The Journal of Experimental Education, 2006, 74(2), 163–190

Utilization of Sample Weights in Single-Level Structural Equation Modeling

DEBBIE L. HAHS-VAUGHN
University of Central Florida

RICHARD G. LOMAX
University of Alabama

ABSTRACT. Complex survey designs often employ multistage cluster sampling designs and oversample particular units to ensure more accurate population parameter estimates. These issues must be accommodated in the analysis to ensure accurate parameter estimation. Incorporation of sample weights in some statistical procedures has been studied. However, research on the behavior of sample weights on estimates, standard errors, and fit measures in latent variable models is negligible, and studies examining methodology on latent variable modeling applications using extant data are rare. Using the Beginning Postsecondary Students Longitudinal Study 1990/92/94, the authors found, with mixed results, that a statistically significant difference exists in estimates and fit indices when weights and designs are applied versus when they are ignored.

Key words: complex samples, design-based approach, design effect adjusted weights, latent variable modeling, sample weights, structural equation modeling

ORGANIZATIONS such as the American Educational Research Association and the Association for Institutional Research offer grants and fellowships, including summer database training opportunities, to introduce researchers to large complex datasets available through the National Center for Education Statistics (NCES) and the National Science Foundation (Association for Institutional Research, n.d.). With increased interest in and access to analyzing national datasets, researchers must understand the challenges associated with modeling complex sample designs (Stapleton, 2002). The common threads that run across most

Address correspondence to: Debbie L. Hahs-Vaughn, University of Central Florida, PO Box 161250, Orlando, FL 32816-1250. E-mail: dhahs@mail.ucf.edu

complex samples—and likewise that may create the most challenge in interpreting and using the data—is that usually the data have been collected by multistage cluster sampling and subsets of the population have been oversampled. These issues associated with complex samples must be accommodated in the analysis to ensure that parametric assumptions are not violated and accurate parameter estimates result (Hahs-Vaughn, 2005).

Multilevel models (i.e., model-based approach) to account for multistage cluster sampling have received much attention recently (Hox & Kreft, 1994; Kaplan & Elliott, 1997b; Muthen, 1994). Although it has been argued that the appropriate statistical method to analyze multistage samples is multilevel modeling, multilevel modeling is not always appropriate for complex surveys (Kaplan & Elliott) and not all researchers may be interested in a multilevel approach. When a multilevel model is not appropriate, a single-level model, or design-based approach, can be employed. However, the researcher is faced with how to apply weights accurately to the data to compensate for oversampled groups and how to effectively address homogeneities that exist due to the multistage cluster sampling (usually through the use of design effects). Research on the behavior of oversampling and cluster sampling on parameter estimates, standard errors, and fit measures in latent variable models that use a design-based approach is negligible at best (exceptions include Hahs, 2003; Kaplan & Ferguson, 1999), and studies that actually use sample weights and design effects in studying the methodology of latent variable modeling applications using extant complex sample data are rare (an exception is Hahs).

Incorporation of sampling weights in various statistical procedures (such as regression and multilevel structural equation modeling [SEM]) to address oversampling has also been the focus of recent research (Kaplan & Ferguson, 1999). Numerous studies have shown that ignoring oversampling by failing to apply survey weights to analyses can lead to biased population parameter estimates (DuMouchel & Duncan, 1983; Korn & Graubard, 1995b; Skinner, Holt, & Smith, 1989). Although Kaplan and Ferguson conducted a bootstrapping simulation to determine the impact of weights in a single group factor analysis model, the sampling model was rather simple and thus somewhat unrealistic. Regardless of this limitation, Kaplan and Ferguson appear to be the first researchers to systematically address weighting issues in a latent variable context. A review of more recent literature does not acknowledge additional studies using survey weights in a methodological study using single-level SEM (i.e., a design-based approach) nor do studies exist that specifically examine the use or omission of weights in an actual SEM application with extant data from a complex survey (although there are many studies that correctly apply weights using national data; e.g., Fan, 2001; Perna, 2003).

The significance of conducting a methodological study that uses extant data rather than simulated data is to illuminate the extent to which simulation study results related to complex samples in SEM play out in an actual model. For example, if oversampling and cluster sampling are ignored, how do parameter estimates, standard errors, and model fit differ? Although simulations have been invaluable in contributing knowledge used by researchers who conduct analyses using complex samples, applying the knowledge of simulations and testing what we know in an actual analysis will complement the body of literature on this topic. The purpose of this study was to analyze a single-level structural equation model using data from an existing complex sample and to compare the results (i.e., parameter estimates, standard errors, fit indices) when the data are analyzed with and without accommodations for oversampling and cluster sampling. We hypothesized that parameter estimates, standard errors, and fit indices of the structural equation model will differ significantly when weights and design effects are applied versus when they are not applied to accommodate the oversampling and cluster sampling, respectively. We used data from the Beginning Postsecondary Students Longitudinal Study 90/92/94. We analyzed the methodological issue of the use of weights in SEM using extant data with the model based in theory. In addition, we provide recommendations on analyzing complex samples with structural equation models.

Next, we present previous research relating to the accommodation of oversampling and cluster sampling in SEM as (a) complex survey designs, (b) disproportionate sampling, and (c) homogeneities within multistage cluster sampling in complex samples.

Complex Survey Design and Structural Equation Modeling

Traditional standard error formulas are based on simple random samples (Lee, Forthofer, & Lorimor, 1989), and an assumption in SEM is independence of observations (Schumacker & Lomax, 1996). National surveys, however, do not usually gather data based on simple random samples. Most times, national surveys have been collected using both multistage cluster sampling (such as sampling schools first and then students, which may create homogeneities within the clusters) and disproportionate sampling (such as oversampling some groups, which creates disproportion selection; Thomas & Heck, 2001). When pre-existing clusters or natural groups are selected as the basis for the sample design, homogeneities exist that negate the assumption of independence (Kish & Frankel, 1974). Units have unequal probabilities of being included in a sample when oversampling is employed (Kaplan & Ferguson, 1999). SEM is a popular tool to study complex interrelationships between and among variables (Kaplan & Elliott, 1997b); however, the homogeneities that exist within clusters and the disproportionate selection probabilities due to oversampled groups must be addressed to ensure correct model specification and accurate parameter estimates (Stapleton, 2002).

Disproportionate Sampling

Sampling weights play a vital role in modeling data by testing and protecting against sampling designs that could cause selection bias and by protecting against misspecification of the model (Pfeffermann, 1993). Ignoring disproportionate sampling results in obtaining parameter point estimates that are biased (Kalton, 1989; Korn & Graubard, 1991, 1995b; Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998; Stapleton, 2002) and poor performance of test statistics and confidence intervals (Pfeffermann). The parameter estimate bias that can result when unequal probability of selection is ignored has been illustrated using regression and SEM (Hahs, 2003; Hahs-Vaughn, 2005; Kaplan & Ferguson, 1999; Korn & Graubard; Lee et al., 1989). Using design weights is often the easiest way to address unequal probability of selection (Stapleton).

The weighted sum of sample observations is an unbiased estimator of the population parameters when

$$E(\sum_{j=1}^{n} y_j / P_j) = \sum_{i=1}^{N} Y_i.$$

In this case, i represents individuals in the population from 1 to N, j is the draw or selection and ranges from a value of 1 to n, and P_j is the selection probability of the jth sample element (Kish & Frankel, 1974). In the most basic case, a sample weight is the inverse of the probability that the observation will be included in the sample (Kaplan & Ferguson, 1999; Muthen & Satorra, 1995). If observation i has the probability p_i of being included in the sample, the sample weight w_i for observation i is $w_i = 1/p_i$ with $p_i = n/N$ for all observations i. Thus, the sample weights sum to the population size, N.

