Research in Higher Education, Vol. 42, No. 5, 2001

ANALYSIS OF LARGE-SCALE SECONDARY DATA IN HIGHER EDUCATION RESEARCH: Potential Perils Associated with Complex Sampling Designs

Scott L. Thomas and Ronald H. Heck

Most large-scale secondary data sets used in higher education research (e.g., NPSAS or BPS) are constructed using complex survey sample designs where the population of interest is stratified on a number of dimensions and oversampled within certain of these strata. Moreover, these complex sample designs often cluster lower level units (e.g., students) within higher level units (e.g., colleges) to achieve efficiencies in the sampling process. Ignoring oversampling (unequal probability of selection) in complex survey designs presents problems when trying to make inferences—data from these designs are, in their raw form, admittedly nonrepresentative of the population to which they are designed to generalize. Ignoring the clustering of observations in these sampling designs presents a second set of problems when making inferences about variability in the population and testing hypotheses and usually leads to an increased likelihood of committing Type I errors (declaring something as an effect when in fact it is not). This article presents an extended example using complex sample survey data to demonstrate how researchers can address problems associated with oversampling and clustering of observations in these designs.

KEY WORDS: survey analysis; complex sample; clustered data.

INTRODUCTION

Research in higher education has benefited tremendously from the increased availability of quality secondary data sets germane to our varied interests. Academic researchers now have available a host of nationally representative data with which they can systematically examine outcomes such as student college

Scott L. Thomas and Ronald Heck are affiliated with the University of Hawai'i at Mānoa.

Address correspondence to: Scott L. Thomas, Department of Educational Administration, University of Hawai'i at Mānoa, 1776 University Avenue, Honolulu, HI 96822.

0361-0365/01/1000-0517$19.50/0

going, persistence, postgraduation labor force participation, faculty pay and workload equity, and faculty entrepreneurialism, to name only a few. Like their academic counterparts, institutional researchers are increasingly using these data to make comparisons to students, administrators, and faculty at their institutions as well as comparisons among institutions themselves.

Perhaps the most widely used secondary data sources in higher education research are those overseen by the U.S. Department of Education through the National Center for Education Statistics (NCES). NCES produces numerous surveys touching on almost every facet of higher education. These surveys are both longitudinal and cross-sectional in nature and include familiar names such as High School and Beyond (HS&B), the National Postsecondary Student Aid Study (NPSAS), the National Study of Postsecondary Faculty (NSOPF), the Beginning Postsecondary Student Longitudinal Study (BPS), the Baccalaureate and Beyond Study (B&B), and the Integrated Postsecondary Education Data System (IPEDS).

The relatively high quality data associated with these and other surveys are enticing to researchers and graduate students alike. A recent review of national secondary data resources informing higher education interests identified 65 conference presentations, 30 published reports, and 91 refereed journal articles that used one or more of the national secondary data sources reviewed (Dey, et al., 1997). In addition to the analyses identified in that review, anecdotal evidence exists that such data are widely used as working examples in graduate-level applied statistics courses taught in schools of education throughout the United States.

The ease of availability of these data combined with well-publicized financial incentives for their use and increasingly sophisticated technology that permits powerful analysis of large data sets has led many to rightfully view such data as an exciting research opportunity. But, while we have evolved significantly from the days of punch cards, 9–track tape downloads, and complicated mainframe programming, a number of important statistical issues associated with analyzing secondary survey data remain. In fact, it would seem that the fruits of these technological advances have in some instances overshadowed a number of basic analytical issues that can confound the interpretation of findings based on many of these large-scale nationally representative data sets. As the richness of such data has great appeal to policymakers and analysts, we argue that ignorance of some of the more fundamental analytical issues associated with the use of these data may often yield results that could mistakenly lead us down the wrong policy paths.

This article addresses two separate but related, longstanding analytical issues associated with the use of data collected through complex sampling designs—designs that are most often used in the large-scale data collection efforts of government agencies. These issues can be classified into two broad areas: (1)

representativeness of the samples analyzed, and (2) the correct assessment of population variances that form the basis for the identification of statistical effects and hypothesis testing. In this article, we address these problems in an applied, relatively nontechnical manner in an effort to illuminate the important, but not so obvious, problems associated with the analysis of many large-scale secondary data sets. The main points are demonstrated using a subsample of data from NCES's Baccalaureate & Beyond survey (B&B:93/94), which collected information on over 11,000 graduates receiving their first baccalaureate degree in the 1992/1993 academic year. The survey was designed to be a nationally representative sample of baccalaureate recipients in the United States.

PROBLEMS WITH LARGE-SCALE SAMPLE SURVEY DATA

There are two basic data collection problems encountered when collecting large-scale sample survey data. First, with many large-scale surveys, there exists no simple sampling frame (i.e., a single list from which we can randomly choose our sample members) for our target population. Second, even if a sampling frame existed, we would most likely want to make sure that we had a sufficient number of respondents with certain characteristics (e.g., certain racial/ethnic groups, those at different types of postsecondary institutions). Knowing that many are interested in the analysis of certain segments of the population, a simple random sample (SRS)—where one randomly chooses a certain number of respondents for a sample—might not yield adequate numbers of observations in the segments of interest.

Both of these data collection problems are addressed by stratified multistage cluster sampling strategies. Such strategies usually involve the oversampling (i.e., sampling certain elements with a higher probability of selection than is the case for others in the sample) of those individuals with certain characteristics that need to be included in sufficient numbers for purposes of analysis. A stratified multistage cluster sample is achieved by first stratifying the population at higher levels. For example, since no comprehensive list of 4-year college students exists, we might compile a list of 4-year institutions that the students we are interested in attend; such lists are readily available. From this list we can then draw a sample of institutions within each of the strata we created (e.g., public; private). From the institutions sampled, we then request lists of students meeting the criteria related to our research interests. The lists of students from each institution can then be stratified into subgroups from which we could draw samples. In this example we have used a two-stage (institutions and students) stratified cluster sample. We could, of course, at either stage oversample within any particular strata. For example, we might wish to ensure that we had a sufficient number of Historically Black Colleges and Universities (HBCUs) in our sample from which we could make generalizations to the population of HBCUs.

