected, the estimates in Table 5 show that for these three data sets precision was lost and all standard errors were overestimated. For the unconditional model, the *deft* values for the ELS, NELS, and SASS data sets were all less than 1.0. The SASS-TFS dependent variable showed the greatest overestimation in the standard error, at about 30 percent for the intercept (the SRS-assumed intercept standard error estimate should be multiplied by .721 to obtain a more appropriate estimate of the intercept standard error). For the NELS data set, the overestimation occurred mainly at level 2, with the

Table 5. Multilevel Root Design Effects (Deft) of Standard Errors for Unconditional Model Parameter Estimates.

Parameter	Database	Number of Dependent Variables Analyzed	Average	Minimum	Maximum
γ_{00}	ECLS-K	3	1.008	.988	1.025
	ELS	6	0.973	.947	0.992
	NELS	4	0.906	.841	0.986
	SASS-TFS	1	0.721	—	—
	TIMSS	1	0.860	—	—
τ_{00}	ECLS-K	3	0.962	.932	0.986
	ELS	6	0.949	.870	1.000
	NELS	4	0.919	.833	0.997
	SASS-TFS	1	0.794	—	—
	TIMSS	1	1.102	—	—
σ^2	ECLS-K	3	0.970	.892	1.048
	ELS	6	0.964	.961	0.966
	NELS	4	0.974	.967	0.985
	SASS-TFS	1	[a]	—	—
	TIMSS	1	1.033	—	—

Note: — Because only one dependent variable was examined for TIMSS and SASS-TFS, minimum and maximum values are not provided. ECLSK = Early Childhood Longitudinal Study-Kindergarten; ELS = Education Longitudinal Study; NELS = National Education Longitudinal Study; SASS-TFS = Schools and Staffing Survey–Teacher Follow-up Study; TIMSS = Trends in International Mathematics and Science Study.

[a]The dependent variable modeled with SASS-TFS data was dichotomous, and therefore no level-1 residual variance is estimated.

standard errors for the intercept and between-school variances needing to be deflated by up to 17 percent. The within-school variance standard errors showed little overestimation. For the ELS data set, the overestimation in standard errors was fairly minor across all three types of parameter estimates on average, with most overestimation occurring for the standard error of the between-school variance estimate.

In the unconditional models run with the ECLS-K and the TIMSS data, because we were ignoring *both* first-stage selection of PSUs and stratification of PSUs, standard errors might have been over- or under-estimated.[6] As shown in Table 5, for the ECLS-K data, the *deft* measures for the intercept and the estimate of within-school variance were found to be both above and below 1.0 depending on the dependent variable examined. In general, the *deft* values showed minimal departure from 1.0, except in the case for one within-school variance measure. For the dependent variable Spring Kindergarten Reading Item Response Theory scale

Table 6. Multilevel Root Design Effects (Deft) of Standard Errors for Conditional Model Parameter Estimates.

Parameter	Database	Average	Minimum	Maximum
γ_{00}	ECLS-K	1.002	0.995	1.010
	ELS	1.049	1.009	1.072
	NELS	1.001	0.997	1.005
	SASS-TFS	0.877	—	—
	TIMSS	1.094	—	—
Level-1 fixed slopes (γ_{10}, γ_{20}, γ_{30}, γ_{40}, etc.)	ECLS-K	1.021	0.912	1.300
	ELS	0.996	0.945	1.055
	NELS	0.998	0.987	1.007
	SASS-TFS	0.872	0.828	0.926
	TIMSS	1.050	0.893	1.184
Level-2 intercept coefficients (γ_{01}, γ_{02}, γ_{03}, γ_{04}, etc.)	ECLS-K	0.984	0.884	1.030
	ELS	1.017	0.953	1.093
	NELS	0.995	0.916	1.014
	SASS-TFS	0.911	0.874	0.958
	TIMSS	1.040	0.798	1.178
τ_{00}	ECLS-K	0.977	0.916	1.024
	ELS	0.982	0.953	1.033
	NELS	0.995	0.986	1.000
	SASS-TFS	0.793	—	—
	TIMSS	1.032	—	—
σ^2	ECLS-K	0.969	0.930	0.998
	ELS	0.949	0.945	0.954
	NELS	0.986	0.982	0.993
	SASS-TFS	[a]	—	—
	TIMSS	1.034	—	—

Note: — Because only one dependent variable was examined for TIMSS and SASS-TFS, minimum and maximum values are not provided. ECLSK = Early Childhood Longitudinal Study-Kindergarten; ELS = Education Longitudinal Study; NELS = National Education Longitudinal Study; SASS-TFS = Schools and Staffing Survey–Teacher Follow-up Study; TIMSS = Trends in International Mathematics and Science Study.

