The Journal of Experimental Education, *2005*, *73*(3), 221–248

A Primer for Using and Understanding Weights With National Datasets

DEBBIE L. HAHS-VAUGHN
University of Central Florida

ABSTRACT. Using data from the National Study of Postsecondary Faculty and the Early Childhood Longitudinal Study—Kindergarten Class of 1998–99, the author provides guidelines for incorporating weights and design effects in single-level analysis using Windows-based SPSS and AM software. Examples of analyses that do and do not employ weights and design effects are also provided to illuminate the differential results of key parameter estimates and standard errors using varying degrees of using or not using the weighting and design effect continuum. The author gives recommendations on the most appropriate weighting options, with specific reference to employing a strategy to accommodate both oversampled groups and cluster sampling (i.e., using weights and design effects) that leads to the most accurate parameter estimates and the decreased potential of committing a Type I error. However, using a design effect adjusted weight in SPSS may produce underestimated standard errors when compared with accurate estimates produced by specialized software such as AM.

Key words: AM, design effects, design effects adjusted weights, normalized weights, raw weights, unadjusted weights, SPSS

LARGE COMPLEX DATASETS are available to researchers through federal agencies such as the National Center for Education Statistics (NCES) and the National Science Foundation (NSF), to name just a few. Although these datasets have become more readily accessible to researchers and statistical programs for analyzing the data have become more user-friendly and capable of handling such large datasets, there are two basic analytical issues that must be addressed by researchers who are interested in generalizing to the population for which the dataset is representative: the "superpopulation" compared with the "finite population," the population from which the sample was taken (Potthoff, Woodbury, &

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

Manton, 1992, p. 383). First, these samples are generally not collected using simple random sampling techniques. Most of these complex samples have been collected by cluster sampling, stratified sampling, or multistage sampling because a simple random sample is not feasible. There is no easily accessible list that exists, for example, of all children who attended kindergarten in 1998–1999 (i.e., the Early Childhood Longitudinal Study–Kindergarten Class of 1998–99, ECLS-K). However, what is available is a list of schools (the second-stage sampling unit) within counties (the primary sampling unit) that serve kindergarten students. From the schools, therefore, students can then be identified (the third and last stage of the sampling unit). Second, subsets of the population have been oversampled. This type of sampling design creates challenges that must be addressed when performing statistical analyses to ensure accurate standard errors and parameter estimates. The underlying assumptions of parametric statistical procedures may be violated if the complex sampling design is not considered in the analyses. However, research addressing methodological issues when dealing with complex samples is negligible (Kaplan & Ferguson, 1999), and research that uses extant data to study sample weights and design effects relating to methodological procedures is scarce (Hahs, 2003).

Survey weights and design effects are appropriate tools through which complex sampling designs can be accommodated. However, guidelines on how to incorporate weights and design effects effectively in common statistical programs are negligible, and guidelines on how to apply knowledge gained from simulation research on this topic to extant data under different weighting options are few (e.g., DuMouchel & Duncan, 1983; Hahs, 2003; Korn & Graubard, 1995; Thomas & Heck, 2001). Thomas and Heck provided guidelines for using weights and design effects in SAS and SPSS, including programming codes for each. However, instructions on applying weights and design effects using Windows-based SPSS are not available. A recent software program freely accessible online, AM, is an alternative for analyses of national samples. However, because it is relatively new, it has not received widespread review in the literature as an option.

My purpose in this article is to assist in filling the void that currently exists on understanding the intricacies of complex samples, specifically as it relates to the use of weights and design effects in Windows-based SPSS and AM. Using the National Study of Postsecondary Faculty (NSOPF:93) and the ECLS-K public microdata, I provide guidelines on how to appropriately incorporate weights and design effects in single-level analysis using Windows-based SPSS and AM. In addition, I provide examples of analyses that do and do not employ weights and design effects to illuminate the differential results of key parameters and standard errors using varying degrees of using or not using the weighting and design effect continuum (no weights, raw weights, relative or normalized weights, and design effect adjusted weights). Recommendations on the most appropriate weighting options are provided. In this study, I focus on researchers using data collected

with a complex sampling design, such as those available through NCES or NSF, and using single-level analyses that require an adjustment to the standard error to ensure accurate parameter estimates and standard errors.