Groundbreaking work for using sample weights when working with samples that have known inclusion probabilities is from Horvitz and Thompson (1952), who found that unbiased estimators (under any sampling design) can be obtained by using sample weights when inclusion probabilities are known (cited in Kaplan & Ferguson, 1999). When there is only limited knowledge about the actual sampling process and limited access to design information, the sampling weight becomes an important model feature; however, there is no clear method to incorporate the weight, and different opinions exist as to whether weights are even relevant (Pfeffermann, 1993). In what has been classified as "pioneering" in the world of complex surveys, Skinner et al. (1989) concluded that weighted estimates should be used because of the robustness they provide for model misspecification and suggested comparing weighted and unweighted parameter estimates as a diagnostic tool to check the adequacy of a model. This recommendation has been suggested by others (e.g., Hahs, 2003) and has been illustrated in practice (e.g., DuMouchel & Duncan, 1983).

As mentioned previously, the sum of the raw weights equals the population size, *N*. Estimates that may be sensitive to sample size, such as test statistics, may be unduly influenced if raw weights are applied to the data (Thomas & Heck, 2001). Normalized weights, also referred to as *relative weights*, preserve the sample size such that they sum to the actual sample size, *n* (Kaplan & Ferguson, 1999; Peng, 2000; Thomas & Heck) and can be computed as the weight divided by the mean weight (Peng; Thomas & Heck). Applying normalized weights ensures that oversampling is accommodated. In what seems to be the only published study using sample weights in single-level SEM, Kaplan and Ferguson used a bootstrapping simulation design to examine single-level SEM parameter behavior when raw and normalized weights were applied versus when no weight was applied. They found that when weights were ignored in disproportional sampling (compared with using raw weights or normalized weights), serious bias was found in latent variable model parameter point estimates as compared to the population values. The most extensive bias was for error variances and the factor variance, while generally small and identical bias was seen in the raw weighting and normalized weighting conditions with diminishing bias as the strata sample sizes increased (Kaplan & Ferguson). Regarding sampling variability, they found that standard errors relative to the standard deviation of the estimates were underestimated for the raw weight and normalized weight condition with raw weights yielding smaller standard errors. Because the weights sum to the sample size in the normalized weight condition, the bias values for the raw weight condition were larger than the normalized weight condition. When no weights were used, the biases were positive and smaller (Kaplan & Ferguson).

In reviewing fit statistics, raw weighting, because the population size is used in the calculation, substantially inflates the likelihood-ratio chi-square statistic, but the chi-square statistic decreased as the strata sample size increased. Normalized weighting generates likelihood-ratio chi-square statistics that are overall closer to the results of the population. Other SEM fit indices, including root mean square error of approximation (*RMSEA*), expected cross-validation index (ECVI), Akaike Information Criterion (AIC), and consistent AIC (CAIC), showed that normalized weights provided mixed results dependent on sample sizes with no weighting and raw weighting providing similar results. Regardless of the weighting procedure used, larger strata sample sizes provided measures closer to the population values. The weighting employed did not affect the goodness-of-fit index (GFI), adjusted GFI (AGFI), nonnormed fit index (NNFI), and comparative fit index (CFI; Kaplan & Ferguson, 1999). Kaplan and Ferguson concluded that using raw sample weights or normalized sample weights alleviated the problem of biased parameter estimates with normalized weights giving an advantage.

As very minimal research has been conducted employing sample weights in single-level SEM and slightly more research has been done in the context of multilevel SEM, the research relating to multilevel SEM is provided in this methodological review of research. Stapleton (2002) conducted a Monte Carlo simulation study to determine how multilevel SEM performed under three conditions: (a) unweighted, (b) weighted using relative weights (normalized to reflect the actual sample size, the multilevel version of Kaplan and Ferguson's [1999] normalized weight), and (c) weighted using effective weights (in which the sum of the weights equals the sum of the weights squared). Using unweighted covariance matrices, Stapleton found that estimates for the path coefficient from the independent variable to the factor were negatively biased in relation to the population values due to lower path values in the oversampled strata. The more unequal the weights, the more biased the results. Group size and number of groups were not related to the path estimate bias. Using relative weighted covariance matrices, more accurate point estimates of the path coefficient were found. However, slight negative biases for the within-group residual and disturbance terms and positive biases for the between-group residual variance estimates were found. Using effective weighted covariance matrices, parameter point estimates were produced that were very close to the population values (within .5%).

Homogeneous Clusters

Multistage sampling is the process of subsampling clusters so that the elements selected are obtained from selecting sampling units in two or more stages (Kish, 1965). Many national datasets involve multistage sampling in which geographic regions are first selected, then institutions, and finally students (e.g., U.S. Department of Education, 2004). When multistage sampling is not addressed in the analysis, standard errors may be underestimated (Hahs-Vaughn, 2005; Kish & Frankel, 1974; Korn & Graubard, 1991, 1995a; Muthen & Satorra, 1995; Stapleton, 2002; Thomas & Heck, 2001). The homogeneities that exist due to multistage sampling can be accommodated through multilevel modeling; however, not all researchers may be interested in multilevel modeling. Likewise, the available datasets may not have the appropriate institution-level variables for the specified model (Kaplan & Elliott, 1997b). In these instances, design effects may be incorporated in the analysis to address clustering.

A design effect is the ratio of the estimated variance to the exact variance. Design effect corrections are needed to adjust the variances produced by the analysis (Peng, 2000). To correct for potential biased standard error estimates, four strategies have been suggested (Thomas & Heck, 2001), including (a) using specialized software, such as AM, SUDAAN, or WESVAR, that provides the option of defining weight, strata, and cluster variables; (b) adjusting the test statistic value (e.g., dividing the *t* test statistic by the square root of the design effect or dividing the *F* statistic by the design effect); (c) adjusting the normalized weight by the design effect (creating a new design effect adjusted weight, which is computed as

$$NORMWT\left(\frac{1}{DEFF}\right),$$

where *NORMWT* is the normalized weight) and applying the design effect adjusted weight to the analysis; or (d) using a more conservative alpha level. The first strategy is the optimal solution (Hahs-Vaughn, 2005; Thomas & Heck). The second and third strategies have been determined to yield roughly equivalent results (Thomas & Heck). We used the third strategy in this study, adjusting the relative weight downward as a function of the design effect by creating a design effect adjusted weight. This strategy (as does the first strategy) allows the researcher to accommodate both oversampled groups and homogeneous clusters in one step.

Method

Data

The database used for this study is the Beginning Postsecondary Students Longitudinal Study (BPS:90/92/94). The NCES makes BPS available in two formats: public use and restricted (licensed) use files (Fitzgerald, Berkner, Horn, Choy, & Hoachlander, 1994). We used the restricted use file.

The BPS:90 is derived from the National Postsecondary Student Aid Study of 1990 (NPSAS:90). The NPSAS survey included undergraduate and graduate students who did, and who did not, receive financial aid. The BPS:90 sample comprises students who began postsecondary education for the first time at any time between July 1, 1989, and June 30, 1990 (Pratt et al., 1996). Both the BPS:90/92 and BPS:90/94 are follow-ups of a selected subset of respondents to the NPSAS:90 student survey. We used the base year (first-time beginning postsecondary students between July 1, 1989, and June 30, 1990), first follow-up (1992), and second follow-up (1994) in the present analysis.