[a]The dependent variable modeled with SASS-TFS data was dichotomous, and therefore no level-1 residual variance is estimated.

score, the *deft* was .892 indicating that the standard error of the within-school residual variance was overestimated when ignoring the first-stage sampling stratification by about 10 percent. Finally, for the one dependent variable that we examined from the TIMSS data set, the intercept standard error was overestimated and needed to be deflated by a factor of .86 while the variance component standard errors were somewhat underestimated.

Turning to the conditional models, the multilevel root design effects are presented in Table 6 for the fixed slope coefficient estimates as well as the intercept and two variance components. Across all data sets, with few exceptions, the *deft* values for the intercept and between-school variances were closer to a value of 1.0 as compared to the unconditional model results. Because level-2 predictor variables tended to include measures used to define strata or could serve as proxies of those measures, there was basically no loss in precision of the estimates when running an analysis without the MPML estimator and therefore the *deft* estimates increased toward 1.0. This conclusion cannot be made definitively, however, given confounds associated with also including level-1 predictor variables in our conditional models. Of interest is that the *deft* values for the parameter estimates based on the SASS-TFS outcome of interest were consistently less than 1.0 and relatively low compared to the estimates for the other data sets. The likely reason is that the popular level-2 predictors included in the model for SASS-TFS (see Table 3), included only one of the variables used in stratification of the sample: enrollment level of the school. The primary stratification variables as defined in the Online Appendix, state and district, were not included in the model and therefore precision in the estimation of the intercept could be expected to be lost. *Deft* estimates for the level-1 variance standard errors across all data sets were very similar in the conditional and unconditional models.

Discussion

This article presents an empirical investigation of the effects of ignoring stratification and first-stage selection in weighted two-level analyses with selected data from the NCES. The findings suggest that the standard errors of parameters in unconditional models might be over- or underestimated, depending on whether the ignored sampling components included stratification at the first stage of sampling or an additional stage of sampling that was not accommodated. In general, given the variables used in this study, the misestimation of the standard errors was not as extreme as presented in prior simulation research (e.g. Asparouhov and Muthén 2006; Rabe-Hesketh and Skrondal 2006). Importantly, the standard error estimates were improved with conditional models, where the conditional models in our examples included fixed effects of stratification variables at level 2. Given these empirical findings, we suggest that inferences from the published multilevel applied research that has been conducted using these public release data files are likely robust even though the more advanced newly available MPML

estimators were not implemented. Note, however, that our findings are only generalizable to the data sets and variables examined here.

Given our findings, we suggest possible steps that applied researchers who do not have access to appropriate estimators might follow to evaluate the extent to which their weighted two-level analyses might be affected by ignored elements of the sampling design. First, it is crucial to evaluate the sampling design used to obtain the data. By understanding the elements of the design, it will be clear what components are not being addressed by a weighted two-level model. A thorough reading of the database user's guide is essential. We strongly encourage researchers to examine descriptive statistics in the data set, such as the number of strata and PSUs within strata. The researcher can then evaluate whether various sampling components can be ignored. As an example, the ECLS-K data collection is reported in publications as being a three-stage sampling design. By reading the user's guide (Tourangeau et al. 2009) in detail and examining the data, it becomes clear that for two-thirds of the data, it is based on a two-stage sampling design so the fact that some schools are clustered in PSUs becomes more of a minor concern.

Second, if stratification is used at the first stage of sampling, the researcher might calculate the informativeness of the stratification for the level-2 cluster (e.g., school) means. Stratum informativeness represents the proportion of variance in the outcome variable that is associated with differences across strata. The informativeness index for stratification can be calculated as follows,

$$\hat{\rho}_{strat} = \frac{MS_{Bs} - MS_{Ws}}{MS_{Bs} + (c-1)MS_{Ws}} \tag{7}$$

where c is the average number of units (e.g., schools) per stratum and the MS_{Bs} (mean square between) and MS_{Ws} (mean square within) values are from a weighted ANOVA with first-stage *stratum* as the grouping factor.[7] This formula can also be used for data that are converted to binary indicators (e.g., dummy codes for nominal categories such as race/ethnicity; Snijders and Bosker 2012:304). To provide an example, we calculated this index for the TIMSS data set variable Math Standardized Score at the school mean level. At the school level, the components of MS_{Bs} and MS_{Ws} were 52.4 and 18.8, respectively, with a c value of 4.17 (there were approximately four schools per stratum on average). These estimates result in an informative index of .30 at the school level and suggest that there is a moderate amount of homogeneity of school average math scores within strata. The stratum informativeness index can then be translated into the amount of standard error