Similarly, we might also want to ensure that our final sample had a sufficient number of students from various racial or ethnic backgrounds. Both objectives can be achieved by stratifying the sampling frame at the appropriate level and then choosing a set number or proportion of schools or students in each stratum (HBCU/non HBCU at the institution stage and minority/nonminority at the student stage). Most NCES survey samples are stratified on many different variables at each level.

While such complex sampling strategies are effective in getting the right numbers of the right types of observations in a sample, they also yield a sample that in its raw form is usually a severe distortion of the population from which it was drawn (i.e., in the example above, a disproportionate number of HBCUs and racial minorities). Hence, providing more weight to these types of institutions and students than is present in the overall population will bias any subsequent results in known and perhaps unknown directions, depending on the type of analysis and outcome of interest.

A second artifact of complex sampling strategies results from the clustered nature of the lower stages (students in this example). If the clusters of students are internally homogeneous—that is, if students within colleges are more similar than students across colleges—then the estimates of overall variance on measures will be lower than would be the case if a simple random sampling strategy were used (Muthen and Satorra, 1995). Simple random sampling requires the researcher to assume that all observations are independent (i.e., that individuals within similar subunits and institutions share no common characteristics or perceptions). Consequently, as similarities among individuals within groups become more pronounced in the sample, estimates of variances and standard errors derived from such data become more biased (Muthen and Satorra, 1995). This internal homogeneity of clusters is measured by calculating an intracluster correlation (ICC) coefficient (this calculation is discussed in a subsequent section). The degree of bias in estimating variances in data collected through cluster samples is a function of the ICC present in the data—the greater the ICC, the larger the resulting bias (Hox, 1998; Muthen and Satorra, 1995).

Commonly used statistical packages such as SAS and SPSS treat any data set as though it were constructed through a simple random sample (i.e., single-stage with equal probability of selection), thus ignoring the complexities associated with data collected through multistage cluster samples. Therein lies the problem we address in this article. We emphasize that researchers should be mindful of the structure of data assembled through complex sampling techniques (e.g., the presence of oversampling and the degree of homogeneity within clusters) and consider appropriate corrective strategies in terms of the analysis of these data. There are two classes of approaches that can be used to address analytical issues associated with the use of data from complex samples: *designed-based* approaches and *model-based* approaches (de Leeuw and Kreft, 1995; Muthen and Satorra, 1995).

In a design-based approach, the approach we present in this article, a single-level analysis can be maintained after adjustments are made for sample design effects including unequal subject selection probabilities and nonindependence of observations resulting from clustered designs (Muthen and Satorra, 1995). When using data from complex samples, the equal weighting of observations, which is appropriate with data collected through simple random samples, will bias the model's parameter estimates if there are certain subpopulations that have been oversampled. Moreover, the analytic methods appropriate for data collected from simple random samples ignore the similarities among individuals in the same institution (i.e., clustering effects). As we will show in this article, adjustments for these problems can be readily made. The analyst using such a design-based approach is conceptually constrained to modeling at a single level of analysis (i.e., students *or* institutions, but not students *and* institutions). This single-level analytic approach is consistent with the majority of techniques included in SAS or SPSS (e.g., simple descriptive analyses, multiple regression, discriminant analysis).

In contrast, model-based approaches (i.e., multilevel regression) directly incorporate the clustered sample design into the analytical models (Muthen and Satorra, 1995). In this approach, for each individual, the total score on a dependent variable is decomposed into an individual, or within-group, component and a between-group component. The decomposition of variables from the sample data into their component parts can be used to compute a within-groups covariance matrix (i.e., the covariance matrix of the individual deviations from the group means) and a between-groups covariance matrix (i.e., the covariance matrix of the disaggregated group means). The variation at each level can then be explained simultaneously with sets of predictors at each level of the data hierarchy (see Bryk and Raudenbush, 1992; Muthen, 1994; Muthen and Satorra, 1995; and Heck and Thomas, 2000, for further discussion of this approach).

By definition, multilevel approaches take the clustered data structure into account when producing estimates and thus obviate the need for further action to deal with problems of variance estimation resulting from clustered data in complex samples. For example, at the individual level, this is accomplished by developing a pooled within-group covariance matrix instead of a conventional covariance matrix based on the total number of individuals in the sample. This equation corresponds to the conventional equation for the covariance matrix of individual deviation scores, except the number of individuals in the sample minus the number of groups $(n - g)$ is used in the denominator of the equation instead of the usual $n - 1$ (see Muthen and Satorra, 1995, for further discussion). This adjustment provides correct degrees of freedom for the individual-level analysis where the assumption of independent observations is not met because of clustering. Such model-based approaches, while effectively dealing with clustered data, still require statistical adjustments for oversampling, however. The multilevel approach is the correct one in cases where the theoretical model calls

for data from more than one level of analysis to be analyzed (e.g., combining data from students and colleges in the same model). These approaches, however, also require specialty software packages to conduct the multilevel analysis (e.g., HLM, MLWin, and MPlus).

DESIGNED-BASED REMEDIES

Weighting for Oversampling

In situations where the researcher wishes to maintain a single-level analysis, most standard software packages can be manipulated though the use of sample weights to adjust for oversampling. Fortunately, most data sets from nationally representative samples also include a weight or a set of weights to adjust for unequal probabilities of selection in the sample design. Consider the B&B:93/94 sampling strategy, which relied on a multistage cluster sample of colleges and students with stratified samples and differential probabilities of selection at each level (NCES, 1995).[1] Institutions were first selected within geographic strata and were then further stratified by control (e.g., public, private, not-for-profit, etc.), and degree offering (e.g., 4-year nondoctorate granting, 4-year doctorate granting, etc.). Consequently, any estimates based on the raw unweighted sample will be biased in the favor of students graduating from schools that were oversampled within particular strata.[2]

If we used these data to learn something about the education-related indebtedness of baccalaureate recipients, our newfound empirical knowledge would be skewed by the disproportionate representation of graduates from particular types of institutions in particular regions. Consider the descriptive results of such an analysis that appear in Table 1. The estimates in this table are based on an unweighted subsample of graduates in the B&B:93/94 sample that reported holding education-related debt upon graduation. Because we have not weighted the sample to account for oversampling (i.e., the unequal probability of selection) within and across strata, this subsample is not representative of the target population—that is, the estimates are incorrect.

