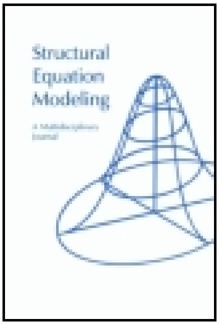


Structural Equation Modeling: A Multidisciplinary Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hsem20

An Assessment of Practical Solutions for Structural Equation Modeling with Complex Sample Data

Laura M. Stapleton

Published online: 19 Nov 2009.

To cite this article: Laura M. Stapleton (2006) An Assessment of Practical Solutions for Structural Equation Modeling with Complex Sample Data, Structural Equation Modeling: A Multidisciplinary Journal, 13:1, 28-58, DOI: 10.1207/s15328007sem1301_2

To link to this article: http://dx.doi.org/10.1207/s15328007sem1301_2

STRUCTURAL EQUATION MODELING, *13*(1), 28–58

An Assessment of Practical Solutions for Structural Equation Modeling with Complex Sample Data

Laura M. Stapleton
University of Maryland—Baltimore County

This article discusses 5 approaches that secondary researchers might use to obtain robust estimates in structural equation modeling analyses when using data that come from large survey programs. These survey programs usually collect data using complex sampling designs and estimates obtained from conventional analyses that ignore the dependencies in complex sample data may not be robust. The results from a simulation study that examined 5 methods of estimation under 6 types of sampling designs and different population conditions are shared and applied analysts are encouraged to consider using the pseudomaximum likelihood for linearization estimation of asymptotic covariance matrices currently available in some software programs for structural equation modeling analyses.

Secondary analyses of large scale datasets, such as those provided by the U.S. National Center for Education Statistics and the Centers for Disease Control and Prevention, are becoming more common and it is therefore important for researchers to understand the difficulties associated with modeling data that come from complex sample designs. Because most large-scale datasets are collected using sample designs other than simple random sampling (SRS), analyzing the data to make inferences to guide decisions can be a complicated endeavor. Tools for modeling these complex sample datasets using regression methods have been developed and analysts can use a range of methods available in such software as SUDAAN, WESVAR, and STATA and with new procedures in SAS and SPSS. However, some software programs for structural equation modeling (SEM) do not have estimation functions that accommodate complex sample data and others have only recently included estimation procedures for complex sample data. This article re-

Correspondence should be addressed to Laura M. Stapleton, Psychology Department, 1000 Hilltop Circle, University of Maryland–Baltimore County, Baltimore, MD 21250. E-mail: lstaplet@umbc.edu

ports on research that evaluated some accessible, practical approaches that secondary researchers can use when analyzing structural models with complex sample data.

Muthén and Satorra (1995) formally discussed the logic of SEM using complex sample data and outlined two approaches: aggregate and disaggregate (Skinner, Holt, & Smith, 1989). They refer to methods that assume SRS (and thus ignore the sampling design) as "conventional." For aggregate modeling (also called "design-based modeling" in the sampling literature), these authors propose a Taylor Series-like solution to variance estimation with estimates from a normal theory covariance matrix. Their Monte Carlo simulation appraisal of this method demonstrated its superiority over the conventional method on estimation of standard errors and chi-square statistics appropriate for inference to the finite population. The authors also discuss disaggregate (or "model-based") analysis in their work and researchers have now used these multilevel model-based SEM techniques to take into account clustering in the sample design (see e.g., Hox, 2002; Kaplan & Elliott, 1997). Many researchers may not be interested in research questions that address within-group and between-group relations separately, seeing the clustering as a nuisance and not part of the causal structure of the data. These researchers would like to make global statements about theoretical models of relations in the finite population. For these research questions, aggregate (design-based) analysis may be appropriate and is the focus of the research discussed in this article.

Methods to model aggregate SEM analyses using complex sample data have not been fully discussed and therefore the research described in this article is intended to provide guidance to researchers using national and international large-scale datasets regarding best methods of variance estimation when modeling data from complex sampling designs. This article briefly discusses the problems associated with the analysis of data that come from complex sampling designs, outlines possible approaches to estimate parameters and sampling variances, and describes results from a simulation study that examined five accessible methods of accommodating complex sample data in SEM analyses under selected conditions.

BACKGROUND

It is not common for large survey and testing programs to collect data based on a SRS design. Comprehensive reviews of the special characteristics of typical complex sample designs are provided by Longford (1995) and by E.S. Lee, Forthofer, and Lorimor (1989). Their treatments cover the topics of clustering, stratification, unequal probabilities of selection, and nonresponse and poststratification adjustment. This study concentrated on a design that incorporates some of these characteristics of a sampling design and, in describing effects and the simulation procedure, the data collection plan of the Early Childhood Longitudinal Study (ECLS) (U.S. Department of Education, 2001) is used as a model.

Briefly, the ECLS sampling design includes three stages of sampling: primary sampling units (PSUs) of single counties or groups of counties, schools within those counties, and students within the selected schools. At the first two stages of sampling, stratification was used in addition to probability proportional to size (PPS) sampling. At the final stage (selection of students), stratification was used with disproportionate sampling across two strata. The first stage stratification was fairly complex, with 24 certainty strata (strata for which a single PSU represents each stratum and thus these 24 PSUs are automatically selected in the sample) and 38 noncertainty strata. These strata were defined by a combination of Census Metropolitan Statistical Area, Census region, proportion of the population of a specific race or ethnicity, size of the PSU, and average per capita income of the PSU. Within the 38 noncertainty strata, PPS sampling was used with paired selection to identify two PSUs per stratum (resulting in a total of 100 PSUs across all certainty and noncertainty strata). Within each of the 100 PSUs, schools were stratified by public or private status and then were sampled with further implicit stratification by size of school and proportion of Asian or Pacific Islander (API) students. Finally, within the 1,280 sampled schools, students were selected using two strata: API students and non-API students. API students were selected at a rate three times greater than non-API students (when population numbers allowed) for a target size of 24 students per school.

A review of some other popular datasets from the National Center for Educational Statistics (NCES) provides examples of sampling designs similar in complexity to the ECLS sampling design. The National Educational Longitudinal Study included a two-stage sampling design, with the selection of schools stratified by urban location, minority enrollment, and public or private status, and then the selection of students within those schools with unequal probabilities across strata (R. Lee, 1990). The Beginning Postsecondary Student Study included two-stage sampling as well, with higher education institutions selected disproportionately from nine strata defined by institution type (Wine, Heuer, Wheeless, Francis, & Dudley, 2002). Additionally, institutions were selected using implicit stratification, a technique that was used to ensure adequate representation across region and size categories. Students within sampled institutions were selected using stratified systematic sampling, with strata defined by student level. A last example dataset, the National Study of Postsecondary Faculty, was collected using a two-stage sampling design. Institutions of higher education were first sampled using PPS sampling within 15 strata defined by institution type (Selfa et al., 1997). Within institutions, faculty members were split into five strata based on race or ethnicity, gender, and teaching discipline and sampled with disproportionate probabilities. All of these large dataset examples include clustering within PSUs as well as stratification and unequal probability of selection at all levels of sampling. To understand the statistical considerations in obtaining parameter estimates and their sampling variance estimates for complex sample designs, the issues of clustering, stratification, and unequal probability of selection are briefly addressed next.

In conventional SEM (and other traditional analysis techniques), an assumption is made that observations are independent and identically distributed. Due to the use of multistage sampling, however, data that come from these complex sampling designs usually have some degree of dependence among observations (E.S. Lee et al., 1989; Skinner et al., 1989). For example, with ECLS, it can be expected that children from the same school will have some similar characteristics. Because traditional estimation of standard errors assumes that the correlation of the errors across individuals is zero, a researcher using clustered data may underestimate the standard error. Any underestimation would subsequently result in inflated Type I error rates (Kish & Frankel, 1974; E.S. Lee et al., 1989). In an SEM context, Muthén and Satorra (1995) demonstrated the standard error bias and effects on the chi-square value that can occur when conventional SEM is applied to data from cluster samples. Analysis of clustered data with conventional SEM, therefore, may lead to the improper rejection of structural models for the finite population and the inappropriate assertion of statistically significant relations where only random covariation exists.