misestimation by calculating an approximate design effect of the school intercept (γ_{00}). When the sample design is based on stratification alone with proportional selection across strata, Kish (1995) shows that the estimated design effect of the mean is:

$$\widehat{deff} = 1 - \hat{\rho}_{strat} \tag{8}$$

Thus, the more homogenous the strata, the smaller the *deff* and the root design effect, *deft*, and therefore, the smaller the adjusted standard errors. Using this formula with our example, we can approximate that if a model was run ignoring the stratification (and assuming there were no additional stages of sampling), the estimated standard error for the school intercept in an unconditional model would overestimated and should be decreased by a factor of the square root of (1–.30), which is .84.

The sampling design for TIMSS, however, includes a stage of selection above selection of schools as well as the stratification which brings us to our third recommendation. To understand the possible impact of ignoring a first stage of selection, one can calculate an informativeness index for PSU clustering. PSU informativeness represents the proportion of variance in the outcome variable that is associated with the PSU clustering. This informativeness index for normally distributed continuous school means can be calculated as,

$$\hat{\rho}_{\text{PSU}} = \frac{MS_{Bp} - MS_{Wp}}{MS_{Bp} + (c-1)MS_{Wp}} \tag{9}$$

where c indicates the average number of level-2 units per PSU and the MS_{Bp} and MS_{Wp} values can obtained from a weighted ANOVA using *PSU* as the grouping factor (Snijders and Bosker:2012:sections 3.3 and 3.4, 17-24). This formula can also be used for data that were converted to binary indicators (e.g., dummy codes for nominal categories such as race/ethnicity; Snijders and Bosker 2012:304). To continue the example, we calculated this index for the TIMSS data set variable Math Standardized Score at the school mean level making the assumption that there was no stratification in the sampling design. At the school level, the components of MS_{Bp} and MS_{Wp} were 36.3 and 18.1, respectively, with a c value of 2.09 (there were approximately two schools per PSU on average). These estimates result in a PSU informative index of .326 at the school level and suggest that there is a moderate amount of homogeneity of school average math scores within PSUs. The PSU informativeness index can also be translated into the amount of standard error misestimation by

calculating an approximate design effect of the school intercept (γ_{00}). When the sample design is based on first-stage clustering alone (and without stratification), Kish (1995) shows that the estimated design effect of the mean is:

$$\widehat{deff} = 1 + (c - 1)\hat{\rho}_{\text{PSU}}. \tag{10}$$

Using our example, the resultant design effect estimate is 1.16, suggesting that standard errors should be inflated by 16 percent to appropriately capture the imprecision introduced by the first-stage sampling. There is currently no guideline to combine these estimates of design effects due to stratification and due to PSU clustering.[8] For the unconditional model intercept for the TIMSS math variable, the standard error may need to be deflated to address stratification and inflated to address clustering, but the applied analyst can assume that the true correction should be somewhere between .84 and 1.16. In fact, from Table 5, we see that the actual design effect of the intercept was 0.86. For those without access to multilevel software with the capacity to include additional sampling design considerations, using this strategy to determine the bounds of needed adjustment of the standard error may be helpful.

Fourth, although sampling weights were not the focus of this article, we suggest that researchers calculate the coefficient of variation of the sampling weight for the level-2 units as the standard deviation of the weight over the mean of the weight. While there is no convenient equation to translate this coefficient into a design effect (Kish 1995), the greater this value increases from 0, the more the standard errors at level 2 may be underestimated. In their chapter on sampling weights within multilevel models (chapter 14), Snijders and Bosker (2012) provide additional guidance in this area.

Finally, depending on the information obtained from the first three steps of the review, consider using two-level software that can accommodate hybrid aggregate–disaggregate analyses. If the design effects due to stratification or PSU clustering as calculated in step 2 are far from a value of 1.0, more confidence in the statistical inference from model estimates would be gained, if the sampling design were more properly accounted for. As of this writing, only Stata's *gllamm* package and M*plus* have the required capability. If the design effects are relatively close to 1.0, as in the empirical analyses included in this article, it may not be crucial to use the appropriate estimator. Inclusion of level-2 variables that were used in the explicit stratification process of the sampling design should be considered as they were shown in these analyses to improve the standard error estimates. Of course,

inclusion of additional variables should not be done if it detracts from the conceptual framework of the model.