The subset used in this study consisted of BPS:90/92/94 students who were first-time beginning students, U.S. citizens, and working on an associate's degree or higher (n = 1,629). Only slightly more than half were women (52%). Students were predominately White (82%), with 8% African American, 6% Hispanic, 4% Asian American, and .5% American Indian/Alaskan Native. First-generation students, of whom neither parent had more than a high school diploma, consisted of 29% of the subset, and non-first-generation students consisted of 71% of the subset.

Measures

The theoretical model for this study is based on a conceptual model of college impact developed by Terenzini, Springer, Yaeger, Pascarella, and Nora (1996) that was tested using data from the National Study of Student Learning (NSSL), a 3-year longitudinal national study of over 4,000 new students who entered 2- or 4-year colleges in 1992. According to Terenzini et al., what is known about first-generation students falls into three broad categories that reflect the chronological college-going process. The first category includes a student's college expectations and his or her resulting preparation and planning for college. The second category relates to the transition between high school or work and college. The third category encompasses the impact of college experiences on persistence and degree attainment. The model incorporates six constructs, including precollegiate traits, curricular patterns, in-class experiences, out-of-class experiences, institutional context, and learning outcomes. We selected those observed variables that most closely replicated the variables used in the original study (Terenzini et al.), although exact duplication was not possible given the different dataset.

Analysis began with a structural equation model using 23 observed variables and 6 latent variables (2 exogenous variables and 4 endogenous variables; Appendix). The latent variables were (a) precollegiate traits, (b) curricular experiences, (c) in-class experiences, (d) out-of-class experiences, (e) institutional context, and (f) learning outcomes.

To create the most parsimonious model, the model was respecified by removing variables with high correlations ($r > .80$) that were making negligible contribution to the model. Variables that are highly correlated do not account for different proportions of variance in the model (Lomax, 2001). The variable *hours of remedial instruction* was removed from the model. More than 80% of students had received no hours of remedial instruction; therefore, we determined that the variable did not substantively contribute to the model. Several variables were also removed because of lack of contribution to the model (e.g., $R^2 < .20$, a small effect) and problems with model specification. The final model is presented in Figure 1 where college experiences represent curricular experiences, in-class ex-

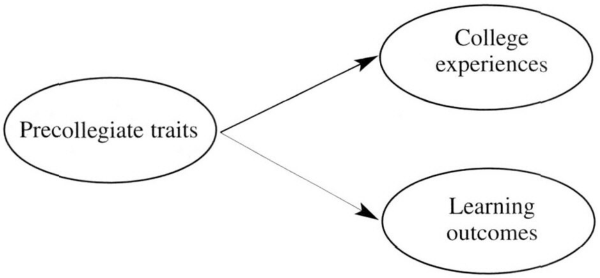

FIGURE 1. Final model of college impact.

periences, and out-of-class experiences. Correlations, means, and standard deviations for the weighted data are presented in Table 1; unweighted data are presented in Table 2.

Results

We performed the analysis using both a design effect adjusted sample weight and ignoring the weight. Because this is a single-level model, applying a weight without also incorporating a design effect correction is inappropriate. Applying the design effect adjusted weight adjusts for both the oversampling and multistage cluster sampling. We compared the confidence intervals of the parameter estimates for the two models to determine the effect that the design effect adjusted weight had on single-level SEM. No overlap in the confidence intervals indicates that the difference is statistically significant, and overlap indicates that the difference is not significant (Schenker & Gentleman, 2001).

Imputation of Missing Values

Imputation was used to obtain a complete dataset. This imputation procedure uses the expected maximization (EM) algorithm method (du Toit & Mels, 2002). With this method of imputation, the substituted value for the missing value is obtained from another case that has a similar response pattern over the other variables in the analysis (Joreskog & Sorbom, 1996). Approximately half of the variables had some missing values, although, on average, there was no more than 5% missing per variable. All missing values were successfully imputed.

Weight and Design Effect

The weight and the design effect are basic elements needed for a complete analysis of survey data (Lee et al., 1989). Because we used data where disproportionate sampling has been used, it was necessary to apply a sampling weight. The weight used was the panel survey weight, which includes longitudinal weights for comparing responses to the 1990, 1992, and 1994 cross-sectional populations based on those students who responded in all surveys (Pratt et al., 1996). Employing sampling weights makes the data analysis results generalizable to the national sample of first-time beginning postsecondary students.

It was also necessary to make a design effect correction to the standard error because the BPS:90/92/94 used a multistage cluster sample. For a similarly constructed survey also conducted by NCES, the NELS:88, when the design effect for a dependent variable used in a study is not reported in the technical reports, the design effect for a similar variable, the average design effect averaged over a set of variables, or the average design effect of the dependent variable averaged

TABLE 1. Correlations, Means, and Standard Deviations of College Impact Between First-Generation and Non-First-Generation College Students (Weighted)

Variable	1	2	3	4	5	6	7	8	9	10
1. Degree attained	—	.468	.560	.350	.262	.398	.154	.092	.278	.293
2. Cumulative grade point average	.438	—	.417	.183	.149	.162	.110	.109	.223	.218
3. Aspirations for education	.574	.313	—	.325	.278	.207	.182	.167	.561	.279
4. Nonacademic experiences	.267	.070	.335	—	.520	.295	.085	.193	.294	.259
5. Academic experiences	.223	.077	.306	.589	—	.241	.057	.112	.201	.133
6. Intensity of enrollment	.443	–.093	.224	.424	.341	—	.082	.059	.077	.257
7. Father's education	.133	.142	–.064	.046	.017	–.060	—	.128	.133	.122
8. Mother's education	.089	.106	–.067	.017	.052	.014	.771	—	.167	.091
9. Expected highest level of education	.214	.092	.476	.263	.260	.126	.072	.017	—	.231
10. Entrance exam score	.228	.081	.272	.230	.244	.240	.265	.319	.249	—
First-generation students										
M	1.77	3.92	3.12	8.96	6.93	2.69	1.58	1.70	4.65	.91
SD	1.81	.96	.88	5.48	3.33	.63	.68	.59	1.74	1.30
Non-first-generation students										
M	2.23	4.00	3.51	10.77	7.58	2.81	4.90	4.34	5.39	1.31
SD	1.85	.97	.69	5.68	3.04	.51	1.98	1.84	1.67	1.58

Note. First-generation students are below the diagonal; non-first-generation students are above the diagonal.