One of the sampling design issues that has received less attention in the methodological research of modeling data from complex sample designs is that of stratification. Each of the example datasets listed previously were sampled by first dividing the elements of the population into several strata. When sample designs include stratification, the stratification usually is intended, at least in part, to help provide more precise, or efficient, estimates of population parameters. Estimates of the sampling variance of statistics made under an assumption of SRS will not necessarily be correct; assuming that the data exhibit some level of homogeneity within strata, the sampling variance estimate based on formulas that assume SRS will tend to be positively biased (Kalton, 1983b; Kish, 1965). The use of stratification in the sampling design and the appropriate sampling variance estimation can result in more efficient estimators of population parameters but the appropriate calculation of sampling variances is difficult in practice. Simple formulas for linearized estimates are available for the sampling variance of estimates of population means and proportions; however, more complicated statistics are not accompanied by such simple solutions to sampling variance estimation (Kish & Frankel, 1974). Although one might hope that use of traditional statistics on data that were collected with cluster sampling (which tends to result in negatively biased standard error estimates) paired with stratification (which tends to result in positively biased standard error estimates) would result in canceled effects on the estimate of the standard error, in practice the negative effects of clustering tend to outweigh the effects of stratification. This phenomenon has not been demonstrated in SEM analyses and one of the goals of this research is to lend clarity to this issue.

A final concern addressed here is that the complex sampling designs that give rise to large-scale survey and testing data tend to include unequal probability of selection of elements, using such strategies as PPS selection of institutions and differential selection rates across strata. PPS sampling is often used in multistage sampling and involves assigning higher rates of selection to the larger clusters. This sampling is typically used when the ultimate number of elements to be chosen from each cluster is a constant (e.g., 24 students from each school for ECLS). With PPS, the overall probability of selection for each element will tend to be similar because the probability of an element in a small cluster being selected is the product of a small probability of selection of the cluster and a large conditional probability within the cluster, and the situation is the reverse for an element in a large cluster (Kalton, 1983b). Differential selection across strata tends to be used for a different purpose; it is often employed to ensure sufficient sample sizes for subgroup reporting. For example, with ECLS, special interest was on reporting efficient estimates for API students, and therefore those students were sampled at a rate three times higher than the rate of other students.

When analyses are undertaken ignoring unequal selection rates of the sampled elements, biased parameter estimates may result when the response variable(s) are correlated with the probability of selection. Using conventional SEM, Kaplan and Ferguson (1999) and Asparouhov (2005) demonstrate the parameter estimate bias that can result when unequal selection rates are ignored. Often, it is simple to incorporate unequal selection probabilities by including sampling design weights in an analysis; however, the inclusion of the weights themselves in the analysis can again result in negatively biased standard error estimates of parameters (Longford, 1995; Potthoff, Woodbury, & Manton, 1992). The use of nonoptimal weighting results in a decrease in information and can be represented as a decrease in the "effective sample size."

Accommodating Complex Sample Data

Given that data do not come from a simple random sample, what are the options available to an analyst desiring to undertake an aggregate SEM analysis? Various general analytic strategies have been developed for estimating parameters and their sampling variances when undertaking design-based analyses with complex sample data. These strategies range from simple adjustments of standard errors that are obtained from conventional analyses to advanced statistical approaches but these advanced techniques have been primarily designed simpler statistics: means, ratios, and regression coefficients. The methods that are the most often suggested to estimate sampling variances are Taylor Series linearized estimates or replication methods, such as jackknife repeated replications, balanced repeated replications, and bootstrapping (Kish & Frankel, 1974; E.S. Lee et al., 1989). Currently, only one of these options, linearized estimation, has been extended from regression analysis to SEM analysis and this programmatic option for modeling with complex sample data is described in a later section.

The first and most simple method to obtain estimates of standard errors that have been corrected for the complex sampling design is to adjust the estimates taken from a conventional analysis using an inflation factor—the design effect (Kish, 1965). The design effect can be viewed as the expected effect on sampling variance estimates due to the sampling design; it is the ratio of the correct sampling variance of a statistic under the complex sampling design to the sampling variance that would have been obtained had SRS been used (Kish, 1965). To adjust for the sampling design in any analysis a researcher could inflate each standard error estimate by the square root of the average design effect of the variables in the analysis (Kalton, 1977). NCES typically provides instructions in its technical manuals on how to undertake this procedure. The example here is taken from the manual for the 1988 National Educational Longitudinal Study (NELS):

> Researchers who do not have access to software for computing accurate estimates of standard errors can use the mean design effects presented in this report to approximate the standard errors of statistics based on the NELS:88 data. Design-corrected standard errors for a proportion can be estimated from the standard error computed using the formula for the standard error of a proportion based on a simple random sample and the appropriate mean root design effect (DEFT):
>
> $$\mathrm{SE} = \mathrm{DEFT} \times \mathrm{SQRT}(p(1 - p)/n)$$
>
> where p is the weighted proportion of respondents giving a particular response, n is the size of the sample, and DEFT is the mean root design effect. Similarly, the standard error of a mean can be estimated from the weighted variance of the individual scores and the appropriate mean DEFT:
>
> $$\mathrm{SE} = \mathrm{DEFT} \times \mathrm{SQRT}(\mathrm{Var}/n)$$
>
> where Var is the sample variance, n is the size of the sample, and DEFT is the mean root design effect. (U. S. Department of Education, 1996, p. 5-25)

Thus, given this approach, a researcher need only obtain parameter estimates and estimates of standard errors from a conventional analysis, determine the square root of the average design effect for the variables in the analysis, and multiply that root design effect by the standard error estimates to obtain more appropriate standard error estimates. This procedure, however, has been claimed to result in fairly conservative estimates of the sampling errors in more complex statistical procedures and can be rather unwieldy if used for complex models with many parameter estimates (Kish & Frankel, 1974). The NELS manual (U.S. Department of Education, 1996) cautions the reader that "more complex estimators show smaller design effects than simple estimators" (p. 5-25) and indicates that this approach can therefore provide conservative estimates of standard errors. An applied example using this procedure with an SEM analysis is provided by Fan (2001) who used this conservative procedure for standard error estimation in his latent growth model using the NELS:88 data. It should be noted that this procedure does not address the effect of the sample design on the chi-square statistic when using conventional SEM on complex sample data as demonstrated in Muthén and Satorra (1995). An option to adjust the chi-square value would be to manually calculate an adjusted test statistic by dividing the conventional chi-square by the average design effect of the dependent variables in the analysis.

A similar approach to the manual design-effect adjustment, and one advocated in some NCES training sessions for analysts who do not have access to specialized software (K. Rust, personal communication, June 14, 2002), is to create a design-effect adjusted sampling weight by dividing the normalized sampling weight for each element by the average design effect for the variables in the analysis. Note that the normalized sampling weights should sum to the sample size. When the design effect is greater than 1.0, dividing the normalized weight by the design effect will result in an average sampling weight less than 1.0 and the sum of these adjusted weights will be less than the sample size. This new sum of adjusted weights is termed the "effective sample size." Use of these adjusted weights will result in an inflation of standard errors and a subsequent decrease in power due to the smaller effective sample size. The manual for the 1988 NELS explains this method:

> One analytic strategy for accommodating complex survey designs is to use the mean design effect to adjust for the effective sample size resulting from the design. For example, one could create a new rescaled, design-effect adjusted weight, which is the product of the inverse of the design effect and the rescaled case weight … and use this new weight to deflate the obtained sample size to take into account the inefficiencies due to a sample design that departs from a simple random sample. Using this procedure, statistics calculated by a statistical program such as SPSS will reflect the reduction in sample size in the calculation of standard errors and degrees of freedom. Such techniques only approximately capture the effect of the sample design on sample statistics. However, although not providing a complete accounting of the sample design, this procedure is a decidedly better approach than conducting an analysis that assumes the data were collected from a simple random sample. The analyst applying this correction procedure should carefully examine the statistical software he or she is using and assess whether the program treats weights in a way that will produce the effect described above. (U.S. Department of Education, 1996, p. 5-26)

This approach is fairly simple and does not require the SEM analyst to use specialized software, to write syntax to undertake more advanced sampling variance estimation methods, or to manually apply an inflation factor to each standard error estimate. Walker and Young (2003) discuss the use of this approach with contingency tables and Thomas and Heck (2001) demonstrate the superiority of using these design-effect adjusted weights with a regression analysis over a method that ignores the sampling design. Additionally, Stapleton (2002) examined the success of this method of using an effective sample size weight in her multilevel SEM simulation study. Using empirical data, Marsh and Yeung (1996) provide an applied example of the use of this design-effect adjusted weighting method with a complex sample dataset in their confirmatory factor analysis. This approach has the simplicity of adjusting all standard errors with just one change in the weights (instead of manually adjusting each standard error estimate), but an analyst must be sure that the software uses the sum of the weights as expected. For example, SAS, starting with version 8, does not use the sum of the weights in the analysis and instead uses the actual sample size and thus cannot be used with this method. However, this method can be used with current versions of SPSS as it uses the sum of the weights in calculation of standard errors. With current SEM software, raw data with these design-effect adjusted weights cannot be used because, similar to SAS, most SEM software uses the sample size in estimation of the asymptotic covariance matrix and not the sum of the weights. To use this procedure with SEM, the analyst would have to calculate a weighted covariance matrix outside of the software and provide this weighted matrix and the effective sample size as the number of observations to the software. Not addressed in other analytic contexts, but relevant in SEM analyses, is that the use of a smaller effective sample size will also result in lower model chi-square test statistics. It is not clear whether this effective sample size adjustment to the chi-square statistic will result in appropriate estimates, however. Given its ease of use, it would be beneficial to determine whether the standard error and chi-square estimates from this method are robust, considering the complexities typically found in complex sample data.