To make these data representative of the target population, we need to apply sample weights to deemphasize the disproportionate contribution of those elements that were oversampled. Two types of sample weights are commonly used in the analysis of survey data: raw (expansion) weights and relative weights. In its most basic form, the raw weight is computed as the reciprocal of an observation's probability of selection. Observations selected with a higher probability (i.e., oversampled) will have a smaller raw weight value. Summing the raw weight across all observations yields the population N:

$$\sum_{j=1}^{n} w_j = N$$

TABLE 1. Unweighted Estimates Based on SRS Assumption

Variable	*N*	Mean	SE	95% CI Lower	95% CI Upper	Σ Weights (effective *N*)	Variable Description
DEBT1000	4285	10.286	0.151	9.989	10.582	4285	Total educational debt/1000($)
FEMALE	4285	0.562	0.008	0.547	0.576	4285	Female dummy variable
NONWHITE	4285	0.148	0.005	0.138	0.159	4285	Nonwhite dummy variable
BATIME	4285	79.350	0.918	77.550	81.151	4285	Months to complete BA degree
TOTCOST	4285	123.256	1.085	121.129	125.383	4285	Total annual cost of attendance
FTFY	4285	0.503	0.008	0.487	0.518	4285	Full-time/full-year attendance dummy variable

While statistical packages vary in the way they use weights to calculate certain statistics, most calculate the weighted mean as follows:

$$\bar{x} = \sum_{j=1}^{n} w_j x_j \,/\, \sum w_j$$

or as the sum of the products of each observation's raw weight and value for x, divided by the sum of the raw weight, w. Notice that the sum of the raw weight ($\Sigma\, w_j$), or the size of the target population, now becomes the effective sample size in this calculation. This is an important point that will be considered subsequently. Weights are easily applied in SPSS using the WEIGHT BY command or in SAS using the WEIGHT subcommand.[3]

As with most large-scale government-related surveys, the B&B:93/94 survey includes a set of raw weights for the analyst to choose from. The weights accompanying most data sets of this type have been adjusted to account also for nonresponse and therefore are considerably more refined than the simple reciprocal of the probability of selection (details of such poststratification refinements are found in accompanying methodological reports). Choice of the proper raw weight depends on the purpose of the analysis (e.g., longitudinal vs. cross-sectional). In the case of the B&B:93/94 example used here, we are interested in estimating debt upon graduation—a cross-sectional analysis. The methodology reports that usually accompany these surveys will generally spell out the specific purpose of each of the weights included in the data file.

Now consider the weighted estimates in Table 2. These estimates have been weighted using the raw cross-sectional weight provided by NCES, BNBWT1. The columns headed "% +/–" represent the percentage increase or decrease (i.e., bias) of each estimate relative to the unweighted estimates shown in Table 1.

In addition to noting the change in the point estimates of the means and SEs, attention should also be paid to the sum of the weights. As explained above, this value will be equal to the size of the target population. These weighted

TABLE 2. Raw Weighted Estimates Based on SRS Assumption

Variable	N	Mean	Mean % +/–	SE	SE % +/–	95% CI Lower	95% CI Upper	Σ Weights (effective *N*)
DEBT1000	4285	10.083	–1.97	0.014	–90.50	10.055	10.111	474718
FEMALE	4285	0.540	–3.92	0.001	–90.50	0.538	0.541	474718
NONWHITE	4285	0.154	+3.64	0.001	–90.50	0.153	0.155	474718
BATIME	4285	82.546	+4.03	0.091	–90.50	82.368	82.724	474718
TOTCOST	4285	120.098	–2.63	0.102	–90.50	119.898	120.299	474718
FTFY	4285	0.472	–6.14	0.001	–90.50	0.471	0.473	474718

point estimates of the means are presumed to be correct for the parameters of interest. Notice, however, that the estimated standard errors reported in Table 2 are dramatically smaller (90.50 percent smaller) than those reported in Table 1. Some statistical packages, such as SPSS, calculate this value using the sum of the weights ($N = 474{,}718$) as the effective sample size.[4] In contrast, SAS version 8 uses the size of the actual sample n in the calculation of the standard error, regardless of the weight applied, and is therefore not sensitive to this problem.[5]

A consequence of using the raw weights supplied with most complex survey data is that, when calculating SE estimates, many statistical packages (SPSS included) are fooled into believing that the sample size is much larger than it really is. While both the raw and relative weights yield the same point estimates for the mean in all software packages, in some packages analyses using the raw weight result in an effective sample size that is the same as the population N. This can seriously compromise calculations that are sample size specific, such as variances and covariances, and leads to incorrect results. The effects of this become an especially critical point when one wishes to test hypotheses using weighted data—most every difference or coefficient becomes significant as a result when using statistical packages that are blind to the actual sample size.

This difficulty can be avoided in any statistical package, however, with a simple correction to the raw weight. In order to preserve the effective sample size while still adjusting for oversampling, we create a *relative* weight by dividing the raw weight by its mean,

$$w_i / \bar{w}$$

where $\bar{w} = \Sigma\, w_i / n$. Consider the new values in Table 3 obtained using the relative weight.