Model- and Design-Based Approaches

Two approaches have been suggested as ways to deal with homogeneous clusters: (a) model-based and (b) design-based approaches (Kalton, 1983). Which approach is selected should be based on the research question. If a researcher is interested in the clustered relationships as well as from the individual, a model-based approach should be pursued. If a researcher is not interested in the clustering effect but wants to focus the analysis on one level, treating the sample as one group, then a design-based approach is appropriate (Thomas & Heck, 2001).

Model-Based Approach

In a model-based approach (Kalton, 1983), the statistical methodology directly incorporates the clustering in the analysis (Heck & Mahoe, 2004; Muthén & Satorra, 1995). The variance of the dependent variable score is partitioned into within- and between-variances, which are explained at each level by including predictor variables hierarchically (Heck & Mahoe). Model-based approaches treat the sample as one group and adjust variances to account for homogeneous clusters (Heck & Mahoe). This process is also known as a *disaggregated approach* because the procedures disaggregate scores from an individual into their respective cluster (Muthén & Satorra; Thomas & Heck, 2001). Examples of tools for model-based approaches include multilevel structural equation modeling and hierarchical linear modeling. By design, a model-based approach negates the need to deal with clustered and potentially homogeneous subsets (Thomas & Heck), although oversampling must still be addressed (e.g., through the use of weights).

Multilevel models are one way to account for multistage sampling, and these have received substantial attention (e.g., Hox & Kreft, 1994; Kaplan & Elliott, 1997; Muthén, 1994). Not all researchers, however, may be interested in multilevel modeling to account for the multistage sampling design (an example of multilevel modeling is reviewing student-level variables as a function of school-level variables, such as climate, policies, and resources; Kaplan & Elliott). Likewise, the available datasets may not have the appropriate institution-level variables for the specified model (Kaplan & Elliott). In those situations, a design-based approach is appropriate.

Design-Based Approach

A design-based approach (Kalton, 1983), also known as an *aggregated ap-*

proach, estimates a best overall model fit by focusing on only one level of the analysis (i.e., single-level analysis; Thomas & Heck, 2001). Common statistical procedures, such as analysis of variance and regression, are examples of design-based tools because they do not automatically accommodate homogeneous clusters. As stated previously, applying a design-based approach without correcting for the possible bias resulting from homogeneous clusters underestimates the true population variance (Hox, 1998). A number of strategies, ranging from using specialized software to using a more conservative alpha level, have been proposed when using design-based approaches that will generate accurate variance estimates (e.g., Peng, 2000; Thomas & Heck).

Design-Based Approach Strategies Using Weights and Design Effects

Weights

Unequal selection probability occurs when elements in the population are sampled at different rates (Stapleton, 2002). A weight, in its simplest form, is the inverse of the probability of selection (Kish, 1965). When the unit of analysis (i.e., an individual) is sampled with unequal probability of selection, the sample weight represents the number of units (i.e., individuals) in the population that each unit (i.e., individual) represents (Korn & Graubard, 1995). Incorporating weights in descriptive or inferential analyses is needed to compensate for the unequal probability of selection, nonresponse and noncoverage, and poststratification (Kalton, 1989) and is often the easiest way to deal with disproportionate sampling (Stapleton). Populations that are oversampled in national datasets have a smaller weight value (Thomas & Heck, 2001). Ignoring disproportionate sampling may result in biased parameter estimates and poor performance of test statistics and confidence intervals (Pfeffermann, 1993) as the weights are required to produce estimates that are representative of the intended population (U.S. Department of Education, 2002). Biased parameter estimates and poor performance have been demonstrated using simulation in single-level structural equation modeling (e.g., Kaplan & Ferguson, 1999) and using extant data in regression (e.g., Korn & Graubard).