Another approach to variance estimation for statistics from complex samples is to estimate sampling variances using linearization. Variance estimates for nonlinear functions are often obtained by creating an approximate linear function, and then the variance of the new function is used as the variance estimate for the nonlinear function. This approach to variance estimation has several names in the literature, including the linearization method, the delta method, Taylor Series approximation (E.S. Lee et al., 1989), and the propagation of variance (Kish, 1965). In the specific case of complex sample data, linearization results in a variance estimate that is a weighted combination of the variation as assessed by the first-order derivatives across PSUs within the same stratum (Kalton, 1983a). Skinner et al. (1989) explain how this basic linearization technique is extended to covariance matrices and Muthén and Satorra (1995) further discuss the extension of this linearization method to covariance structures with complex data. Specifically, Muthén and Satorra propose that robust normal theory estimation can be generalized to data from complex samples. Assuming a three-stage sampling design with stratification at the first stage and p variables, the p^*-dimensional data vector of distinct elements of the sample covariance matrix for any given PSU can be calculated as

$$d_{ij} = \sum_{k=1}^{K} \sum_{l=1}^{L} w_{ijkl} \begin{bmatrix} (y_{ijkl1} - \overline{y}_1)(y_{ijkl1} - \overline{y}_1) \\ (y_{jkli2} - \overline{y}_2)(y_{ijkl1} - \overline{y}_1) \\ (y_{ijkl2} - \overline{y}_2)(y_{ijkl2} - \overline{y}_2) \\ \vdots \\ (y_{ijklp} - \overline{y}_p)(y_{ijklp} - \overline{y}_p) \end{bmatrix}$$

where w_{ijkl} is the inverse of the selection probability for the lth element (e.g., student) in the kth segment (e.g., school) within the jth PSU (e.g., county group) in the ith stratum (e.g., Census region), also where the $\overline{y}$ variables are a $p \times 1$ vector of simple weighted means defined as

$$\overline{y} = n^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} w_{ijkl} y_{ijkl}$$

and n is the sum of the weights. The vector of distinct elements of the sample covariance matrix across all strata and PSUs is then defined as

$$s_T = n^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J} d_{ij}$$

and is equivalent to calculating a simple weighted covariance matrix across all elements in the sample (disregarding the sample structure). The vector of parameter estimates, $\hat{\theta}$, is determined through minimization of the likelihood function:

$$\ln L = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl} \ln f(y_{ijkl} \,|\, \hat{\theta})$$

where

$$f(y_{ijkl} \,|\, \hat{\theta}) = (2\pi)^{-p/2} \,|\, \boldsymbol{\Sigma} \,|^{-1/2} \;\; \exp\{-\frac{1}{2} tr\{\boldsymbol{\Sigma}(\hat{\theta})^{-1}(y_{ijkl} - \boldsymbol{\mu}(\hat{\theta}))(y_{ijkl} - \boldsymbol{\mu}(\hat{\theta}))^{-1}\}$$

The asymptotic covariance matrices of s_T and the vector of normal theory estimates, $\hat{\theta}$, however, are not determined in the conventional manner.

Given the q-dimensional vector of normal theory estimates, $\hat{\theta}$, the standard errors for estimates from complex sample data are given via the asymptotic covariance matrix, $a\,\mathrm{cov}(\hat{\theta}) = \mathbf{I}^{-1}\mathbf{\Gamma}\mathbf{I}^{-1}$ where $\mathbf{I}$ denotes the information matrix, $-E\left[\dfrac{\partial^2 \ln L}{\partial\theta\partial\theta'}\right]$, and $\mathbf{\Gamma}$ is estimated by

$$\hat{\mathbf{\Gamma}} = \sum_{i=1}^{I} J(J-1)^{-1} \sum_{j=1}^{J} (d_{ij} - d_{i.})(d_{ij} - d_{i.})'$$

(Binder & Roberts, 2003; Muthén & Satorra, 1995; Skinner et al., 1989).

This equation for $\hat{\mathbf{\Gamma}}$ can be viewed as a measure of the pooled variability across the I strata of the J PSU estimates. Robust estimates of chi-square statistics are also based on this estimate of the asymptotic covariance matrix; specifically, a scaled chi-square statistic can be obtained given a function of the asymptotic covariance matrix of $\hat{\theta}$ and the information matrix.

$$\chi^2_{\mathrm{robust}} = \frac{q}{tr[\mathbf{I} \quad a\,\mathrm{cov}(\hat{\theta})]}\chi^2_{NT}$$

(Scientific Software International, n.d.).

If there are no sampling design effects, then the elements on the diagonal of the matrix resulting from the product of the information matrix and the asymptotic covariance matrix will each be equivalent to one and the scaling factor will therefore be one. As of this writing, this pseudomaximum likelihood (PML) procedure has been implemented in both M*plus* version 3.11 (Asparouhov, 2004) and LISREL, version 8.7 (Scientific Software International, n.d.) and has been termed PML estimation. Given the speed with which improvements are made to statistical software, it is expected that other SEM programs will be implementing this PML estimation process in the near future.

The implementation of this method assumes that the PSUs are selected with replacement and at a constant selection probability (Muthén & Muthén, 1998) and simulations have demonstrated the robustness of this approach under conditions meeting the assumptions (Asparouhov, 2004; Muthén & Satorra, 1995). It remains unclear how robust this method might be under conditions of typical NCES sampling designs, including selection without replacement, paired selection of PSUs, PPS selection, stratification, and three stages of sampling.

A final method of estimation of sampling variances that is typically used for means and regression analyses are replication techniques. These methods involve repeated sampling from the original sample and subsequent analysis in each replicate sample and the empirical distribution of the parameter estimates across these replicate samples is used to approximate the sampling variance of the parameter estimate. These methods have been employed with great success in regression analysis contexts and are available in several software packages as WESVAR, SUDAAN, and STATA but are not available in SEM software for complex sample data. Thus, these methods are not currently feasible to undertake for the typical secondary researcher and therefore are not considered in this article.

An analyst has several accessible options for undertaking an aggregate SEM analysis on complex sample data: a conventional analysis (ignoring the sampling design), a conventional analysis paired with adjustment of standard errors and chi-square based on the design effect, the use of a design-effect adjusted sampling weight, and the use of the linearization approach. Each of these procedures has advantages and disadvantages and the evaluation of each under typical sampling design conditions used in large-scale survey programs is of interest here.

METHOD

The study described in this article sought to assess the accuracy and feasibility of the following five methods: the conventional method, the conventional method with design-effect adjustments of standard errors, the design-effect adjusted weighting method, the linearization method with cluster identifiers, and the linearization method with stratum and cluster identifiers. The research included a Monte Carlo simulation study, using as a base the ECLS sample design. An examination of the robustness of parameter estimates and of standard error estimates and fit statistics that result from the use of the previously mentioned methods under selected sampling design conditions were of specific interest.

For this study, three finite target populations of data were created and six sampling designs were applied to each. The population data were generated using a four-stage structure (fictitious regions, geographic counties, schools, and students) and each population consisted of 60 regions with 12 counties per each region. Within each of the 720 counties, hypothetical school data were generated with two types of schools—“private” and “public.” In each county, data for 7 private schools and 23 public schools were generated. Finally, within each of the 21,600 schools, fictitious student data records were generated with varying numbers of students per school. Private schools were designed, on average, to have 47 students (with a standard deviation of 11) and public schools had an average size of 130 (with a standard deviation of 30). Within each school, student data were generated to reflect a two-stratum structure: 70% of the students were generated to be from a “majority” group and 30% from a “minority” group. Each population contained more than 2,300,000 observations.