By using the relative weight, the estimates of the means in Table 3 have been corrected for oversampling in the design and can be considered correct (i.e., the same as those found in Table 2). Similarly, the relative weighted SE estimates

TABLE 3. Relative Weighted Estimates Based on SRS Assumption

Variable	*N*	Mean	SE	95% CI Lower	95% CI Upper	Σ Weights (effective *N*)
DEBT1000	4285	10.083	0.150	9.789	10.377	4285
FEMALE	4285	0.540	0.008	0.525	0.554	4285
NONWHITE	4285	0.154	0.006	0.143	0.165	4285
BATIME	4285	82.546	0.956	80.672	84.419	4285
TOTCOST	4285	120.098	1.076	117.989	122.208	4285
FTFY	4285	0.472	0.008	0.457	0.487	4285

in Table 3 are now correct, *assuming that a simple random sample was used to collect the data*. Put another way, in the case where a simple random sample was used, hypothesis tests using these relative weighted values would yield accurate results. We know, however, that this assumption is not valid when using data from the B&B:93/94 survey.

DESIGN EFFECTS

To this point in this article, we have focused on issues related to oversampling (i.e., the unequal probability of selection). We know, however, that a multistage cluster sample was used in the sample design of B&B:93/94 and that, as argued previously, there may exist homogeneity within the clusters (colleges) that would lead to underestimated SE values reported in Table 3. We now turn our attention to the impact of clustering in complex samples. The discussion throughout the remainder of the article assumes that problems outlined in the previous section relating to oversampling have been addressed by applying a relative weight in the analysis.

As a rule of thumb, the more similar are observations within their respective clusters the greater will be the underestimation of the true variability in the population (Hox, 1998). To examine the degree to which there exists intracluster homogeneity, we can examine the variance components of our outcome variable. The relative homogeneity of the clusters can be determined by partitioning the variance in the outcome measure into its within-cluster and between-cluster components. The partitioning is accomplished using the equivalent of a one-way ANOVA with random effects where the sample cluster variable (or primary sampling unit—institutions), SCHLID, is treated as a random factor with 583 levels (i.e., the number of institutions in the first-stage sample). From these components we can calculate an ICC coefficient by:

$$Var(\textit{between clusters}) \,/\, Var(\text{between clusters} + \textit{within clusters}).$$

The ICC should be zero when the data are independent; thus, its magnitude depends on characteristics of the variable measured and the attributes of the groups. In the presence of ICC, the impact of cluster sampling on the operating alpha level can be substantial. However, in the absence of substantial ICC (e.g., where the ICC is somewhat less than .05), there is little need to adjust for the design effect associated with this clustering. In such cases where the observations are nearly independent, traditional multiple regression analysis using appropriately weighted data will provide accurate estimates of the parameters and standard errors.

Partitioned variance coefficients can be obtained for the relative weighted data in SAS using the PROC MIXED routine or in SPSS using the VARCOMP

routine.[6] Each package offers several methods for estimating these components. Efficient estimation of the components requires that random errors are independent, normally distributed, and have constant variance (Bryk and Raudenbush, 1992). These assumptions are unrealistic, however, given that we are using a random coefficients model to estimate the intercepts (means) across the sample of first-stage units (colleges); such a model requires a more complex error structure. For this reason, we recommend using the maximum likelihood (i.e., full information maximum likelihood, restricted maximum likelihood) method of estimating these components (see Bryk and Raudenbush [1992] or Dempster, Laird, and Rubin [1977] for a complete consideration of these issues). Using the relative weighted data and restricted maximum-likelihood estimation (which uses the appropriate degrees of freedom and yields more precise estimates of the level-2 variance components when the number of units in the study may be small), SPSS provides the estimates shown in Table 4.

From these variance components, the ICC can be calculated as 18.669 / (18.669 + 80.783) = .188. The ICC of .188 (substantially above .05) requires the analyst to further consider the sample design effect.[7]

The choice of how to deal with effects of the internal homogeneity of the clusters depends, in part, on the aims of the research. A research question that involves modeling relationships from the level of the cluster as well as from the individual requires a model-based or multilevel approach that disaggregates individuals' scores within their specific clusters. In contrast to disaggregated approaches that rely on multilevel modeling techniques, as we have suggested, aggregated, design-based approaches restrict the analyst to focusing on one level of analysis to estimate a best overall model. Aggregated approaches to model estimation treat the sample as though it were a single group and then, by a variety of alternative procedures, adjust variances to account for homogeneity within clusters. Note, for example, that the regression routines found in standard statistical packages such as SPSS and SAS are aggregate approaches that do not, on their own, account for intracluster homogeneity. Again, routines such as these assume the data being analyzed were collected through a simple random sample and neither require, nor allow, further information about stratification or clustering in the sample design.

As we have argued, applying such aggregate simple random sample ap-

TABLE 4. Estimated Variance Components

Component	Estimate
Var(SCHLID)	18.669
Var(Error)	80.783

Dependent Variable: DEBT1000 Method: Restricted Maximum Likelihood Estimation.

proaches to data collected through complex sample designs biases estimated variances to the degree that there exists homogeneity among the clusters in the sample (Muthen and Satorra, 1995). If the analyst chooses to employ an aggregate analytical strategy using data collected through a complex sample design, then she or he must find a way to gauge and correct for potential bias in the estimates.

There are a number of ways to get correct variance estimates in aggregate (single-level) analyses. The most frequently used are Taylor expansion (linearization),[8] balanced repeated replication (BRR), and jack-knifing and bootstrapping techniques (e.g., see Wolter, 1985, and Rust, 1985). These techniques are used in specialty software packages that have been specifically designed to produce standard error estimates for data from complex samples (e.g., SUDAAN, WesVarPC, PCCARP). However, given the specialized nature of such packages, the financial cost, difficulty of use, or some combination of these factors, many analysts have been discouraged from incorporating them into their research efforts.