TABLE 2. Correlations, Means, and Standard Deviations of College Impact Between First-Generation and Non-First-Generation College Students (Unweighted)

Variable	1	2	3	4	5	6	7	8	9	10
1. Degree attained	—	.470	.574	.332	.238	.330	.214	.146	.274	.236
2. Cumulative grade point average	.442	—	.431	.154	.130	.161	.168	.124	.208	.241
3. Aspirations for education	.567	.323	—	.350	.256	.242	.256	.251	.553	.322
4. Nonacademic experiences	.267	.077	.360	—	.447	.270	.138	.150	.301	.205
5. Academic experiences	.215	.076	.299	.530	—	.212	.062	.077	.173	.065
6. Intensity of enrollment	.258	–.019	.119	.358	.275	—	.097	.099	.148	.162
7. Father's education	.140	.187	.063	.128	.090	.017	—	.262	.199	.184
8. Mother's education	.074	.074	-.001	.069	.071	.087	.742	—	.172	.138
9. Expected highest level of education	.201	.098	.517	.243	.241	.027	.048	.037	—	.236
10. Entrance exam score	.249	.119	.343	.231	.180	.074	.221	.250	.249	—
First-generation students										
M	2.02	3.96	3.26	10.58	7.80	2.86	1.59	1.73	4.91	1.23
SD	1.83	1.00	.82	5.47	3.00	.43	.69	.56	1.75	1.43
Non-first-generation students										
M	2.57	4.16	3.60	12.56	8.14	2.89	5.03	4.52	5.69	1.81
SD	.181	.99	.62	5.47	2.73	.38	1.96	1.85	1.64	1.74

Note. First-generation students are below the diagonal; non-first-generation students are above the diagonal.

over subgroups of the independent variable is appropriate to use (Huang, Salvucci, Peng, & Owings, 1996). The BPS:90/92/94 technical report (Pratt et al., 1996) provides design effects for only two dependent variables, both of which we used in this study: (a) highest undergraduate degree attained and (b) persistence/attainment. Analysis using NCES data that is on a national level can use the mean of the design effects to make an adjustment (Peng, 2000). To compute the design effect appropriate for this study, therefore, we averaged the average design effect of these two dependent variables over subgroups of independent variables. The resulting average design effect was 2.4721. In this study, we made the design effect correction by computing and applying a design effect adjusted weight. The design effect adjustment was used in combination with the raw weights to create a design effect adjusted weight by dividing the raw weights by the mean weight and mean design effect. The weight used thereby reflects the effective sample size—the original sample size adjusted for clustering (Kaplan & Ferguson, 1999; Peng) and is one of the suggested strategies for simultaneously addressing oversampling and cluster sampling (Thomas & Heck, 2001).

Analyses

Measurement and structural models were produced. The confidence intervals of the parameter estimates of the single sample models using design effect adjusted weights (for ease in reference, *design effect adjusted weights* will be referred to as *weights* throughout the analyses section) and models not using weights were compared to determine the effect that sample weights had on single-level SEM.

Baseline models. All model tests were based on the asymptotic covariance matrix and used weighted least squares (WLS) in LISREL version 8.5. WLS is also the estimation method recommended when data have mixed scales, such as ordinal and continuous (Byrne, 1998).

In model testing, it is suggested that the measurement model be assessed first and then the structural model (Anderson & Gerbing, 1988; Byrne, 1998; Schumacker & Lomax, 1996). We assessed adequacy of the measurement model by reviewing parameter estimates and overall fit of the measurement model (Byrne). The fit of the measurement models in this study was assessed by reviewing how well the models were represented by the observed variables. These indicators and criteria for adequate fit included factor loadings with correct direction (positive or negative) loading on the appropriate latent variable, small standard errors, significant test statistics, and moderate or strong squared multiple correlations (R^2).

We generated four single sample models in this analysis: (a) first-generation students with design effect adjusted weight applied, (b) first-generation students without design effect adjusted weight applied (i.e., unweighted), (c) non-first-generation students with design effect adjusted weight applied, and (d) non-first-generation students without design effect adjusted weight applied (i.e., unweighted). The measurement and structural model assessments for each subset relative to the weighted and unweighted model are described first. Comparison of confidence intervals for the weighted versus unweighted baseline models are then presented. Sample sizes of the unweighted models reflect the actual sample size. Sample sizes differ for the weighted models reflecting the effective sample size because of the weighting. Evaluating the models to determine acceptable levels of model fit followed the recommendations of Hu and Bentler (1995). Values closer to 1 for GFI, AGFI, and the NFI and values less than .05 for *RMSEA* and ECVI indicate good model fit. Confidence intervals for *RMSEA* and ECVI are provided in Table 3 for both the weighted and unweighted models.

First-generation students, weighted. The subset of first-generation students when weights are applied resulted in a sample size of 474. Each of the observed variables in the final measurement model had a significant ($p < .01$) positive loading on the appropriate latent variable. All structural paths were significant ($p < .01$). The GFI, AGFI, ECVI, and NFI indicate an overall good fit. The *RMSEA* and root mean square residual (*RMR*), however, suggest a poor fit. The final measurement and structural model and goodness of fit indices for first-generation students when weights are applied are presented in Table 4.

First-generation students, unweighted. The subset of first-generation students when no weights are applied resulted in a sample size of 1,170. Each of the observed variables in the final measurement model had a significant ($p < .01$) positive loading on the appropriate latent variable. The model was reviewed for fit, and all paths were significant ($p < .01$). The GFI and AGFI indicate a relatively good fit. However, the other goodness of fit indices (*RMR*, *RMSEA*, NFI, and ECVI) indicate poor fit. The final measurement and structural model and goodness of fit indices for first-generation students when weights are not applied are presented in Table 4.

TABLE 3. Single-Sample Fit Indices 90% Confidence Intervals

	90% confidence interval	
Sample	*RMSEA*	ECVI
First-generation, weighted	.144, .169	.898, 1.188
First-generation, unweighted	.0992, .116	.408, .533
Non-first-generation, weighted	.0815, .0983	.298, .403
Non-first-generation, unweighted	.0590, .0698	.149, .198

TABLE 4. Measurement and Structural Models and Goodness of Fit Indices, First-Generation Students

		Estimate		*t*		*SE*		R^2	
Latent variable	Observed variable	W	U	W	U	W	U	W	U
		Measurement model							
Precollegiate traits	Father's education	1.000	1.000	—	—	—	—	.659	.696
	Mother's education	1.000	.905	16.264	16.755	.0615	.0504	.659	.570
	Expected highest level of education	.657	.688	15.495	17.641	.0424	.0390	.336	.390
	Entrance exam score	.650	.601	16.064	16.190	.0405	.0371	.291	.252
College experiences	Academic experiences (1992)	1.000	1.000	—	—	—	—	.413	.490
	Intensity of enrollment	.920	.955	8.538	14.575	.1080	.0655	.350	.447
	Nonacademic experiences (1992)	.877	.692	16.072	7.930	.0545	.0872	.318	.235
Learning outcomes	Degree attained	1.000	1.000	—	—	—	—	.562	.529
	Aspirations for education (1994)	.854	1.036	27.661	33.248	.0319	.0379	.500	.529
	Cumulative grade point average	.744	.661	17.223	17.961	.0432	.0368	.350	.253

		Estimate		*t*		*SE*		R^2	
Path to	Path from	W	U	W	U	W	U	W	U
		Structural model							
College experiences	Precollegiate traits	.337	.342	9.119	10.225	.0369	.0334	.181	.166
Learning outcomes	College experiences	.536	.240	6.016	4.473	.0890	.0537	.362	.302
Learning outcomes	Precollegiate traits	.283	.474	5.151	11.307	.0550	.0419	.362	.302

Notes. W = weighted results; U = unweighted results. — = fixed parameter. Weighted: χ^2 = 428.290; GFI = .938; AGFI = .903; RMR = .160; *RMSEA* = .154; NFI = .805; ECVI = .990. Unweighted: χ^2 = 506.283; GFI = .963; AGFI = .943; *RMR* = .145; *RMSEA* = .107; NFI = .764; ECVI = .467.

Non-first-generation students, weighted. The subset of non-first-generation students when weights are applied resulted in a sample size of 1,155. All structural paths were significant ($p < .01$). The GFI, AGFI, and NFI indicate a relatively good fit. The *RMSEA*, as do other indices, indicates a nearly acceptable fit. The final measurement and structural model and goodness-of-fit indices for non-first-generation students when weights are applied are presented in Table 5.