Data were generated to reflect the population model in Figure 1; however, the three populations varied with regard to the amount of clustering within schools, counties, and regions as well as the strength of the path from the exogenous manifest variable to the endogenous factor. In Population 1, there were no grouping effects (across region, county, and school) and therefore the intraclass correlation was 0. Also, there was no difference in the strength of the path from the exogenous

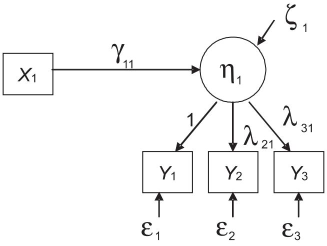


FIGURE 1 Population model of interest.

variable to the factor across school and student types. In Populations 2 and 3, the intraclass correlation was generated to be at a high level, .5, with each of the three levels of grouping (region, county, and school) accounting for a third of the grouping variance. In Population 2, there was no difference in the strength of the path from the exogenous variable to the factor across school and student types. In Population 3, "private school" students and "majority" students had higher levels of the direct effect. For Populations 1 and 2, γ_{11} was generated at .35 (a standardized value of .5), but in Population 3, γ_{11} was generated at .35 with a .07 increment each for students in private schools and majority students. For example, a minority student from a public school had a generating γ_{11} value of .35, a minority student from a private school had a generating γ_{11} value of .42 and a majority student from a private school had a generating γ_{11} value of .49.

Although population values were designed to be known values, because of random error and random sizes of schools in the generating process, the empirical population values differed slightly from the generating values. Table 1 displays the parameter values for the three populations in the study. The generating γ_{11} value for Population 3 was an estimate; because each combination of school and student stratum was associated with a different γ_{11}, the overall γ_{11} was estimated to be a weighted average of the γ_{11} across the strata, given expected population sizes. The estimated generating ψ_1 value was a function of this γ_{11} estimate.

TABLE 1
Population Parameter Values

		λ_{21}	λ_{31}	$Var(\varepsilon_1)$	$Var(\varepsilon_2)$	$Var(\varepsilon_3)$	γ_{11}	ψ_1
Population 1	Generating values	1.000	1.000	51.000	51.000	51.000	0.350	36.750
	Empirical values	0.999	0.999	50.959	51.080	50.982	0.350	36.784
Population 2	Generating values	1.000	1.000	51.000	51.000	51.000	0.350	36.750
	Empirical values	1.048	1.039	49.175	50.884	50.905	0.337	32.866
Population 3	Generating values	1.000	1.000	51.000	51.000	51.000	0.406	32.516
	Empirical values	1.054	1.034	49.219	50.656	51.095	0.391	32.416

The generating parameters of ICCs of 0 and .5 were chosen to investigate the range of bias that might occur with typical data; it would be rare to find ICCs above .5 in cross-sectional research. Table 2 contains empirical estimates from the ECLS data, showing the extent of clustering and the design effects of the sampling variance of the mean for seven typical variables. Of this set of variables, the greatest amount of cluster effects were found in the variables WKSESL, a derived measure of socioeconomic status, and CIRGSCAL, a general knowledge score. About 40% of the variability in these variables can be attributed to group level differences. Also included in Table 2 is an indication of the amount of variability accounted for by each level of the sample design (and these values drove the decision in the simulation to split the variability evenly between region, county, and school levels). Design effects ranged from about 4 to almost 9 for these variables. Note that age usually is not associated with large design effects in most complex sample datasets, but grouping effects on

TABLE 2
Empirical Estimates of ICCs and Design Effects
for Selected ECLS Variables

			Sources of Variability				
Variable	*Description*	*Overall ICC*	*Region*	*County group*	*School*	*Individual*	*Design effect*
C1RGSCAL	General knowledge IRT scale score	.38	.19	.14	.06	.62	8.778
C1RMSCAL	Math IRT scale score	.23	.10	.09	.05	.77	6.747
C1RRSCAL	Reading IRT scale score	.29	.12	.11	.05	.71	7.471
TOTALTV	User-computed variable estimating total TV hours during a week	.12	.03	.04	.05	.88	6.128
WKSESL	Continuous measure of SES (derived from income and occupation)	.40	.15	.14	.11	.60	8.366
P1CHLBOO	How many books child has (reported by parents)	.22	.08	.09	.05	.78	7.255
R1KAGE	Age at assessment (in months)	.10	.05	.02	.02	.90	4.036

Note. ECLS = Early Childhood Longitudinal Study; ICC = intraclass correlations; IRT = item response theory; SES = socioeconomic status; TV = television.

age are found in these data because the ECLS assessments were not administered on a given date to all sample elements but were instead spread out across a 3-month period as the interviewers needed to travel from site to site.

Once population data were generated, six different sampling designs were imposed on the population data and these designs are listed in Table 3 from the most simple to the most complex. Although not reflective of many large surveys, the simpler designs offer baseline estimates of the effects of analyzing each type of sampling design and demonstrate the principles of complex sample data analysis. The most complex design is intended to be similar to the design used in ECLS data collection. All sampling was conducted using *PROC SURVEYSELECT* statements in the SAS programming language. The first design, Scenario A, was a simple random sample of students, under which 14,400 observations were selected without replacement at random from more than 2.3 million student observations in the population. Scenario B was also a single stage selection of 14,400 students, but in this case, students were first split into two strata, majority and minority, and then students were selected with equal probability from each stratum (so 30% of the sample was minority students and 70% were majority students). Scenario C was very similar to Scenario B but unequal probability of selection was used such that 50% of the sample consisted of majority students and 50% were minority students. These first three sampling designs were quite simple and reflect the strategies that are more likely used with small local surveys than large, national and international

TABLE 3
Sample Designs for Simulation

Scenario	*"Counties"*	*"Schools"*	*"Students"*
A	—	—	Simple random sampling
B	—	—	Two strata of students, equal probability of selection
C	—	—	Two strata of students, unequal probability of selection
D	—	Two strata of schools, PPS selection of schools	Simple random sampling within schools
E	—	Two strata of schools, PPS selection of schools	Two strata of students, unequal probability of selection
F	Sixty regional strata of counties, two counties selected per stratum using (Brewer) PPS selection	Two strata of schools, unequal probability of selection, PPS selection	Two strata of students, unequal probability of selection

Note. PPS = probability proportionate to size.

studies. Scenario D involved two stages of sampling. First, schools were separated into strata by public and private status and then 720 schools were selected using probability proportionate to size sampling. Once the schools were selected, 20 students were randomly selected without replacement from each of the schools. Scenario E differed from D in the selection of students. In Scenario E, the students in the 720 selected schools were divided into majority and minority strata and then unequal selection rates were imposed such that 10 students were randomly chosen from each stratum. Finally, Scenario F replicated some of the sampling design of ECLS. First, all counties were separated into 60 regional strata. Within these region strata, two counties (or PSUs) were chosen using PPS sampling. Within the 120 selected counties, schools were divided into public and private strata and were selected with unequal probability from the two strata, using PPS selection. Finally, within the 720 schools, 10 students were chosen from each of the majority and minority strata using unequal selection probabilities.

Once the specified sampling design was imposed on the population data to result in selection of a sample of 14,400, five methods of analysis were undertaken. Method 1 entailed robust maximum likelihood estimation ignoring any stratification or clustering in the sampling design (previously referred to as the "conventional" method) but incorporating the sampling weights in the estimation of parameters. Method 1b consisted of manual adjustment of the standard error estimates from Method 1 by multiplying each estimate by the square root of the average design effect of the variables in the analysis. In addition, the chi-square test statistic was adjusted by dividing it by the average design effect. The average design effect was determined by obtaining an estimate of the sampling variance of the mean of each of the three dependent variables assuming SRS and also obtaining an estimate of the sampling variance of the means of each of the three variables using SAS *PROC SURVEYMEANS*, which approximates the sampling variance with Taylor series expansion using strata, PSU, and weight information (SAS Institute Inc., 1999). The design effect for each variable was then estimated by the ratio of the Taylor series expansion estimate of the sampling variance over the SRS-assumed sampling variance. In applied research, an analyst can obtain these estimates of design effects from the technical report that accompanies the large-scale database or can estimate them as done here using available software. The three dependent variable design effects were then averaged to obtain one composite estimate of the sampling design effect.