Recent software developments at both SAS and SPSS, however, promise to put more accessible complex sample tools within the reach of mainstream higher education researchers.[9] SAS version 8 includes two procedures and associated documentation in this area: PROC SURVEYMEANS and PROC SURVEYREG. Both procedures allow the analyst to adjust standard error estimates for sample design effects. The SAS procedures employ the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Fuller, 1975; Woodruff, 1971). With knowledge of and access to the sampling stratum variable and the cluster identification variable, which are usually clearly identified in the user's manuals for the data set, the analyst can employ either of these procedures to produce correct standard error estimates. Again, most secondary data sets from government agencies contain these design variables.

Table 5 contains the results of a relative weighted analysis in which a 16-category stratification variable (BNBSTRAT) and 583 category cluster variable (SCHLID) were designated (both of these variables can be found on the data file provided by NCES). Notice that the point estimates of the mean are consistent with the previous weighted estimates and that the standard errors are significantly larger than those reported in Tables 2 and 3. The inflation of the SE estimates adjusts out the bias resulting from intracluster homogeneity.

The ratio of this larger complex sample SE (squared) to the original relative weighted simple random sample SE (squared) defines what is known as the design effect or DEFF (Kish, 1965):

TABLE 5. Relative Weighted Estimates Based on Complex Sample Assumption*

Variable	*N*	Mean	SE	Σ Weights	DEFF
DEBT1000	4285	10.083	0.209	4285	1.94
FEMALE	4285	0.540	0.011	4285	2.08
NONWHITE	4285	0.154	0.011	4285	4.15
BATIME	4285	82.546	1.842	4285	3.72
TOTCOST	4285	120.098	1.932	4285	3.22
FTFY	4285	0.472	0.012	4285	2.52

*SEs adjusted for 16 strata and 583 clusters.

$$DEFF = \frac{SE^2_{COMPLEX\ SAMPLE}}{SE^2_{SRS}}$$

Both the DEFF and its square root or DEFT (also called the root design effect) are useful for adjustments that can be made either prior to hypothesis testing or after traditional hypothesis tests have been conducted. Most methodology reports provide design effect values for key variables in the data set as well as an overall mean value.

Once the design effect has been determined, ideally in terms of the outcome variable, it is possible to adjust future analyses by this value to compensate for underestimation of standard errors. Again, this is a critical adjustment in terms of hypothesis testing. If standard errors are underestimated by not taking the complex sample design into account, there exists a greater likelihood of finding erroneously "significant" parameters in the model than the a priori established alpha value indicates. For example, Barcikowski (1981) showed that the alpha of a *t* test performed at alpha .05 is inflated to .11, with ICC = .20 and a cluster size of 10. With a common group size of 25 and an ICC of .10, the operating alpha level would be .29 for a test performed at the standard level of alpha = .05 (Hox, 1998). Obviously, not adjusting for the effects of clustering can produce misleading results of parameter significance.

Consider a standard OLS regression model with total debt regressed on a set of predictors. These results are presented in Table 6 and were calculated using the relative weight but ignoring the complex sampling design. This output is what would be produced by the standard regression routines using relative weighted data in SAS or SPSS. Note the size of the standard error and the level of significance of each of the parameters in the model. For example, there are three predictor variables that are significant at alpha < .05 and one that is significant at $p < .10$. From the analysis of variance components in the previous section, we know that the outcome, total debt, has a fair degree of intracluster

TABLE 6. Relative Weighted Regression Estimates Based on SRS Assumption: Y = DEBT1000

Variable	B	SE	T	Prob(T)
INTERCEPT	7.174***	0.390	18.39	<0.0001
FEMALE	–0.523*	0.294	–1.78	0.0751
NONWHITE	–0.096	0.405	–0.24	0.8127
BATIME2	–0.006**	0.002	–2.34	0.0195
TOTCOST	0.034***	0.002	13.78	<0.0001
FTFY	–0.766**	0.351	–2.18	0.0291

*$p < .10$, **$p < .05$, ***$p < .01$.

homogeneity (ICC = .188) and that this has most likely biased downward the standard errors in Table 6, heightening the potential of committing Type I errors. We should therefore take corrective action with this model.

CORRECTIVE STRATEGIES

Using sample weights corrects for oversampling but not for similarities among individuals in clusters. There are at least four corrective alternatives that can be considered to account for the effects of clustered samples (in order of precision): (1) estimate the model using special software/procedures that account for the sample design, (2) adjust the estimated standard errors in regression upward as a function of a known DEFT value, (3) manipulate the effective sample size by adjusting the relative weight downward as a function of a known DEFF value, or (4) leave everything as is but evaluate each parameter in terms of a more conservative critical alpha value (e.g., .01 or .001 instead of .05). It should be stressed that alternative 1, using special software procedures, is by far the most accurate and preferable.

Alternative 1: Use Specialized Software Packages or Routines to Analyze the Data

Many specialty software packages (e.g., SUDAAN, WesVar, PCCARP) have a regression function that allows the specification of linear models such as the one reported in Table 6. These packages are often either expensive or very difficult to use properly. As mentioned previously, mainstream statistical packages are now beginning to incorporate routines for analyzing complex survey data. For example, the aforementioned SAS procedure, PROC SURVEYREG, enables the development and testing of regression models under complex sampling assumptions.[10] Specifying the same stratification and cluster variables used

to generate the results reported in Table 5, we used SAS SURVEYREG to reproduce the regression model in Table 6. Note, in Table 7, that the slope parameter estimates remain the same, while the standard error estimates are considerably larger.

The larger standard error estimates translate into smaller *t* ratios and result in substantive changes in our interpretation of the model's effects. Ignoring the complex sample design in the previous analysis (Table 6) led to the rejection of the hypotheses (at $\alpha = .05$) that $\beta_{FTFY} = 0$ and (at $\alpha = .10$) that $\beta_{FEMALE} = 0$. Given knowledge of the design effects (Table 7), we can see that such action based on the results assuming a SRS reported in Table 6 would have led us to commit at least one and possibly two Type I errors (i.e., false rejection of the null hypothesis), depending on our choice of alpha. It is worth noting that, when treated as a single-level model, these results are very close to those produced by multilevel modeling packages such as HLM. Deviations between the two approaches at this stage will most often result from the ability of most multilevel software programs to adjust estimates not only for clustering but also for within-unit reliabilities (see Heck and Thomas [2000] for a complete discussion of multilevel estimates).