The steps above are derived from both theory and empirical investigation. These suggestions should be evaluated with simulation methods. When we determined the design effects for the unconditional and conditional models, we made the assumption that the MPML standard error estimates were unbiased. Although simulation research has suggested this is the case (Asparouhov and Muthén 2006; Rabe-Hesketh and Skrondal 2006), the data conditions in these empirical analyses may not have matched the simulation conditions. A larger issue in ignoring the sampling design highlighted in Asparouhov and Muthén (2006), the overestimation of the likelihood ratio test value, leading to improper rejection of appropriate models, was not evaluated in this study. Future simulation research should verify the dire repercussions suggested in that article.

In this article, we sought to document the extent to which the exclusion of sampling design information, beyond the clustering of students or teachers in schools, would affect the inference made from two-level analysis models with NCES data. We found that there is little effect with the analyses and data sets used here. In fact, no differences in inference regarding the statistical significance of any individual parameter estimates would have been made in any of the analyses we conducted here.

Authors' Note

The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D110050 to the University of Maryland.

Supplemental Material

Supplementary material for this article is available online.

Notes

1. Probability proportional to size sampling involves selecting at higher probabilities those clusters that contain more lower-level units; such designs are often used when fixed numbers of elements are desired to be selected within clusters (Kish 1965).
2. Implicit stratification refers to systematic selection (every X sample frame unit) down a list ordered by the implicit stratification variable (Kish 1965).
3. With-replacement and without-replacement designs refer to whether, once selected into a sample, the unit is available for selection into the sample again (Kish 1965).
4. Note that we are assuming that any differences in estimates found when ignoring or fully accommodating the complex sampling design using listwise deletion will hold when using other types of missing data accommodation. It is not known if this assumption is plausible.
5. For the SASS-TFS data set, the estimation of standard errors was conducted somewhat differently given available information. See the SASS-TFS section in the Online Appendix for more information.
6. We also examined standard error estimates when accounting for only a portion of the sampling design for ECLS-K and for TIMSS and, as expected, the intercept standard error estimates had deft values less than 1.0 when the stratification was accounted for but the clustering was not, and values over 1.0 when the clustering was accounted for but the stratification was not.
7. This formula assumes a balanced design. An alternate calculation for c should be used if unbalanced: $c = \frac{N^2 - \sum_{j=1}^{J} n_j{}^2}{N(J-1)}$ where J references the number of clusters and N reflects the total sample size.
8. Recent work by Lohr (2014) derived the design effect for regression coefficients under conditions of PSU clustering only. Importantly, she considered the context of random slopes, an issue not addressed here.

References

Asparouhov, T. and B. Muthén. 2006. "Multilevel Modeling of Complex Survey Data." Pp. 2718-26 in *Proceedings of the American Statistical Association*. Seattle, WA: American Statistical Association.

du Toit, M. and S. du Toit. 2008. "Multilevel Structural Equation Modeling." Pp. 435-78 in *Handbook of Multilevel Analysis*, edited by J. de Leeuw and M. Meijer. New York: Springer.

Grilli, L. and M. Pratesi. 2004. "Weighted Estimation in Multilevel Ordinal and Binary Models in the Presence of Informative Sampling Designs." *Survey Methodology* 30:93-103.

Hox, J. J. 2002. *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ingels, S. J., D. J. Pratt, J. E. Rogers, P. H. Siegel, E. S. Stutts, and J. A. Owings. 2005. *Educational Longitudinal Study of 2002: Base-year to First Follow-up Data File Documentation*. Washington, DC: National Center for Education Statistics.

Jenkins, F. 2008. *Multilevel Analysis with Informative Weights*. Retrieved June 26, 2012 (http://www.amstat.org/sections/srms/proceedings/y2008/Files/301419.pdf).

Kalton, G. 1983. "Models in the Practice of Survey Sampling." *International Statistical Review* 51:175-88.

Kaplan, D. and P. R. Elliott. 1997. "A Didactic Example of Multilevel Structural Equation Modeling Applicable to the Study of Organizations." *Structural Equation Modeling: A Multidisciplinary Journal* 4:1-24.

Kish, L. 1965. *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kish, L. 1995. "Methods for Design Effects." *Journal of Official Statistics* 11:55-77.

Kish, L. and M. R. Frankel. 1974. "Inference from Complex Samples." *Journal of the Royal Statistical Society (B)* 36:1-37.

Kovacevic, M. S. and S. N. Rai. 2003. "A Pseudo Maximum Likelihood Approach to Multilevel Modeling of Survey Data." *Communications in Statistics* 32:103-21.