Korn and Graubard (1995) provided an example of weighted compared with unweighted analyses. Using the 1988 National Maternal and Infant Health Survey, the authors presented four regression and logistic regression examples of weighted and unweighted estimates using SUDAAN software to estimate standard errors. They illustrated how weighted and unweighted models differ when a model is misspecified, when a covariate is omitted, and when interactions with a covariate are not included. In these examples, differences in weighted and unweighted estimators of association could be eliminated by changing the model. However, in the last example presented, it was shown that that strategy is not always possible. A cross-classification provided weighted and unweighted mean

birthweights by mothers' smoking status. The unweighted analysis overestimated the mean birthweight difference between mothers who did and mothers who did not smoke because low-birthweight babies were oversampled. As suggested by Korn and Graubard, the potential for bias that may result if weights are ignored needs to be balanced with the increased variability of estimators when designing a strategy for incorporating weights. The Korn and Graubard study points to a broader reflection in light of weighted versus unweighted samples in that the unweighted sample is only a "collection of individuals that represents no meaningful population" (Kalton, 1989, p. 583). The results of analyses from unweighted samples cannot be generalized to any population other than that which was included in the original sample (i.e., the finite population). In most cases, this defeats the purpose of using a dataset that is nationally representative of some underlying population.

Once researchers have decided to incorporate sample weights, what do they do next? The methodology reports that accompany the datasets provide intimate detail into the calculation of weights, the various weights that are contained within the dataset, and how to determine which weight is appropriate given the type of analyses planned. For example, the ECLS-K methodology report provides a table that lists the weight variable name and corresponding detail on the type of analyses for which that weight is appropriate. For example, weight C123CW0 is designed to be used for analysis of "child direct assessment data from fall- AND spring-kindergarten AND fall-first grade, alone or in conjunction with any combination of a limited set of child characteristics (e.g., age, sex, race-ethnicity)" (Tourangeau et al., 2002, p. 5). Once the appropriate weight variable is determined, it is left to the researcher to apply it effectively in the analyses to correct for oversampling in the design. How to effectively apply the weight can be interpreted in terms of understanding the differences between raw, relative or normalized, and design effect adjusted weights.

Raw weights. The weight provided in the dataset is a raw weight. The sum of the raw weights is the population size, *N* (West & Rathburn, 2004). Therefore, estimates derived from the application of raw weights to the data will be done based on the population size, *N,* rather than on the actual sample size, *n* (Kaplan & Ferguson, 1999). Any estimates that are sensitive to sample size (e.g., standard errors, test statistics), therefore, will be affected when using the raw weight (Kaplan & Ferguson). Statistical packages such as SPSS treat the sum of the weights as the actual sample size. Thus, if the raw weight is used, tests of inference most likely will be significant because the software is interpreting the population rather than the sample size (Thomas & Heck, 2001).

Relative or normalized weights. Some authors, including Longford (1995), Pothoff et al. (1992), and Thomas and Heck (2001), choose to label this weight *relative*; others refer to it as *normalized* (e.g., Kaplan & Ferguson, 1999). Be-

cause of its reference as normalized by NCES (West & Rathburn, 2004), I adopt *normalized* as the label in this article, recognizing that the terms *relative* and *normalized* are essentially interchangeable. Regardless of what the weight is labeled, it is calculated by dividing the raw weight by its mean, thereby preserving the sample size (Peng, 2000; Thomas & Heck). Normalized weights sum to the actual sample size, *n* (Kaplan & Ferguson; Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998). Normalized weights address sample size sensitivity issues while still incorporating sample weights (Kaplan & Ferguson). Applying the normalized weight in the analyses ensures that the standard error estimates are correct *given a simple random sample* (Thomas & Heck). Researchers should remember, however, that complex samples do not use simple random sample designs. Therefore, an additional step to include design effects in conjunction with the normalized weighted must be incorporated to recognize and adjust for potential dependence among observations.