Alternative 2: Adjust Estimated Standard Errors by a Known DEFT Value

While the previous approach is the most appropriate for the analysis of data from complex samples, there exist a number of alternatives to approximate and adjust for resulting biases. If the analyst does not have the luxury of estimating the model directly in a complex sample environment, she or he can also adjust the standard errors estimated under the simple random sample assumption (Table 6). The standard errors in that model can be multiplied by the root mean

TABLE 7. Relative Weighted Regression Estimates Based on Complex Sample Assumption: Y = DEBT1000

Variable	B	SE	T	Prob(T)	DEFF
INTERCEPT	7.174***	0.468	15.34	<0.0001	1.44
FEMALE	−0.523	0.357	−1.47	0.1434	1.48
NONWHITE	−0.096	0.521	−0.18	0.8540	1.66
BATIME	−0.006**	0.003	−2.08	0.0383	1.27
TOTCOST	0.034***	0.004	9.25	<0.0001	2.22
FTFY	−0.766	0.482	−1.59	0.1122	1.88

*$p < .10$, **$p < .05$, ***$p < .01$.

design effect or DEFT (i.e., which we can calculate as the square root of the DEFF value for the outcome variable in Table 3).

$$DEFT = \frac{SE_{COMPLEX\ SAMPLE}}{SE_{SRS}} = \sqrt{DEFF}$$

These DEFT adjusted standard errors can then be used to calculate new *t* ratios. The trick in the absence of specialized software is getting an estimate of the design effect itself. Fortunately, for most NCES surveys one can either: (1) use NCES's widely available public use Data Analysis System (DAS) to estimate the design-adjusted standard error of any variable, or (2) resort to gross effects found in tables at the end of each methodology report.[11] However obtained, the DEFF value can then be converted to a DEFT value by which we can adjust each standard error in Table 6 and reevaluate the hypotheses.

Table 8 shows the results of such an adjustment using a DEFF value of 2.26. To illustrate the preceding point about locating the appropriate DEFF value, we took this from the B&B:93/94 methodology report (the corresponding DEFT value is 1.48; NCES, 1996, p. 49). It is worth noting that these design effect values reported by NCES are larger than those we reported in Table 3. There are two reasons for these differences: (1) the NCES values are means across the variables they consider in their report, and (2) the NCES values are computed on the full sample and not the specific subsample we are using. Thus the analyst should be aware of such potential differences.

Due in part to the larger DEFF value used, this approach yields more conservative SE estimates in our example than those provided by SAS PROC SURVEYREG. This technique, however, tends to yield more conservative estimates even when using comparable DEFF values. Given that this is generally the case, the analyst using this approach could consider evaluating each parameter at a more liberal alpha level (e.g., .05 rather than .01).

TABLE 8. DEFT Adjusted, Relative Weighted SRS Regression Estimates: Outcome = DEBT1000

Variable	B	SE	T	Prob(T)
INTERCEPT	7.174***	0.573	12.51	<0.0001
FEMALE	–0.523	0.434	–1.20	0.2400
NONWHITE	–0.096	0.599	–0.16	0.8750
BATIME	–0.006	0.004	–1.56	0.1209
TOTCOST	0.034***	0.004	9.33	<0.0001
FTFY	–0.766	0.520	–1.47	0.1201

*$p < .10$, **$p < .05$, ***$p < .01$.

Alternative 3: Adjust the Relative Weight to Alter the Effective Sample Size

Another alternative for computing more accurate standard errors is to alter the effective sample size by adjusting the relative weight downward as a function of the overall design effect. In some software packages, however, one cannot apply this method (SAS version 8 for example, as mentioned in a previous section, always treats the effective sample size as *n* regardless of how the sum of the sample weight is manipulated). In packages such as SAS that do not allow this approach, the analyst must rely on one of the other techniques for adjusting for complex sample design effects. Assuming that the software package being used does pay attention to the sum of the weight, as is the case with SPSS, adjustments using this approach are made by multiplying the relative weight by the reciprocal of the DEFF value and then reweighting the data with this DEFF adjusted relative weight.

$$\frac{1}{DEFF} * NORMWT$$

In the absence of substantially different variances across groups, the results of this approach (summarized in Table 9) will be roughly equivalent to those found by adjusting the individual parameters using the DEFT value in the previous example (see Table 8).

Alternative 4: Alter the Evaluation Criteria (alpha)

Finally, in the absence of an approximate DEFF value the analyst is still obliged to acknowledge the potential bias associated with estimates produced under SRS assumptions. This acknowledgment should be informed by the ICC

TABLE 9. DEFF Adjusted Weighted Regression Estimates Assuming SRS: Y = DEBT1000

Variable	B	SE	T	Prob(T)
INTERCEPT	7.174***	.587	12.22	<0.0001
FEMALE	–0.523	.442	–1.18	0.2369
NONWHITE	–0.096	.609	–0.16	0.8749
BATIME	–0.006	.004	–1.55	0.1207
TOTCOST	0.034***	.004	9.16	<0.0001
FTFY	–0.766	.528	–1.45	0.1471

*$p < .10$, **$p < .05$, ***$p < .01$.

of the outcome variable, which could be used as a guide in determining the potential bias.[12] The evaluation criterion can be adjusted according to the ICC—where higher ICCs should lead to lower alpha values. Unfortunately, there exists little empirical work assessing this relationship with large numbers of groups of unequal size to provide a firm framework in which to consider such adjustments.