Method 2 included providing a weighted sample covariance matrix to the SEM software. The matrix was calculated using design-effect adjusted weights and the effective sample size (the sum of the adjusted weights) was provided to the software in place of the total number of sample observations. These design-effect adjusted weights were calculated as

$$w^* = w \times \frac{n}{\sum w} \times \frac{1}{\text{deff}}$$

where w^* is the design-effect adjusted weight, w represents the raw sampling weight (as appropriate given the sampling design condition), n represents the number of observations in the sample, and deff is the average estimated design effect for the dependent variables in the analysis (as described previously). Note that providing a covariance matrix calculated outside of the SEM software is required as most software will use the actual sample size in the estimation of the asymptotic covariance matrix instead of the sum of the weights (and thus effective sample size analyses with raw data cannot be undertaken with current SEM software).

Method 3 used the linearization approach, as described earlier, with PML estimation, but only with cluster identification where appropriate. Because secondary researchers may not always have access to variables that define the survey design, it was of interest to determine the robustness of estimates ignoring stratification information. Note that this method is relevant only for Sampling Scenarios D, E, and F (see Table 3).

Method 4 used PML estimation as well, but with all identification of strata, cluster, or both where appropriate. This method assumes that the researcher has access to any stratum and cluster variables that were used at the first stage of selection in the survey design. It was expected that this method of estimation would provide the most robust estimates across all sampling designs and populations.

All analyses were undertaken with M*plus* version 3.11 (Muthén & Muthén, 2004). Additionally, PML estimation (Methods 3 and 4) were undertaken with LISREL 8.7 to assess the comparability of the estimates from M*plus* and LISREL. The sampling process and analysis with the five methods were repeated 500 times for each of the sampling design and population combinations (for a total of 18 conditions). Because sampling weights to address unequal probability of selection were used for Methods 1a, 1b, 3, and 4 and a weighted matrix was provided for Method 2, key parameter estimates were expected to be unbiased with all approaches. Main interest was in the accuracy of standard error estimates and the effects on tests of fit. The accuracy of parameter and standard error estimates was determined by computing the percentage relative bias across the 500 replications. Fit was assessed by examining mean chi-square values and by comparing model chi-square rejection rates to the nominal ($\alpha = .05$) level.

RESULTS

All analyses resulted in converged and admissible solutions. In the sections that follow, first parameter estimate bias is discussed, then standard error bias, followed by test statistic information. Before examining the robustness of the estimates from the five methods, a review of the design effects associated with each of the six sampling designs is provided.

Sample Design Effects

Table 4 contains the average estimated design effects for variable means across the 500 replications for the dependent variables for each sample design and population combination. Design effects are not applicable in Scenario A, the SRS design. Scenario B, which included stratification of students by majority and minority status and the subsequent use of SRS with equal probability within each stratum, resulted in design effects of 1.0 for Populations 1 and 2. This null sampling design effect was anticipated because efficiencies are gained with stratification only when the variable on which data are stratified is related to the response variable and, for Populations 1 and 2, responses did not differ for majority and minority students. In Population 3, however, there were subpopulation differences across the two strata. Thus, for Sampling Scenario B under Population 3, the design effect was slightly less than 1.0, indicating that the sampling variances from this design are smaller than those expected from a SRS of the same size and this efficiency was gained by the use of stratification across majority and minority status.

Scenario C included stratification of students by majority and minority status with selection within these two groups at disproportionate rates. This nonoptimal selection (and subsequent use of sampling weights in analyses) led to less efficient estimates of the parameters than would have been found with SRS in all of the populations as evidenced by the design effects greater than 1.0. Note that the design effect for Population 3 was slightly lower than that of Populations 1 and 2 due to the informative stratification.

A more complex sampling situation, Scenario D, included the use of two-stage sampling. Schools were first stratified by public/private status and PPS sampling was used within each of these strata, then students were selected from the sample schools using SRS. Interestingly, in Population 1, this sampling design resulted in slightly more efficient estimates than SRS on average over the 500 replications.

TABLE 4
Empirical Average Design Effects of the Mean Across 500 Replications

Sampling Design	*Population 1* *ICC = 0.0* *No Strata* *Differences*	*Population 2* *ICC = 0.5* *No Strata* *Differences*	*Population 3* *ICC = 0.0* *Strata* *Differences*
A: SRS	—	—	—
B: Strat SRS	1.000	1.000	0.976
C: Strat unequal	1.162	1.163	1.135
D: 2-stage SRS	0.972	10.132	9.791
E: 2-stage complex	1.134	10.186	9.829
F: 3-stage complex	1.201	24.222	23.880

Note. ICC = intraclass correlations; SRS = simple random sampling.

When clusters are formed at random, as was done for Population 1, it is possible that the clusters may be internally more heterogeneous than the sample overall, resulting in a rare cluster design effect less than one (Kalton, 1983b). When there was homogeneity within schools (Populations 2 and 3), design effects much greater than 1.0 were obtained as expected. Again, Population 3 had slightly lower design effects given the differences across school strata present in the population.

Scenario E was similar to Scenario D; however, within schools students were selected at disproportionate rates across the two strata. Reflecting this nonoptimal sampling within schools, the design effects were somewhat inflated as compared to those of Scenario D across all populations.

Finally, Scenario F was quite complex, with sampling of counties, schools within those counties, and students within the schools, with disproportionate selection and stratification at each level of sampling. In Population 1, because the clustering was heterogeneous, the design effects are quite minimal, 1.210. However, Populations 2 and 3 demonstrate substantial design effects under this sampling design, above 20; the estimates of sampling variance for a mean assuming SRS would be over 20 times too small. Comparing the estimates found in Table 4 with the empirical estimates from ECLS in Table 2, the design effects of sampling scenarios D and E for Populations 2 and 3 are most similar to the ECLS empirical estimates.

Parameter Estimate Bias

Tables 5 and 6 provide average bias information for parameter estimates of the loadings and the direct effect, and the residual and disturbance variance, respectively. Parameter point estimates are equivalent for Methods 1, 3, and 4 and thus results are shown in just one column. As expected, the estimates of λ and γ were unbiased under all conditions for all methods (see Table 5). Estimates of residual and disturbance variances, however, were biased using Method 2 under conditions when design effects were greater than 1.0 (see Table 6). In all populations, in the more complex designs (sampling designs C, D, E, and F) the estimates were negatively biased. Given that the calculation of a weighted variance is not invariant to the scale of the weight, applying a deflated adjustment of the weight directly affects the magnitude of the variance and covariance estimates in the sample matrix. The estimates of the residual and disturbance variances were unbiased using Methods 1, 3 and 4 (see Table 6).

Standard Error Estimate Bias

As shown in Tables 7 and 8, negative standard error bias was found for Method 1 under conditions when data were homogeneous within clusters (Populations 2 and 3) and multistage sampling was utilized (sampling scenarios D, E, and F). Similar

TABLE 5
Average Parameter Estimate Bias in Direct Effects (λ_{21}, λ_{31}, and γ_{11})

Population	*Sampling Design*	*Methods 1, 3, and 4*	*Method 2 Design Effect Weighted Analysis*
ICC = 0.0	A: SRS	–0.000	—
	B: Strat SRS	–0.001	–0.001
	C: Strat unequal	0.000	0.000
	D: 2-stage SRS	–0.000	–0.001
	E: 2-stage complex	0.000	–0.000
	F: 3-stage complex	0.000	–0.000
ICC = 0.5 and no strata differences	A: SRS	0.000	—
	B: Strat SRS	0.000	0.000
	C: Strat unequal	0.000	0.000
	D: 2-stage SRS	0.001	0.001
	E: 2-stage complex	–0.000	–0.000
	F: 3-stage complex	–0.000	–0.000
ICC = 0.5 and strata differences	A: SRS	–0.000	—
	B: Strat SRS	0.000	0.000
	C: Strat unequal	–0.000	–0.000
	D: 2-stage SRS	0.001	0.000
	E: 2-stage complex	0.001	0.001
	F: 3-stage complex	0.001	0.001

Note. ICC = intraclass correlations; SRS = simple random sampling.

levels of standard error bias were exhibited for both direct effects and variance estimates and these biases were as great as –68%. Method 1b, the manual application of an inflation factor to each standard error estimate from Method 1, tended to overinflate the standard errors, as anticipated. Depending on the sampling scenario, the standard errors were between 33% and 58% too large when there was cluster sampling and homogeneity in the clusters. Method 2, the use of design effect adjusted weights, resulted in overestimation of the standard errors for direct effects but somewhat less overestimation as compared to Method 1b in the more complex sampling designs. Standard error estimates for the residual and disturbance variances were also biased with Method 2, but the bias was not consistent as it was confounded with the parameter estimate bias discussed previously. Method 3, the linear approximation estimation with cluster information only, provided fairly robust standard errors for sampling Scenarios D and E (two-stage sampling with only two strata of schools at the first stage) but provided somewhat positively biased estimates of standard errors under Scenario F, when paired selection of PSUs (counties) was undertaken within each of 60 regional strata at the first stage of sampling. Because strata were not taken into account, the standard error estimates are somewhat overestimated (by as much as 17% under the conditions ex-