Non-first-generation students, unweighted. The subset of non-first-generation students when no weights are applied resulted in a sample size of 2,755. All structural paths were significant ($p < .01$). Overall, the fit indices suggest a reasonable fit. The final measurement and structural model and goodness-of-fit indices for non-first-generation students when weights are not applied are presented in Table 5.

Final parameter estimates of weighted and unweighted groups. Standardized solutions of first-generation and non-first-generation students, both with and without weights, are provided in Figures 2 and 3.

Baseline Model Confidence Interval Results

Results discussed include single-sample measurement model parameter estimate confidence intervals and structural model parameter estimate confidence intervals.

Confidence intervals of parameter estimates. In reviewing 90% confidence intervals of parameter estimates in the measurement models, we found a significant difference for weighted versus unweighted models for one variable in the first-generation student model and four variables in the non-first-generation student model. Confidence intervals for parameter estimates were estimated as $\beta \pm 1.645(s_b)$, where s_b is the standard error (Lomax, 2001). If there is no overlap in the confidence intervals, there is a significant difference between the weighted and unweighted models (Schenker & Gentleman, 2001). In the model for first-generation students, we found a significant difference in the parameter estimate for the variable *aspirations for education* when comparing weighted versus unweighted models. In the model for non-first-generation students, we found a significant difference in the parameter estimate for the variables *mother's education, entrance exam scores, expected highest level of education*, and *nonacademic experiences* when comparing weighted versus unweighted models.

In reviewing the 90% confidence intervals of the parameter estimates in the structural models, we found a significant difference for two paths for first-generation and non-first-generation students when looking at weighted versus unweighted models. Specifically, both paths to *learning outcomes* (from college experiences and from precollegiate traits) were significantly different when

TABLE 5. Measurement and Structural Models and Goodness of Fit Indices, Non-First-Generation Students

		Estimate		*t*		*SE*		R^2	
Latent variable	Observed variable	W	U	W	U	W	U	W	U
		Measurement model							
Precollegiate traits	Father's education	1.000	1.000	—	—	—	—	.5230	.163
	Mother's education	1.246	.848	8.349	13.232	.1490	.0641	.0777	.117
	Entrance exam score	1.974	1.132	9.922	15.660	.1990	.0723	.2040	.217
	Expected highest level of education	3.156	1.749	10.315	18.503	.3060	.0945	.5480	.539
College experiences	Academic experiences (1992)	1.000	1.000	—	—	—	—	.3410	.248
	Intensity of enrollment	.714	.708	13.141	9.116	.0543	.0776	.1740	.124
	Nonacademic experiences (1992)	1.159	1.505	21.732	15.979	.0534	.0942	.4590	.561
Learning outcomes	Degree attained	1.000	1.000	—	—	—	—	.5830	.568
	Aspirations for education (1994)	1.091	1.082	44.628	55.064	.0244	.0196	.6250	.607
	Cumulative grade point average	.695	.705	30.042	29.860	.0231	.0236	.2680	.270

		Estimate		*t*		*SE*		R^2	
Path to	Path from	W	U	W	U	W	U	W	U
		Structural model							
College experiences	Precollegiate traits	1.722	.737	8.618	10.965	.200	.0672	.435	.358
Learning outcomes	College experiences	.218	.211	2.956	2.886	.0739	.0731	.760	.743
Learning outcomes	Precollegiate traits	2.554	1.429	8.182	13.213	.3120	.1080	.760	.743

Notes. W = weighted results; U = unweighted results. — = fixed parameter. Weighted: χ^2 = 360.663; GFI = .975; AGFI = .960; RMR = .0627; *RMSEA* = .0898; NFI = .867; ECVI = .34. Unweighted: χ^2 = 433.982; GFI = .986; AGFI = .977; *RMR* = .0587; *RMSEA* = .0643; NFI = .838; ECVI = .172

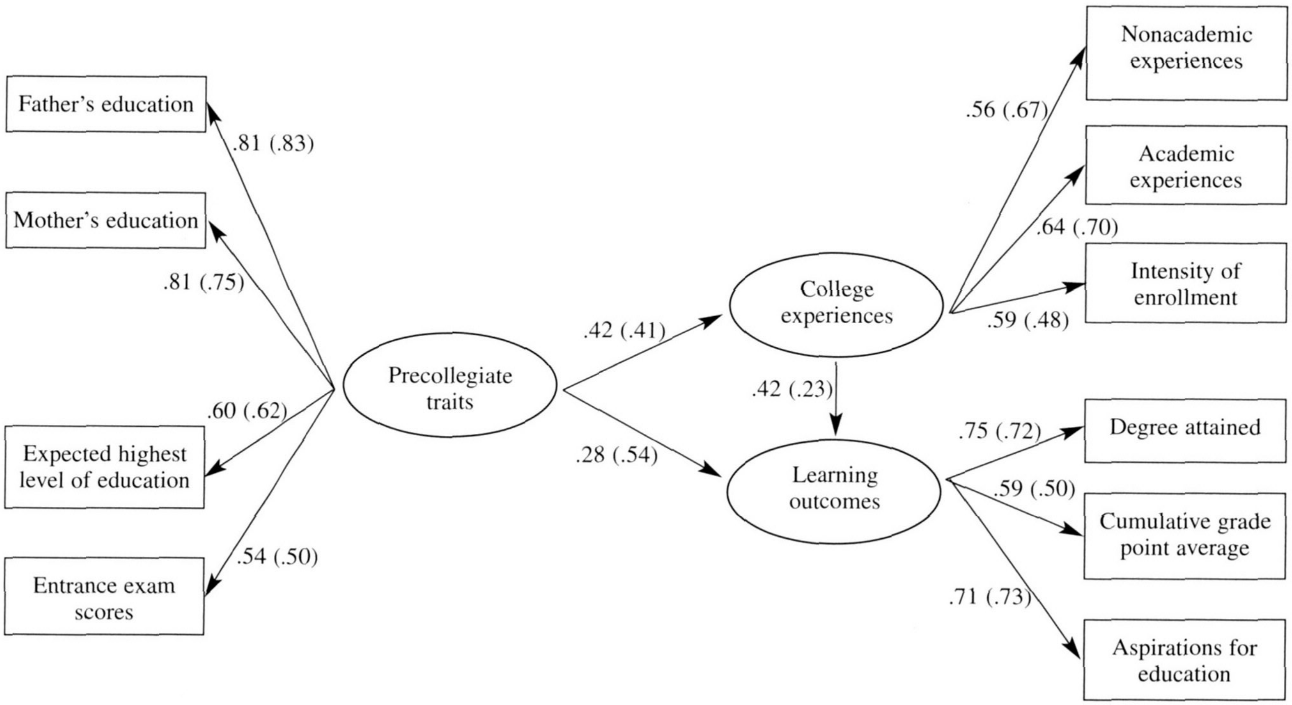

FIGURE 2. Standardized solution of first-generation students. ***Note.*** **First-generation students with weights are outside parentheses.**