Design effect adjusted weights. The assumption of independent and identically distributed observations is required for estimating accurate standard errors (Lee, Forthofer, & Lorimor, 1989). Complex sample data usually have some degree of dependence among the observations because of the multistage or clustered sample design (Stapleton, 2002). A comprehensive approach using weights and design effects in tandem compensates for dependence along with disproportionate sampling and is detailed as one of the strategies for using design effects.

Design Effects

Multistage sampling is the process of subsampling clusters so that the elements 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 (Pratt et al., 1996). It is possible that the observations within clusters are more alike in some ways compared with observations in other clusters (Hox & Kreft, 1994). Because of the similarities within clusters, the assumption of independence is negated (Kish & Frankel, 1974), and the true population variance will be underestimated (Hox, 1998; Selfa et al., 1997). Whether the assumption is mildly, moderately, or severely violated, the reported probability intervals will reflect that same level in its underestimation (i.e., mild, moderate, or severe underestimation; Kish & Frankel). The design effect measures the impact of departing from simple random sampling on sample estimate precision and is the ratio of the estimated variance of a statistic derived from considering the sample design to that derived from the formula for simple random samples (Selfa et al.).

As with the weights, the methodology reports that accompany the datasets provide lengthy tables of standard errors and design effects. For example, in the NSOPF:93 methodology report, average design effects (DEFF) and the average

of the square root of the design effects (DEFT) for total respondents and for 30 subgroups are presented for 30 randomly selected dichotomized items from the faculty and institution questionnaires (Selfa et al., 1997). For groups formed by subdividing those provided in the methodology reports (and when the variable used to subdivide cuts across institutions), design effects will generally be smaller because they are less affected by clustering than larger subgroups. Therefore, using the subgroup mean DEFT is a conservative approach to estimating standard errors. For comparisons between subgroups, if the subgroups cut across institutions, then the design effect for the difference between the subgroup means will be slightly smaller than the design effect for the individual means. The estimate of the variance of the difference will be less than the sum of the variances from which it is derived. Using DEFT is also conservative in calculating standard errors for complex estimators, such as correlation and regression coefficients compared with simple estimators (Kish & Frankel, 1974). Regression coefficients most often will have smaller design effects than comparisons of subgroups, and comparisons of subgroups often will have smaller design effects than means (Selfa et al.).

Various strategies have been proposed to accommodate for the homogeneous clusters in single-level analyses using design effects: (a) specialized software, (b) adjusted test statistic, and (c) normalized or relative weight adjusted by DEFF (Thomas & Heck, 2001). A final and last resort strategy when no known DEFF is available is to adjust the alpha level to a more conservative evaluation criterion (Thomas & Heck). A discussion of the three most desirable strategies follows, and examples using extant data are presented.

Strategy 1: Specialized software. Software packages such as WesVar, SUDAAN, and STATA are designed to accommodate complex sampling assumptions (e.g., Thomas & Heck, 2001; West & Rathburn, 2004) but are not widely used by mainstream researchers because of their cost or difficult use (Thomas & Heck). A relatively new and freely accessible software package, AM, is now available, although it is still in Beta testing (AM, n.d.). Made available by the American Institutes for Research, AM is designed to be user-friendly and to accommodate large-scale assessments. Using Taylor-series approximation, this software automatically provides appropriate standard errors for complex samples. Although the range of statistical tools in AM is not as broad as more popular packages such as SPSS, it does offer regression, probit, logit, and cross-tabs, among others (AM).