TABLE 6
Average Parameter Estimate Bias of Residual Variances
[Var(ε_1), Var(ε_2), Var(ε_3), and ψ_1]

Population	*Sampling Design*	*Methods 1, 3, 4*	*Method 2* *Design Effect* *Weighted Analysis*
ICC = 0	A: SRS	0.000	—
	B: Strat SRS	0.000	0.000
	C: Strat unequal	–0.000	–0.140
	D: 2-stage SRS	–0.000	0.031
	E: 2-stage complex	–0.000	–0.116
	F: 3-stage complex	–0.000	–0.154
ICC = 0.5 and no strata differences	A: SRS	–0.000	—
	B: Strat SRS	0.000	–0.000
	C: Strat unequal	0.000	–0.140
	D: 2-stage SRS	–0.002	–0.902
	E: 2-stage complex	–0.001	–0.902
	F: 3-stage complex	–0.000	–0.958
ICC = 0.5 and strata differences	A: SRS	–0.001	—
	B: Strat SRS	–0.000	0.024
	C: Strat unequal	–0.001	–0.120
	D: 2-stage SRS	–0.001	–0.898
	E: 2-stage complex	–0.002	–0.898
	F: 3-stage complex	–0.002	–0.958

Note. ICC = intraclass correlations; SRS = simple random sampling.

amined here). Differences between the bias across Populations 2 and 3 were negligible. Finally, Method 4, the linear approximation estimation with cluster and strata information, provided robust standard error estimates under all sampling designs. At most, the standard errors were overestimated by 7% under the most complex sampling design examined in this study.

Model Fit

Model fit information is provided in Table 9 in the form of chi-square statistic rejection rates at a nominal alpha of .05. Under the Population 1 condition, rejection rates were close to the expected level of 5% for all estimation methods. However, under conditions of cluster homogeneity (Populations 2 and 3) the estimate of the chi-square statistic using Method 1 led to rejection of the model too often when the data were collected with a multistage sample (sampling scenarios D, E, and F). Rejection rates of 50% to 60% were found with the two-stage sampling design and rates of 75% were found with the three-stage design.

TABLE 7
Average Standard Error Estimate Bias for Direct Effects (λ_{21}, λ_{31}, and γ_{11})

Population	*Sampling Design*	*Method 1 Conventional Analysis*	*Method 1b Conventional Analysis w/ Adjusted Standard Errors*	*Method 2 Design Effect Weighted Analysis*	*Method 3 Linearized (PML) Cluster*	*Method 4 Linearized (PML) Cluster and Stratum*
ICC = 0.0	A: SRS	0.016	—	—	—	
	B: Strat SRS	–0.001	–0.001	–0.000	—	–0.001
	C: Strat unequal	–0.016	0.061	–0.015	—	–0.016
	D: 2-stage SRS	0.007	–0.008	–0.007	0.005	0.005
	E: 2-stage complex	0.006	0.071	–0.007	0.005	0.005
	F: 3-stage complex	0.009	0.104	0.008	0.005	0.002
ICC = 0.5 and no strata differences	A: SRS	0.023	—	—	—	—
	B: Strat SRS	–0.055	–0.055	–0.055	—	–0.055
	C: Strat unequal	0.009	0.088	0.007	—	0.009
	D: 2-stage SRS	–0.576	0.348	0.348	0.015	0.015
	E: 2-stage complex	–0.539	0.472	0.364	0.021	0.021
	F: 3-stage complex	–0.682	0.560	0.422	0.165	0.070
ICC = 0.5 and strata differences	A: SRS	0.000	—	—	—	—
	B: Strat SRS	0.053	0.040	0.041	—	0.052
	C: Strat unequal	–0.032	0.031	–0.043	—	–0.033
	D: 2-stage SRS	–0.574	0.331	0.332	0.011	0.010
	E: 2-stage complex	–0.535	0.459	0.357	0.024	0.024
	F: 3-stage complex	–0.682	0.553	0.425	0.147	0.059

Note. ICC = intraclass correlations; PML = pseudomaximum likelihood; SRS = simple random sampling.

TABLE 8
Average Standard Error Estimate Bias for Residual Variances [Var(ε_1),Var(ε_2),Var(ε_3), & ψ_1]

Population	*Sampling Design*	*Method 1 Conventional Analysis*	*Method 1b Conventional Analysis w/ Adjusted Standard Errors*	*Method 2 Design Effect Weighted Analysis*	*Method 3 Linearized (PML) Cluster*	*Method 4 Linearized (PML) Cluster and Stratum*
ICC = 0.0	A: SRS	–0.005	—	—	—	
	B: Strat SRS	–0.004	–0.004	–0.004	—	–0.004
	C: Strat unequal	–0.013	0.064	–0.009	—	–0.013
	D: 2-stage SRS	–0.032	–0.046	–0.555	–0.033	–0.033
	E: 2-stage complex	0.021	0.087	–0.481	0.020	0.020
	F: 3-stage complex	–0.024	0.068	–0.842	–0.026	–0.030
ICC = 0.5 and no strata differences	A: SRS	0.001	—	—	—	—
	B: Strat SRS	–0.009	–0.009	–0.002	—	–0.009
	C: Strat unequal	0.024	0.104	0.030	—	0.024
	D: 2-stage SRS	–0.574	0.356	0.370	0.019	0.019
	E: 2-stage complex	–0.540	0.468	0.375	0.016	0.016
	F: 3-stage complex	–0.683	0.557	–0.266	0.153	0.055
ICC = 0.5 and no strata differences	A: SRS	0.015	—	—	—	—
	B: Strat SRS	0.041	0.028	0.022	—	0.040
	C: Strat unequal	0.027	0.094	0.013	—	0.026
	D: 2-stage SRS	–0.570	0.344	0.331	0.023	0.023
	E: 2-stage complex	–0.531	0.469	0.356	0.035	0.035
	F: 3-stage complex	–0.677	0.578	–0.259	0.169	0.073

Note. ICC = intraclass correlations; PML = pseudomaximum likelihood; SRS = simple random sampling.

TABLE 9
Chi-Square Rejection Rates (Expressed as a Percentage)

Population	*Sampling Design*	*Method 1 Conventional Analysis*	*Method 1b Conventional Analysis w/ Adjusted* χ^2	*Method 2 Design Effect Weighted Analysis*	*Method 3 Linearized (PML) Cluster*	*Method 4 Linearized (PML) Cluster and Stratum*
ICC = 0.0	A: SRS	4.4	—	—	—	—
	B: Strat SRS	4.0	3.8	3.8	—	4.0
	C: Strat unequal	4.0	2.8	4.6	—	4.4
	D: 2-stage SRS	6.6	7.0	7.2	6.8	6.8
	E: 2-stage complex	4.8	3.2	5.4	5.2	5.2
	F: 3-stage complex	6.8	4.0	8.0	7.2	6.6
ICC = 0.5 and no strata differences	A: SRS	6.4	—	—	—	—
	B: Strat SRS	3.4	3.4	3.4	—	3.4
	C: Strat unequal	5.6	3.8	5.4	—	5.6
	D: 2-stage SRS	61.6	0.4	0.4	4.8	4.8
	E: 2-stage complex	48.8	0.0	0.2	4.4	4.4
	F: 3-stage complex	75.0	0.2	0.4	2.2	3.4
ICC = 0.5 and no strata differences	A: SRS	3.6	—	—	—	—
	B: Strat SRS	7.0	7.2	7.4	—	7.0
	C: Strat unequal	6.0	3.4	6.2	—	6.0
	D: 2-stage SRS	57.4	0.2	0.2	4.2	4.2
	E: 2-stage complex	54.8	0.4	1.0	5.8	5.8
	F: 3-stage complex	74.6	0.0	0.2	2.2	3.4

Note. ICC = intraclass correlations; SRS = simple random sampling.

Both Methods 1b and 2 resulted in inappropriately low chi-square values when clustering was a characteristic of the sampling design and the population contained homogeneity within clusters. The design-effect adjustment was too conservative in its estimates of standard errors and thereby led to an overcorrection of the chi-square. Finally, Methods 3 and 4 appear to provide fairly accurate estimates of chi-square under all sampling scenarios.