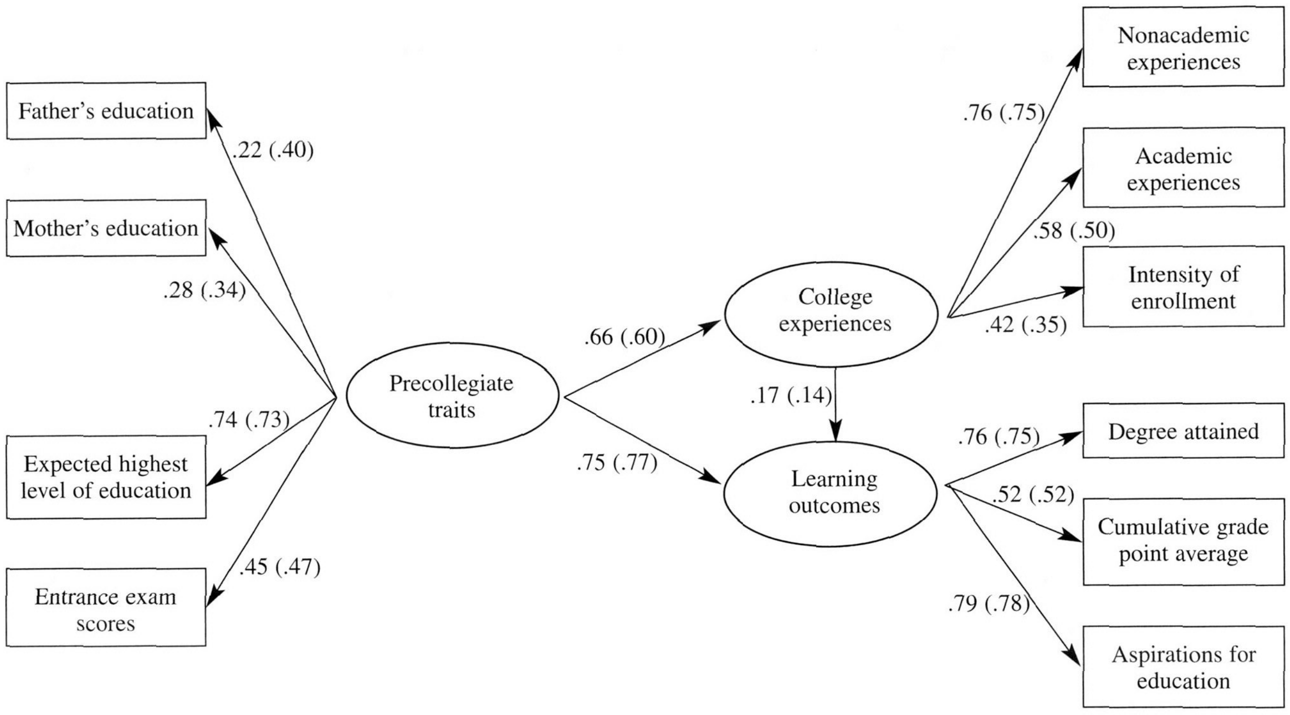

FIGURE 3. Standardized solution of non-first-generation students. ***Note.*** **Non-first-generation students with weights are outside parentheses.**

reviewing the weighted versus unweighted models for first-generation students. The paths from *precollegiate traits* to *college experiences* and to *learning outcomes* were significantly different when reviewing weighted versus unweighted models for non-first-generation students. Comparisons of the measurement and structural model confidence intervals are presented in Table 6.

Discussion

National datasets usually employ multistage sampling and disproportionate sampling. This sampling design, however, violates the assumption of independence and creates problems with traditional standard error formulas that are based on simple random samples (Lee et al., 1989). Incorporating sample weights and design effects in the analysis is one way to address these issues when specialized software is not available (Hahs-Vaughn, 2005; Thomas & Heck, 2001). In the one published study that analyzed the effect of weights in single-level SEM, Kaplan and Ferguson (1999) found that when weights are not used, serious bias in the latent variable model resulted. The impact of the use or lack of use of weights in SEM based on the present study and reviewed in context of previous research is presented for (a) effects on parameter estimates and (b) effects on standard errors.

Weighted Versus Unweighted: Effects on Parameter Estimates

The confidence intervals provide evidence that a significant difference exists in parameter estimates of design effect adjusted weighted versus unweighted models. However, differences were most pronounced for non-first-generation students with significant differences affecting four of the seven nonfixed factor loadings and two of the three structural coefficients. Of all parameters estimated for first-generation students, only one factor loading and two structural coefficients were significantly different based on confidence intervals. Mixed results were found in relation to biased parameter estimates. For some parameter estimates, ignoring oversampling and cluster sampling by failing to apply design effect adjusted weights resulted in bias. However, for other parameter estimates, ignoring weights did not result in bias—this is seen especially in the first-generation model in which only one factor loading was significantly different for the weighted versus unweighted models. Differences between the first-generation and non-first-generation models may be due to sample size as the first-generation model was substantially smaller (n = 474) compared with non-first-generation students (n = 1,155). Greater bias was seen in the structural coefficient parameter estimates.

The importance in the significant differences between the weighted and unweighted models relates to generalization. The unweighted models reflect,

TABLE 6. Single Sample Measurement and Structural Model Parameter Estimate Confidence Intervals

		First generation		Non-first generation	
Latent variable	Observed variable	Weighted	Unweighted	Weighted	Unweighted
		Measurement model			
Precollegiate traits	Father's education	1.000	1.000	1.000	1.000
	Mother's education	1.1011, .8988	.8221, .9879	1.0009, 1.4911*	.7426, .9538*
	Expected highest level of education	.5872, .7197	.6238, .7522	2.6526, 3.6593*	1.5935, 1.9046*
	Entrance exam score	.5834, .7166	.5400, .6620	1.6467, 2.3013*	1.0131, 1.2509*
College experiences	Academic experiences (1992)	1.000	1.000	1.000	1.000
	Intensity of enrollment	.7423, 1.0977	.5486, .8354	.6247, .8033	.5804, .8357
	Nonacademic experiences (1992)	.7874, .9666	.8473, 1.0627	1.0713, 1.2467*	1.3500, 1.6600*
Learning outcomes	Degree attained	1.000	1.000	1.000	1.000
	Aspirations for education (1994)	.8488, .9065*	.9737, 1.098*	1.0509, 1.1311	1.0498, 1.1142
	Cumulative grade point average	.6729, .8151	.6005, .7215	.6570, .7215	.6661, .7438

		First generation		Non-first generation	
Path to	Path from	Weighted	Unweighted	Weighted	Unweighted
		Structural model			
College experiences	Precollegiate traits	.3163, .4377	.2871, .3969	1.393, 2.051*	.6260, .8475*
Learning outcomes	College experiences	.3896, .6824*	.1517, .3280*	.0964, .3396	.0908, .3312
Learning outcomes	Precollegiate traits	.1925, .3735*	.4051, .5429*	2.0408, 3.0672*	1.2513, .1777*

*$p < .10$.

among others, oversampled groups, and thus the unweighted results will mirror the sample and not the population. Where the groups differ for the weighted and unweighted models (in the measurement model and the structural coefficients) is the point at which the researcher can generalize to either the sample or the population. It should be noted, however, that this method of examining overlapping confidence intervals is more conservative (increased chance of Type I error) and is less powerful (increased chance of Type II error) than the standard technique of testing significance under the assumptions of consistency, asymptotic normality, and asymptotic independence of estimates (Schenker & Gentleman, 2001).