CONCLUSION

We have shown that special consideration needs to be given to sampling issues when analyzing data collected through complex sample designs. These sample designs are frequently used in large-scale data collection efforts aimed at interests in postsecondary education. Two issues have been considered in this article: weighting for unequal subject representation and standard error corrections for intracluster correlations. From our analysis it is clear that failure to use sample weights will always (in the absence of simple random sampling), to some degree, lead to incorrect estimates of population parameters. Thus, a first conclusion is that sample weights should always be used when analyzing data from complex samples. Of the two weights considered in this article, the relative weight (NORMWT) is the least problematic. While both the raw weight (RAWWT) and the relative weight (NORMWT) produce the same point estimates in descriptive analyses, some software packages are fooled into thinking that the sum of the raw weight is the effective sample size. In packages that succumb to this problem, this results in a gross understatement of the standard error and leads to erroneous decisions when evaluating hypotheses. We therefore recommend use of a relative weight for most analyses.

After appropriately weighting complex survey data, one may also need to make adjustments to standard errors due to the clustered nature of the sample. We recommend the use of complex sample survey software for estimating population parameters and standard errors with this type of data. While standard software packages can be used to provide approximately equivalent results as those obtained from software especially designed for analyzing data from complex samples, such approaches are only approximate and fail to capitalize on the actual characteristics of the data set. Our analysis demonstrates that not taking the sample design into account results in a potentially substantial overstatement of the effects of parameters in predictive models. Thus, a second conclusion is that ignoring complex sample designs leads to overly conservative estimates of standard errors and the heightened potential for committing Type I errors in hypothesis testing.

Finally, while we have purposefully dealt only with single-level formulations (i.e., individual-level models), we stress that the choice of strategies for dealing with complex sample data rests largely on the conceptualization of the study. Cases in which one seeks to use data from the level of the cluster as well as the

individual—for example, where one wants to understand the effects of organizational characteristics of colleges on students—usually require a multilevel statistical treatment. While the weighting issues considered in this article apply to multilevel models as well, such techniques deal with the effects of clustered samples in different ways, both theoretically and practically.

Acknowledgments. We are indebted to Samuel Peng at the National Center for Education Statistics and George Marcoulides for their helpful comments on an earlier draft. Any errors and omissions are our own.

APPENDIX A. EXAMPLES OF SAS CODE USED

New variables are most easily computed in the initial DATA step in SAS. To compute the relative weight (NORMWT) one first needs to find the mean value of the raw weight (BNBWT1) for the sample being used.

```
PROC  MEANS MEAN STDDEV STDERR N;
  VAR BNBWT1;
RUN;
```

This yields a mean value of 110.7860. The relative weight is calculated in a DATA step by dividing the raw weight by its mean:

```
DATA  DEBT;
  SET SASUSER.DEBTONLY;
  NORMWT=BNBWT1/110.7860;
RUN;
```

Raw weighted descriptive analyses and relative weighted parameter estimates in models can be generated using the WEIGHT subcommand in SAS. For raw weighted analyses the user would use BNBWT1:

```
PROC  MEANS MEAN STDDEV STDERR N;
  VAR DEBT1000 FEMALE NONWHITE BATIME2 TOTCOST FTFY;
  WEIGHT BNBWT1;
RUN;
```

For relative weighted analyses, the user would use the newly created NORMWT variable (see above):

```
PROC  MEANS MEAN STDDEV STDERR N;
  VAR DEBT1000 FEMALE NONWHITE BATIME2 TOTCOST FTFY;
  WEIGHT NORMWT;
RUN;
```

To account for the complex sample design, the user needs to recompute these estimates using the PROC SURVEYMEANS routine. This requires knowledge of any strata and clusters that exist in the data. The Baccalaureate & Beyond data set contains a variable for each of these dimensions: BNBSTRAT for the strata and SCHLID for the clusters. Again, the weight is applied using the WEIGHT subcommand:

```
PROC  SURVEYMEANS MEAN STD STDERR SUMWGT;
  CLUSTER SCHLID;
  VAR DEBT1000 FEMALE NONWHITE BATIME2 TOTCOST FTFY;
  STRATA BNBSTRAT;
  WEIGHT NORMWT;
RUN;
```

The PROC SURVEYREG routine in SAS uses the Taylor Linearization method to provide the correct standard error estimates for regression analyses. The / DEFF option on the MODEL statement provides design effect values for each of the parameters. Again, the weight is applied using the WEIGHT subcommand:

```
PROC  SURVEYREG;
  CLUSTER SCHLID;
  MODEL DEBT1000= FEMALE NONWHITE BATIME2 TOTCOST FTFY /
    DEFF;
  STRATA BNBSTRAT ;
  WEIGHT NORMWT;
RUN;
```

SAS's PROC MIXED routine allows the user to partition the variance in the outcome variable into its within- and between-unit components. The intracluster correlation is calculated using these values. The METHOD = ML option on the PROC MIXED statement specifies maximum likelihood estimation and the WEIGHT subcommand applies the relative weight NORMWT:

```
PROC  MIXED METHOD=ML;
  CLASS SCHLID;
  MODEL DEBT1000=;
  RANDOM SCHLID ;
  WEIGHT NORMWT;
RUN;
```

APPENDIX B. EXAMPLES OF SPSS CODE USED

To compute the relative weight (NORMWT) one first needs to find the mean value of the raw weight (BNBWT1) for the sample being used. The user needs to make sure that no weight is being used when calculating this mean:

```
WEIGHT OFF.
DESCRIBE BNBWT1.
```

This yields a mean value of 110.7860. The relative weight is calculated by dividing the raw weight by its mean:

```
WEIGHT OFF.
COMPUTE NORMWT=BNBWT1/110.7860.
```

Raw weighted descriptive analyses and relative weighted parameter estimates in models can be generated using the WEIGHT BY statement in SPSS. For raw weighted analyses the user would use BNBWT1:

```
WEIGHT BY BNBWT1.
DESCRIBE DEBT1000 FEMALE NONWHITE BATIME TOTCOST FTFY.
```

For relative weighted analyses, the user would use the newly created NORMWT variable (see above):

```
WEIGHT BY NORMWT.
DESCRIBE DEBT1000 FEMALE NONWHITE BATIME TOTCOST FTFY.
```

Caution: Note that in SPSS the user needs to invoke the WEIGHT BY statement only once. All subsequent analyses will be conducted applying the specified weight until the WEIGHT OFF statement is run.