It is important to note what the chi-square statistic is being used to test. In this case of an aggregate analysis, the adjusted chi-square is testing whether the model is plausible within the finite population that the sample was drawn to represent. That is, if SRS had been undertaken instead of a complex sampling design, is the model an appropriate representation of the relations within the sample covariance matrix? However, a high value of the chi-square test statistic from a conventional analysis can also alert the researcher to the issue that there are differences across clusters or strata within the finite population and these differences may deserve some attention. The analyst must address the theoretical concern of whether the grouping is truly a nuisance or whether the grouping contributes to the causal mechanism behind the relations among the constructs. In the absence of an experimental or longitudinal study, there are no statistical tests to determine whether the grouping contributes to the causal relations and the analyst must argue his or her position on theoretical grounds. At a minimum, an analyst should consider running both a weighted conventional analysis and an analysis using the linear approximation and examine the discrepancies in the chi-square values and subsequent interpretation of model fit. If the model is plausible under the linearized estimation and is not plausible under the conventional analysis, the researcher should be sure to make explicit his or her rationale for treating the grouping as a nuisance instead of using analysis methods that model the grouping relations, such as multilevel SEM.

AN EXAMPLE ANALYSIS

To demonstrate the practical differences that might be found in running a structural equation model with complex sample data under the methods discussed here, the model shown in Figure 1 was imposed on data from the ECLS dataset. Specifically, the theoretical model hypothesized that hours watching television per week had an effect on a child's achievement level. Achievement was considered a latent factor, indicated by three assessment scores on reading, math, and general knowledge. These variables were previously discussed and the design effects for the means of these variables are shown in Table 2. For this analysis, data were included for 12,744 kindergarten children, from 89 regional strata, clustered in 518 counties. The cluster sizes ranged from 1 to 204 children. To undertake the analysis, a new PSU variable had to be created. In the ECLS dataset, the variable *Y2COMPSU* represents the sequentially numbered PSU within the *i*th stratum. For input to M*plus*, each PSU must have a different identifier (and therefore the second PSU in Stratum 1 and the second PSU in Stratum 2 could not both have a *Y2COMPSU* value of "2"). Methods 3 and 4 were run using the M*plus* syntax shown in Appendix A; however, Method 3 did not identify the *STRAT* variable. Method 1 used *ESTIMATOR=MLR* (for Robust Maximum Likelihood) and only identified the *WEIGHT* variable. For Methods 1b and 2, an estimate of the design effect had to be obtained, and the design effects of the mean of the three dependent variables in the analysis, *C1RRSCAL*, *C1RGSCAL*, and *C1RMSCAL*, were used to obtain an average design effect of 7.665. For Method 1b, the square root of this design effect was used as an inflation factor for the standard errors from Method 1. For Method 2, the sampling weights in the dataset were adjusted by first dividing the raw sampling weight by the mean of the weights (to create a normalized sampling weight) and then this normalized weight was divided by the average design effect of 7.665. A weighted covariance matrix was calculated and this weighted matrix and the effective sample size of 1,667 (the sum of the design-effect adjusted weights) were provided to M*plus* for estimation of the model parameters.

Estimates from these analyses are provided in Table 10. As expected, the chi-square value from the conventional analysis is higher than the values for the other methods that take the sampling design into account. With $df = 2$, the test statistic would lead to rejection of the hypothesized model, as least on an exact fit basis, for the conventional analysis, as well as under the PML estimation (Methods 3 and 4). Both of the adjusted chi-square values from the methods involving design-effect adjustments, Method 1b and 2, would have resulted in a failure to reject the hypothesized model, but from the previous simulation results, it is clear that these chi-square values are too low and are overcorrected.

Looking at the parameter estimates, it should be clear that Methods 1, 3, and 4 all provide the same estimates. Additionally, Method 2, the design-effect weighted estimate, provides the same λ and γ estimates as the other methods and the only difference is in the estimate of the variances, both of the latent factor and the residuals.

Finally, looking at the standard error estimates, as expected, the smallest standard errors are found with the conventional analysis (apart from the standard errors that are confounded with the bias in the variance estimates in Method 2). The standard errors associated with Methods 1b, 3, and 4 are all larger than those seen for Method 1, with the manual adjustment method, Method 1b, providing the most conservative estimates. This manual adjustment, as seen in the simulation results, results in overly conservative estimates of the standard error. The standard errors for Methods 3 and 4 are very similar, with Method 4 having somewhat lower standard errors than Method 3 for most parameter estimates. This difference is a result of the precision gained from acknowledging that the data come from a stratified sampling design. A comparison of the standard errors from Method 1 and Method 4 provides us with the average root design effect for the SEM estimates of 1.438; on average, the standard errors from the conventional analysis needed to be in-

TABLE 10
Estimates from the Example ECLS Analysis

Estimate	*Method 1* *Conventional* *Analysis*	*Method 1b* *Conventional Analysis* *w/ Adjusted Standard* *Error and χ^2*	*Method 2* *Design Effect* *Weighted* *Analysis*	*Method 3* *Linearized* *(PML) Cluster*	*Method 4* *Linearized* *(PML) Cluster* *and Stratum*
Model χ^2	18.010	2.350	4.681	13.050	12.540
Parameter estimates					
λ_{21}	.840	—	.840	.840	.840
λ_{31}	1.233	—	1.233	1.233	1.233
γ	–0.143	—	–0.143	–0.143	–0.143
ζ	52.165	—	6.800	52.165	52.165
δ_{11}	34.752	—	4.532	34.752	34.752
δ_{22}	18.004	—	2.348	18.004	18.004
δ_{33}	24.196	—	3.155	24.196	24.196
Std. error estimates					
λ_{21}	0.016	0.044	0.025	0.035	0.031
λ_{31}	0.021	0.058	0.036	0.024	0.023
γ	0.014	0.039	0.023	0.016	0.016
ζ	1.600	4.430	0.382	3.358	2.708
δ_{11}	0.858	2.376	0.208	0.999	1.006
δ_{22}	0.575	1.592	0.126	1.165	0.991
δ_{33}	1.377	3.813	0.234	1.576	1.387

Note. ECLS = Early Childhood Longitudinal Study; PML = pseudomaximum likelihood.

flated by 44% to provide a more accurate representation of the variability due to sampling, assuming that these linearized estimates from Method 4 are appropriate. Given the sample size here, there were no differences in substantive interpretation—all parameters were found to be significantly different from zero, under any of the methods.

DISCUSSION

Although the results shared here only include the more accessible approaches to modeling complex sample data (and not replication methods), the information provided should be helpful to secondary analysts. If one is working with data from a complex sample, the use of normalized weights should result in unbiased estimates of population parameters for all the scenarios examined in this study. Note, however, that if such a weighted conventional analysis is undertaken ignoring the complex design, the standard errors of those parameter estimates could be negatively biased by as much as 68%, given the conditions examined in this simulation study (and 44% given the empirical analysis). Manual adjustment of standard errors by inflation by the root design effect tended to overinflate the standard errors by as much as 58% in the conditions examined in the simulation study.

From the results found here, use of the design-effect adjusted weighting scheme, as used with univariate and regression statistics, is not recommended for use with SEM. Although the estimates of parameters that are usually of the most interest, loadings and direct paths, were unbiased with this method, the estimates of residual and disturbance variances were negatively biased. In addition, the standard errors of all estimates were positively biased and chi-square test statistics tended to lead to acceptance of the null hypothesis more often than desired given the nominal alpha level.

The linearization method, using PML estimation, with only cluster identifiers appears to provide both robust estimates of parameters and of standard errors under conditions when two-stage sampling is used with very few strata, given the simulation results. However, when paired selection of PSUs was undertaken from many informative strata (a more typical design for most large-scale surveys), the standard error estimates were somewhat positively biased (as high as 17%) with this method, and the chi-square statistics were somewhat low on average. As expected, the linearization method with stratum identifiers, cluster identifiers, or both as appropriate appears to provide both robust estimates of parameters and of standard errors under all conditions studied here.

Given these results, the user is encouraged to consider a linearization approach using PML and is further encouraged to obtain all survey design information that would aid in analysis, including first-stage stratum and cluster identifiers. Currently, the software programs M*plus* and LISREL can accommodate such analyses and the syntax and instructions for running the analyses for the model used in this simulation and the empirical example are provided in the Appendixes. Future research may consider comparing the performance of the PML estimation with the replicated sampling variance estimation methods typically used for regression-type analyses: jackknife replicates, balanced repeated replicates, and bootstrapping. It also should be noted that Wolter (1985) argued that linear approximation "may not be satisfactory with highly skewed data" and may only be robust when samples are "sufficiently large" but provides no indication of the magnitude of "sufficiently large" (pp. 226–227). Further examination of the robustness of linearized estimates under varying distributional and sample size conditions may also be warranted.