Weighted Versus Unweighted: Effects on Standard Errors

The analysis of parameter estimates on standard errors for first-generation students shows that for five of the seven observed variables that are not fixed and for all of the structural coefficients, the standard errors are higher for the design effect adjusted weighted models. For non-first-generation students, four of the seven observed variables and all of the structural coefficients in the weighted models have higher standard errors. Differences in standard errors then influence the final results, including the test statistic, t. Although it is impossible to determine the actual population standard errors in this study given the nature of using an extant dataset, the findings presented do suggest a substantial impact on standard errors when weights are used compared with when they are ignored in the analyses. Specifically, the unweighted data produce underestimated standard errors.

Recommendations

Consideration must always be given to the design of a sample and how that may affect the analysis. This thought has been echoed by others (Kalton, 1983, 1989; Korn & Graubard, 1991; Thomas & Heck, 2001). When using a design-based approach in structural equation modeling, accommodation of both oversampling and homogeneous groups is needed to ensure that the estimates are accurate. In this study, we illustrated how estimates may differ when the complex design is not accommodated in analyses. Various strategies have been recommended to accommodate issues inherent in complex samples. The use of specialized software is the most appropriate way to effectively handle complex samples (Thomas & Heck). Researchers may be familiar with such software as WESVAR, AM, and SUDAAN. However, these software programs are not designed for SEM analyses, and it was not until recently that a specialized software program became available for SEM analysis of complex samples. The newest version of LISREL (v. 8.7) is designed so that weights, strata, and cluster variables can be defined directly. This alleviates the researcher having to create a design effect adjusted weight. However, it is anticipated that, because of the newness of this version's release, many SEM researchers do not have access to it yet. In those cases, analysis of complex samples in single-level SEM can be performed by applying a design effect adjusted weight to the data in PRELIS. Regardless of the approach to handle the design features of complex samples (specialized software or design effect adjusted weight), it is the accommodation of these issues that is critical—without which the results are biased and may lead to incorrect decisions (Hahs-Vaughn, 2005).

Weights are always recommended if the results of the study are to be generalized to the population representative in the survey. When weights are not used, the results can be generalized only to the sample of students who completed the survey (Hahs, 2003) because more weight is provided in the analyses to the oversampled groups, thus distorting the true population (Thomas & Heck, 2001). If and when researchers choose not to weight their data, they must recognize that their decision has given disproportionate weight to the oversampled groups in the dataset and the conclusions cannot generalize to the population intended because the results reflect disproportionately some groups of individuals. These recommendations are important not only for researchers who analyze complex samples but also journal editors and editorial review boards who review studies that report results from complex samples. Best practices for authors in reporting accommodation of oversampling and homogeneous clusters in results and journals in requirements for authors who use complex samples can be found in Hahs-Vaughn (in press). As stated by Kalton (1989), an unweighted sample represents only "a collection of individuals that represents no meaningful population" (p. 582).

Conclusion

There are various reasons why the results of this study do not completely parallel previous research. First, we used WLS estimation. Simulated research by Kaplan and Ferguson (1999) used maximum likelihood estimation. Second, we used an extant dataset, and previous studies related to methodological issues of complex samples in structural equation modeling have been conducted using simulated data. Simulated studies also concentrated on accommodating oversampled groups through the use of weights without addressing the homogeneities that may exist between groups due to the cluster sampling (i.e., design effect adjustment).

Third, Korn and Graubard (1995a) found that weighted and unweighted estimators may differ if a covariate is not included in the model. The results of this study may differ under other models and using different variables. The possibility that the results occurred because of model misspecification cannot be ruled out. This is obviously a limitation of using extant data to study methodological issues. However, the core of this study is to illustrate differential estimates that may occur when complex samples are appropriately analyzed by addressing oversampling and cluster sampling versus when these elements are ignored. Although there were some model fit indices that suggested less than acceptable fit, the results still add value to illuminating methodological considerations with complex sample.

Fourth, the sample size for first-generation students is substantially less than non-first-generation students in both the design effect adjusted weighted and unweighted models. Kaplan and Ferguson (1999) found that parameter estimate bias decreased as sample size increased when using the same population model. Differences in model fit from weighted to unweighted models may be due in part to not only the application or lack of application of weights and design effects but also the change from actual to effective sample size when weights and design effects are applied. In this study, we used different subsets of the sample (i.e., first-generation and non-first-generation students) and the sizes of the groups differed, and the group with the larger sample size (i.e., non-first-generation students) resulted in greater differences between weighted and unweighted models. Again, this may be due in part to the differences between the groups, however, and not necessarily the sample sizes. Although conclusions cannot be drawn from a single case as represented in this study, this research does contribute to understanding the effect of weights using actual data, thus complementing and contributing to simulation research that has examined methodological issues of weighting.

The use of national datasets usually requires weights to compensate for oversampling some populations; however, examining methodological issues of weighting in structural equation model has been studied by few (Hahs, 2003). Although analysis of weights within a single-level SEM context has been done, this research has not been fully exploited—and in fact, few published studies exist that specifically look at the methodology of weights in single-level SEM (e.g., Kaplan & Ferguson, 1999). In Kaplan and Ferguson's study, findings were limited in the simplicity of the model and the use of simulated data. There also appears to be a very limited number of methodological studies that have addressed the research using extant datasets rather than simulation. Those that have applied this research have done so in the context of multilevel models (e.g., Hill & Goldstein, 1998) or statistical methods other than structural equation modeling (DuMouchel & Duncan, 1983; Hahs-Vaughn, 2005; Korn & Graubard, 1991, 1995a). This study helps fill a void in analysis of weights in single-level structural equation modeling by providing researchers a foundation from which to examine how the application (or lack) of employing survey weights and design effects in a single-level structural equation model may effect estimates produced when analyzing complex survey data. Specifically, this study illustrates that, although it may be likely that there will be differences in parameter estimates, standard errors, and model fit when using weights versus not using weights, these differences may not be cross-cutting. Thus, this study also helps reiterate that a good starting place in working with complex samples is the suggested process of reviewing weighted and unweighted model fit prior to making decisions on which results to report (Skinner et al., 1989) and illustrates the type of errors and incorrect decisions that may be made if the design of the sample is not addressed in the analysis. Recommendations to SEM researchers are provided. This study may be especially valuable to SEM researchers who are less familiar with the methodological requirements of analyzing complex samples.

The findings presented are especially important to federal agencies that have numerous large datasets available for use by the public. As these datasets are promoted more widely, an increasingly diverse group of individuals is accessing the data. This study provides one tool for dataset program officers to help users of single-level structural equation modeling understand the implications of addressing, and conversely not addressing, the design features of complex samples.

REFERENCES

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411–423.

Association for Institutional Research. (n.d.). *AIR summer institutes*. Retrieved July 19, 2004, from http://www.airweb.org/page.asp?page=473

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

du Toit, S., & Mels, G. (2002). *Supplementary notes on multiple imputation*. Retrieved September 21, 2005, from http://www.ssicentral.com/lisrel/techdocs/imputation.pdf

DuMouchel, W. H., & Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association, 78*(383), 535–543.

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

Fitzgerald, R., Berkner, L., Horn, L., Choy, S., & Hoachlander, G. (1994). *Descriptive summary of the 1989–90 postsecondary students two years after entry: Contractor report* (No. 94-386). Washington, DC: National Center for Education Statistics.

Hahs, D. L. (2003). *The utilization of sample weights in structural equation modeling: An application using the Beginning Postsecondary Students Longitudinal Study 1990/92/94*. Tuscaloosa: University of Alabama.

Hahs-Vaughn, D. L. (2005). A primer for using and understanding weights with national datasets. *The Journal of Experimental Education, 73*(3), 221–248.