To compute the design effect adjusted weight (DEFFWT), the user needs to divide the relative weight (NORMWT) by a known value for the survey design effect (again, this value can be obtained through the methodology report accompanying most complex survey data sets or, for most NCES data sets can be generated using the Data Analysis System software distributed by NCES (see footnote X in text).

```
WEIGHT OFF.
COMPUTE DEFFWT=NORMWT/1.96.
```

The design effect adjusted weight can then be applied in the same fashion as the other weights:

```
WEIGHT BY DEFFWT.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT totdebt
  /METHOD=ENTER FEMALE NONWHITE BATIME2 TOTCOSTA FTFY.
```

The VARCOMP procedure in SPSS allows the user to partition the variance in the outcome variable into its within- and between-unit components. The intracluster correlation is calculated using these values. The METHOD = ML option specifies maximum likelihood estimation:

```
WEIGHT BY NORMWT.
VARCOMP
  DEBT1000 BY SCHLID
  /RANDOM = SCHLID
  /METHOD = ML.
```

NOTES

1. The Baccalaureate & Beyond Study tracks the experiences of a cohort of college graduates who received a bachelor's degree during the 1992–1993 academic year. The B&B:93/94 was drawn from the much larger 1993 National Postsecondary Student Aid Study (NPSAS:93). The NPSAS:93 sample, while representative and statistically accurate, was not a simple random sample (NCES, 1995).
2. Conceptually, all analyses can be considered as weighted. An unweighted analysis is actually one in which all observations are weighted equally with a weight of 1.0.
3. In the interest of making these points as accessible as possible, we have included in Appendices A and B examples of the actual SAS and SPSS code used to produce the results reported throughout the article.
4. Recall that standard error decreases as a function of sample size.
5. While SAS version 8 computes the standard error using the correct sample size, basic variance estimates remain problematic and should be checked carefully before using.
6. The SAS VARCOMP procedure does not allow the use of weights. However, weighted variance components can be calculated in SAS using the PROC MIXED routine.
7. This exercise is largely demonstrative at this point, and, as will become clear in a subsequent section, is usually only conducted when there is little information known about the magnitude of potential bias resulting from the clustering of observations.
8. This method goes by several different names in the literature, including the linearization method, the delta method (Kalton, 1983), and the propagation of variance (Kish, 1965).
9. While SAS version 8 incorporates these new procedures into its base program, SPSS version 10 relies on an interface with a version of the WesVar program, thus still requiring purchase of this add-on and programming knowledge of an additional piece of software.

10. This SAS procedure is limited to standard regression models. Some of the more developed complex sample software packages such as SUDAAN and WesVar provide the researcher with a much wider variety of routines (e.g., multinomial logistic regression for ordinal and binary data, proportional hazards models, etc.).
11. While NCES's DAS software provides a convenient way to obtain design effect values, the reader is cautioned that these values are abnormally conservative. Those interested in this approach are advised to contact the NCES staff for further information.
12. This is an instance when knowledge of the ICC is instrumental. In such instances, the researcher can calculate the ICC in a fashion similar to that presented in an earlier section of this article.

REFERENCES

Barcikowski, R. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics* 6: 267–285.

Bryk, A. S., and Raudenbush, S. W., (1992). *Hierarchical Linear Models*. Newbury Park, CA: Sage Publications, Inc.

de Leeuw, J., and Kreft, I. G. (1995). Questioning multilevel models. *Journal of Educational Statistics* 20: 171–189.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 30: 1–38.

Dey, E. L., Hurtado, S., Rhee, B. S., Inkelas, K. K., Wimsatt, L. A., and Guan, F. (1997). *Improving Research on Postsecondary Outcomes*. Palo Alto, CA: National Center for Postsecondary Improvement, Stanford University.

Fuller. (1975). Regression analysis for sample survey. *Sankhyα* 37: 117–132.

Heck, R. H., and Thomas, S. L. (2000). *An Introduction to Multilevel Modeling Techniques*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

HLM [Computer software]. (1999). Chicago: Scientific Software International.

Hox, J. J. (1998). Multilevel modeling: when and why. In I. Balderjahn, R. Mathar, and M. Schader (eds.), *Classification, Data Analysis, and Data Highways*, pp. 147–154. New York: Springer Verlag.

Kalton, G. (1983). *Introduction to Survey Sampling*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-035. Beverly Hills: Sage Publications, Inc.

Kish, L. (1965). *Survey Sampling Principles*. New York: Marcel Dekker, Inc.

MLwinN [Computer software]. (1999). London: Institute of Education.

MPlus [Computer software]. (1999). Los Angeles: Muthén & Muthén.

Muthen, B. O., and Satorra, A. (1995). Complex sample data in structural equation modeling. In P. Marsden (ed.), *Sociological Methodology*, pp. 267–316. Washington, DC: American Sociological Association.

National Center for Education Statistics. (1995). *Methodology Report for the 1993 National Postsecondary Student Aid Study*. Washington, DC: Author.

National Center for Education Statistics. (1996). *Baccalaureate and Beyond Longitudinal Study: 1993/94 First Follow-up Methodology Report*. Washington, DC: Author.

PCCARP [Computer software]. (1989). Ames, IA: Statistical Laboratory, Iowa State University.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics* 4: 381–397.

SAS [Computer software]. (1999). Cary, NC: SAS Institute, Inc.

SPSS [Computer software]. (1999). Chicago: SPSS, Inc.

SUDAAN [Computer Software]. (1999). Research Triangle Park, NC: Research Triangle Institute.
WesVar Complex Samples [Computer Software]. (1998). Rockville, MD: Weststat.
Wolter, K. M. (1985). *An Introduction to Variance Estimation*. New York: Springer.
Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association* 66: 411–414.

Received February 25, 2000.