Lee, V. E., D. T. Burkham, D. D. Ready, J. Honigman, and S. J. Meisels. 2006. "Full-day Versus Half-day Kindergarten: In Which Program Do Children Learn More?" *American Journal of Education* 112:163-208.

Lohr, S. L. 2014. "Design Effects for a Regression Slope in a Cluster Sample." *Journal of Survey Statistics and Methodology* 2:97-125.

Martin, M. O., K. D. Gregory, and S. E. Stemler. 2000. *TIMSS 1999 Technical Report*. Chestnut Hill, MA: International Study Center.

Muthén, L. K. and B. O. Muthén. 2011. *Mplus User's Guide*. 6th ed. Los Angeles, CA: Muthén & Muthén.

O'Connell, A. and B. McCoach. 2008. *Multilevel Analysis of Educational Data*. Greenwich, CT: Information Age Publishing.

Palardy, G. J. 2008. "Differential School Effects among Low, Middle, and High Social Class Composition Schools: A Multiple Group, Multilevel Latent Growth Curve Analysis." *School Effectiveness and School Improvement* 19:21-49.

Pfeffermann, D., C. J. Skinner, D. J. Holmes, H. Goldstein, and J. Rasbash. 1998. "Weighting for Unequal Selection Probabilities in Multilevel Models." *Journal of the Royal Statistical Society, Series B* 60:23-40.

Rabe-Hesketh, S. and A. Skrondal. 2006. "Multilevel Modeling of Complex Survey Data." *Journal of the Royal Statistical Society, Series B* 60:23-56.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2004. "Generalized Multilevel Structural Equation Modelling." *Psychometrika* 69:167-90.

Rasbash, J., F. Steele, W. J. Browne, and H. Goldstein. 2012. *A User's Guide to MLwiN, v2.26*. Bristol, UK: Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S. W. and A. S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage.

Raudenbush, S. W., A. S. Bryk, Y. F. Cheong, R. T. Congdon, and M. du Toit. 2011. *HLM 7: Hierarchical Linear and Nonlinear Modeling*. Lincolnwood, IL: SSI Scientific Software International.

SAS Institute Inc. 2013. *SAS/STAT® 13.1 User's Guide*. Cary, NC: SAS Institute Inc.

Snijders, T. A. B. and R. J. Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. London, UK: Sage.

Spencer, B. D., M. R. Frankel, S. J. Ingles, K. A. Rasinski, R. Tourangeau, and J. A. Owings. 1990. *National Education Longitudinal Study of 1988: Base Year Sample Design Report*. Washington, DC: National Center for Education Statistics.

SPSS (Statistical Package for the Social Sciences). 2002. "Linear Mixed Effects Modeling in SPSS: An Introduction to the MIXED Procedure." Technical report LMEMWP-1002, SPSS, Inc., Chicago.

Stapleton, L. M. 2008. "Variance Estimation Using Replication Methods in Structural Equation Modeling with Complex Sample Data." *Structural Equation Modeling: A Multidisciplinary Journal* 15:183-210.

Sterba, S. K. 2009. "Alternative Model-based and Design-based Frameworks for Inference from Samples to Populations: From Polarization to Integration." *Multivariate Behavioral Research* 44:711-40.

Tourangeau, K., C. Nord, T. Le, A. G. Sorongon, M. Najarian, and E. G. Hausken. 2009. *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K): Combined User's Manual for the ECLS-K Eight-grade and K-8 Full Sample Data Files and Electronic Codebooks*. Washington, DC: National Center for Education Statistics.

Tourkin, S. C., K. W. Pugh, S. E. Fondelier, R. J. Parmer, C. Cole, B. Jackson, T. Warner, G. Weant, E. Walter, K. Gruber, and L. Zhao. 2004. *1999-2000 Schools and Staffing Survey (SASS) Data File User's Manual*. Washington, DC: National Center for Education Statistics.

Wolter, K. M. 1985. *Introduction to Variance Estimation*. New York: Springer-Verlag.

Author Biographies

Laura M. Stapleton is an associate professor in Measurement, Statistics, and Evaluation at the University of Maryland and serves as the associate director of the Research Branch of the Maryland State Longitudinal Data System Center. Prior to earning her PhD in Measurement, Statistics, and Evaluation from the University of

Maryland in 2001, she was an economist at the Bureau of Labor Statistics and, subsequently, conducted educational research at the American Association of State Colleges and Universities and as an associate director of institutional research at the University of Maryland.

Yoonjeong Kang is a psychometrician at American Institutes for Research in Washington, DC. Dr. Kang's research centers on methodological investigations and application in statistical modeling, with a focus on models and procedures associated with structural equation modeling, multilevel modeling, and item response theory.