Hahs-Vaughn, D. L. (in press). Weighting omissions and best practices when using national datasets in educational research. *Professional File.*

Hill, P. W., & Goldstein, H. (1998). Multilevel modeling of educational data with cross-classification and missing identification units. *Journal of Education and Behavioral Statistics, 23*(1), 117–128.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*, 663–685.

Hox, J. J., & Kreft, I. G. G. (1994). Multilevel analysis methods. *Sociological Methods and Research, 22*(3), 283–299.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 76–99). Thousand Oaks, CA: Sage.

Huang, G., Salvucci, S., Peng, S., & Owings, J. (1996). *National Educational Longitudinal Study of 1988 (NELS:88) research framework and issues* (Working Paper No. 96-03). Arlington, VA: Synetics for Management Decisions.

Joreskog, K. G., & Sorbom, D. (1996). *PRELIS 2 user's reference guide*. Chicago: Scientific Software International, Inc.

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

Kalton, G. (1989). Modeling considerations: Discussion from a survey sampling perspective. In D. Kasprzyk, G. Duncan, G. Kalton, & M. Singh (Eds.), *Panel surveys* (pp. 575–585). New York: Wiley.

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

Kaplan, D., & Elliott, P. R. (1997b). A model-based approach to validating education indicators using multilevel structural equation modeling. *Journal of Education and Behavioral Statistics, 22*(3), 323–347.

Kaplan, D., & Ferguson, A. J. (1999). On the utilization of sample weights in latent variable models. *Structural Equation Modeling, 6*(4), 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.

Korn, E. L., & Graubard, B. I. (1991). Epidemiologic studies utilizing surveys: Accounting for the sampling design. *American Journal of Public Health, 81*(9), 1166–1173.

Korn, E. L., & Graubard, B. I. (1995a). Analysis of large health surveys: Accounting for the sampling design. *Journal of the Royal Statistical Society, Series A (Statistics in Society), 158*(2), 263–295.

Korn, E. L., & Graubard, B. I. (1995b). Examples of differing weighted and unweighted estimates from a sample survey. *American Statistician, 49*, 291–305.

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

Lomax, R. G. (2001). *An introduction to statistical concepts for education and behavioral sciences*. Mahwah, NJ: Erlbaum.

Muthen, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22*(3), 376–398.

Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25*, 267–316.

Peng, S. S. (2000, July). *Technical issues in using NCES data*. Paper presented at the AIR/NCES National Data Institute on the Use of Postsecondary Databases, Gaithersburg, MD.

Perna, L. (2003). The status of women and minorities among community college faculty. *Research in Higher Education, 44*(2), 205–240.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61*(2), 317–337.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B, 60*(1), 23–40.

Pratt, D. J., Whitmore, R. W., Wine, J. S., Blackwell, K. M., Forsyth, B. H., Smith, T. K., et al. (1996). *Beginning postsecondary students longitudinal study second follow-up (BPS:90/92/94) final technical report*. Washington, DC: National Center for Education Statistics.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician, 55*(3), 182–187.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Erlbaum.

Skinner, C. J., Holt, D., & Smith, T. M. F. (Eds.). (1989). *Analysis of complex samples*. New York: Wiley.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling, 9*(4), 475–502.

Terenzini, P. R., Springer, L., Yaeger, P. M., Pascarella, E. T., & Nora, A. (1996). First-generation college students: Characteristics, experiences, and cognitive development. *Research in Higher Education, 37*(1), 1–22.

Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education, 42*(5), 517–540.

U.S. Department of Education. (2004). *User's manual for the ECLS-K third grade public-use data file and electronic code book* (No. NCES 2004-001). Washington, DC: Author.

APPENDIX
Variables

Precollegiate Traits

Compared to others, above average ability index (Scale: 0–12)
Composite variable of 12 items that asked students to compare themselves with others on academic ability, drive to achieve, writing ability, and similar items; the higher the score, the more indicators the students ranked themselves as above average compared to peers.

Compared to others, below average ability index (Scale: 0–8)
Composite variable of 8 items that asked students to compare themselves with others on academic ability, drive to achieve, writing ability, and similar items; the higher the score, the more indicators the students ranked themselves as below average compared to peers.

Entrance exam score (Scale: 0–5)
0 = did not take the ACT or SAT
1 = marginally or not qualified
2 = minimally qualified
3 = somewhat qualified
4 = highly qualified
5 = very highly qualified

Expected highest level of education (Scale: 1–8)
1 = less than 1 year trade school
2 = 1–2 years trade school
3 = 2+ years trade school
4 = less than 2 years college
5 = 2 years or more college
6 = bachelor's degree
7 = master's degree
8 = PhD/professional degree

Father's education (Scale: 0–7)
0 = unknown
1 = less than high school
2 = high school graduate
3 = trade school
4 = less than 2 years college
5 = 2 or more years college
6 = bachelor's degree
7 = postgraduate/professional

Mother's education (Scale: 0–7)
0 = unknown
1 = less than high school
2 = high school graduate
3 = trade school
4 = less than 2 years college
5 = 2 or more years college

6 = bachelor's degree
7 = postgraduate/professional

Mother work outside home (Scale: 0–8)
Higher values represent more time spent outside the home working with the highest value representing that the mother is not present in the home.

Percentage of income to poverty (Scale: 0–8,834)
Values of 100 or less indicate that the family is at or below the poverty level.

Socioeconomic status (SES; Scale: 1–96)
Composite variable combining parents' occupation, dependents' family income, and things in the home; lower values represent lower SES.

Intensity of enrollment (Scale: 1–3)
1 = part-time both years
2 = part-time and full-time
3 = full-time both years

In-Class Experiences

Academic experiences (Scale: 0–12)
Composite variable that measured how often the student talked with faculty, met with advisor, and participated in study groups; higher values represent more frequent contact.

Remedial hours of instruction (0–1,200)
Remedial hours of instruction in math, reading, study skills, and writing; higher values represent more hours of remedial instruction.

Out-of-Class Experiences

Hours worked per week (Scale: 0–70)
Average hours worked per week during the months enrolled in school

Nonacademic experiences (0–28)
Composite variable that measured how often the student had social contact with faculty, participated in school clubs, and similar nonacademic experiences; higher values represent more frequent contact.

Institutional Context

Control and enrollment of institution (Scale: 1–14)
Public/private sector and enrollment indicator

Satisfaction with first institution (Scale: 0–5)
Average satisfaction with NPSAS institution; higher values represent greater satisfaction.

Learning Outcomes

Application to graduate school (Scale: 0–1)
0 = not interested, not applied
1 = applied to graduate school

Aspirations for education (Scale: 1–4)
1 = trade school
2 = 2–3 years of college
3 = bachelor's degree
4 = advanced degree

Cumulative grade point average (Scale: 1–6)

1 = less than Cs
2 = mostly Cs
3 = Bs and Cs
4 = mostly Bs
5 = As and Bs
6 = mostly As

Degree attained (Scale: 0–4)
0 = none
1 = attained unknown degree
2 = certificate
3 = associate's
4 = bachelor's

Job satisfaction (Scale: 0–21)
Composite variable of job satisfaction (e.g., pay and fringe benefits, promotion opportunities, job security, and similar items); higher values represent greater satisfaction.

Job/education application (Scale: 0–12)
Composite variable of the application of what was learned in school to the job, whether job was similar to education, and similar items; higher score represents more correspondence between the job and the education.

Persistence (Scale: 0–3)
First type of departure from persistence track; higher values represent greater persistence.