ACKNOWLEDGMENTS

This research was supported by a grant from the American Educational Research Association which receives funds for its AERA Grants Program from the U.S. Department of Education's National Center for Education Statistics of the Institute for Education Sciences and the National Science Foundation under NSF Grant RED–9980573. Opinions reflect those of the author and do not necessarily reflect those of the granting agencies. I would like to particularly thank Dr. Tihomir Asparouhov and two anonymous reviewers for their helpful comments on an earlier draft of this manuscript.

REFERENCES

Asparouhov, T. (2004). *Stratification in multivariate modeling.* M*plus* Web Notes: No. 9. Los Angeles: Muthén & Muthén.

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

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

Fan, X. (2001). The effect of parent involvement on high school students' academic achievement: A growth modeling analysis. *Journal of Experimental Education, 70,* 27–61.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kalton, G. (1977). Practical methods for estimating survey sampling errors. *Bulletin of the International Statistical Institute, 47,* 495–514.

Kalton, G. (1983a). Models in the practice of survey sampling. *International Statistical Review, 51,* 175–188.

Kalton, G. (1983b). *Introduction to survey sampling.* Thousand Oaks, CA: Sage.

Kaplan, D., & Elliott, P. R. (1997). A didactic example of multilevel structural equation modeling applicable to the study of organizations. *Structural Equation Modeling, 4,* 1–24.

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

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

Kish, L., & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society Series B, 36,* 1–37.

Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). *Analyzing complex survey data.* Newbury Park, CA: Sage.

Lee, R. (1990). *User's manual: NELS:88 base year student component data file.* Washington, DC: National Center for Education Statistics.

Longford, N. T. (1995). *Model-based methods for analysis of data from 1990 NAEP trial state assessment.* Washington, DC: National Center for Education Statistics, 95–696.

Marsh, H. W., & Yeung, A. S. (1996). The distinctiveness of affects in specific school subjects: An application of confirmatory factor analysis with the National Educational Longitudinal Study of 1988. *American Educational Research Journal, 33,* 665–689.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide* (3rd ed.). Los Angeles: Muthén & Muthén.

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

Potthoff, R. F., Woodbury, M. A., & Manton, K. G. (1992) "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association, 87,* 383–396.

SAS Institute. (1999). *On-line SAS documentation.* Retrieved February 10, 2004, from http://www.id.unizh.ch/software/unix/statmath/sas/sasdoc/stat/chap61/sect1.htm#idxsvm0001

Scientific Software International. (n.d.) *Analysis of structural equation models for continuous random variables in the case of complex survey data.* Retrieved November 24, 2004 from www.ssicentral.com/Lisrel/techdocs/compsem.pdf

Selfa, L. A., Suter, N., Myers, S., Koch, S., Johnson, R. A., Zahs, D. A., et al. (1997). *1993 National Study of Postsecondary Faculty (NSOPF–93): Methodology report.* Washington, DC: National Center for Education Statistics.

Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of complex surveys.* Chichester, England: Wiley.

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

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

U.S. Department of Education. (1996). *National Education Longitudinal Study: 1988–1994 methodology report.* Washington, DC: National Center for Education Statistics.

U. S. Department of Education. (2001). *ECLS-K, base year public-use data file, kindergarten class of 1998–99: Data files and electronic code book: (Child, teacher, school files).* Washington, DC: National Center for Education Statistics.

Walker, D. A., & Young, D. Y. (2003). Example of the impact of weights and design effects on contingency tables and chi-square analyses. *Journal of Modern Applied Statistical Methods, 2,* 425–432.

Wine, J. S., Heuer, R. E., Wheeless, S. C., Francis, T. L., & Dudley, K. M. (2002). *Beginning Postsecondary Students Longitudinal Study: 1996–2001 (BPS:1996/2001) methodology report.* Washington, DC: National Center for Education Statistics.

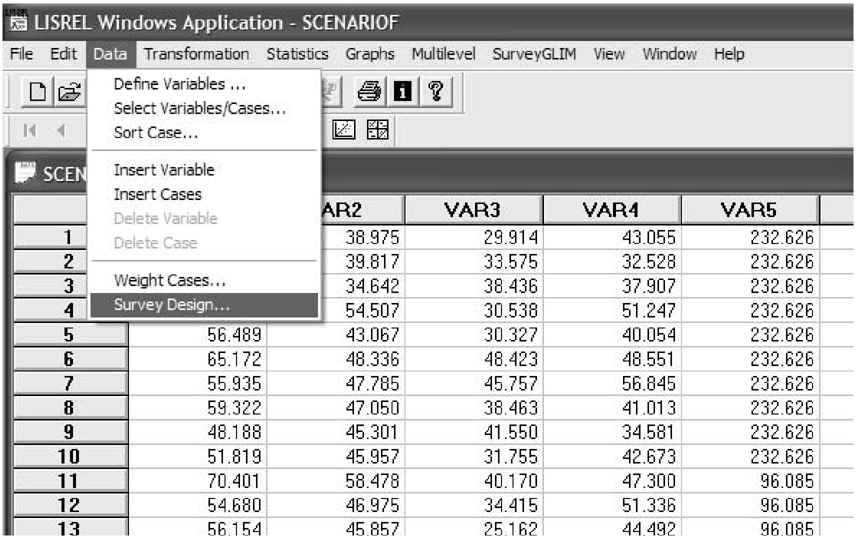
Wolter, K. M. (1985). *Introduction to variance estimation.* New York: Springer-Verlag.

APPENDIX A
Syntax to run M*plus* Complex Sample Aggregate Analysis

```
TITLE: METHOD7
DATA: FILE IS “C:\temp\SCENARIOF.DAT";
VARIABLE:
  NAMES ARE V1 V2 V3 V4 WEIGHT SCHOOL COUNTY REGION;
  USEVARIABLES ARE V1 V2 V3 V4 WEIGHT COUNTY REGION;
  WEIGHT IS WEIGHT;
  CLUSTER IS COUNTY;
  STRAT IS REGION;
ANALYSIS: TYPE IS COMPLEX;
MODEL:
  F1 BY V2 V3 V4;
  F1 ON V1;
OUTPUT:
  STANDARDIZED SAMPSTAT;
```

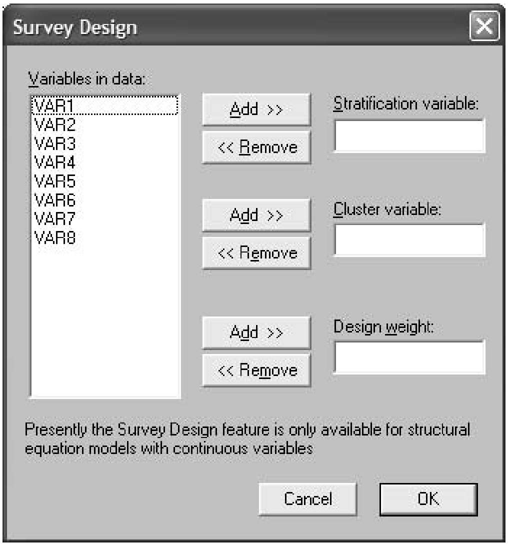
APPENDIX B
Screen Shots to Identify Complex Sampling Design in .psf File Creation in LISREL 8.7

After opening a .psf file, click on *Data* on the menu bar, then click on *Survey Design.*



		AR2	VAR3	VAR4	VAR5
1		38.975	29.914	43.055	232.626
2		39.817	33.575	32.528	232.626
3		34.642	38.436	37.907	232.626
4		54.507	30.538	51.247	232.626
5	56.489	43.067	30.327	40.054	232.626
6	65.172	48.336	48.423	48.551	232.626
7	55.935	47.785	45.757	56.845	232.626
8	59.322	47.050	38.463	41.013	232.626
9	48.188	45.301	41.550	34.581	232.626
10	51.819	45.957	31.755	42.673	232.626
11	70.401	58.478	40.170	47.300	96.085
12	54.680	46.975	34.415	51.336	96.085
13	56.154	45.857	25.162	44.492	96.085

Then identify any stratification, cluster, and sampling weight variables that are present on the dataset. Note that the cluster variable refers to the PSU and not the ultimate cluster.



Presently the Survey Design feature is only available for structural equation models with continuous variables

This complex sample design will then always be used in modeling with this .psf file (i.e., linearization estimates of standard errors and adjusted chi-square values will always be undertaken). To undertake a conventional analysis, the .psf will have to be resaved without the *Survey Design* information.