Strategy 2: Adjusted test statistic. Using DEFT or DEFF, test statistic values can be adjusted. In *t* tests, the test statistic should be divided by DEFT (West & Rathburn, 2004). In *F* tests, the test statistic should be divided by DEFF (West & Rathburn). This strategy for accommodating homogeneous clusters tends to be conservative, and thus a more liberal alpha level may be desirable (Thomas & Heck, 2001). Adjusting the test statistic by DEFF or DEFT and conducting the

analysis with a design effect adjusted weight produces approximately equivalent parameters if the variances across groups are fairly equivalent (Thomas & Heck).

Strategy 3: Normalized weight adjusted by DEFF. Another alternative is to adjust the weight so that the adjusted standard error is derived (West & Rathburn, 2004). In this strategy, the effective sample size is altered by adjusting the normalized weight downward as a function of the overall design effect (Peng, 2000; Thomas & Heck, 2001; West & Rathburn). The adjusted weight is calculated by dividing the normalized weight by DEFF. The analyses are then conducted by applying the new, adjusted weight.

Weights and Design Effects Applied in SPSS Using NSOPF:93

Converting a normalized weight (which corrects the sample for disproportionate sampling) or design effect adjusted weight (which corrects the sample for oversampling *and* multistage sampling; i.e., design effect Strategy 3) from a raw weight in SPSS is a relatively simple process. This example uses the NSOPF:93 public microdata. Given that this is the public-use file, the variables provided are those that were found to pose no potential threat of disclosure; thus, it is a limited dataset. The NSOPF:93 was designed to collect data that are nationally representative of instructional faculty and staff and noninstructional faculty at public or nonproprietary 2-year and above postsecondary institutions. The NSOPF:93 used a multistage sample of institutions and faculty with stratified samples and differential probabilities of selection. The first stage of sampling was at the institution level in which institutions were stratified according to a cross-classification of control by type, specifically two levels of control (public and private), and nine types of institutions based on Carnegie classification. The NSOPF institutional sample frame was drawn from the Integrated Postsecondary Education Data System. The second stage of sampling was at the faculty level, using lists of faculty and instructors obtained from the institutions identified through the first stage of sampling. Using NSF and National Endowment for the Humanities (NEH) analytical objectives, faculty groups that were oversampled included full-time females, Black non-Hispanics and Hispanics, Asian/Pacific Islanders, and faculty in four NEH-designated disciplines. Thus, if analyses are produced that do not appropriately apply weights, the results will be biased in favor of the groups that were oversampled (Selfa et al., 1997).

Raw Weight

The first objective in working with dataset weights is to determine which weight is appropriate. Within the NSOPF:93 public dataset, there are approximately 130 ordinal variables based on faculty responses, 1 faculty respondent raw weight variable (WEIGHT), and 32 replicate weights that can be used to calculate standard errors. The replicate weights can be used in procedures such as Jackknife Repeated Replication (JRR), Taylor-series estimation, or Balanced Re-

peated Replications (BRR; Stapleton, 2002). As discussed previously, the methodology reports are helpful in understanding which raw weight is the appropriate selection given the type of analyses (e.g., cross-sectional, longitudinal, multilevel). Given that this dataset has only 1 weight, selecting the appropriate weight is relatively straightforward.

Normalized Weight

The normalized weight can be calculated in two ways: (a) using the derived mean weight or (b) using the population and sample size. The normalized weight can be computed by dividing the raw weight by its mean (Peng, 2000; Thomas & Heck, 2001),

$$w_N = \frac{w_i}{\bar{w}},$$

where w_N is the normalized weight, w_i is the raw weight for the *i*th observation, and $\bar{w}$ is the mean weight. In this example, the NSOPF:93 mean weight (derived by using any of the appropriate descriptive statistics options in SPSS, such as frequencies, descriptive, or explore) is 40.10729. A normalized weight (NORMWT) can be computed easily in SPSS (Figure 1).

Using sample and population sizes in place of the mean value, the normalized weight also can be derived as the product of the raw weight and the ratio of the sample size to the population size (West & Rathburn, 2004),

$$w_N = w_i(n/N),$$

where w_N is the normalized weight, w_i is the raw weight for the *i*th observation, n is the sample size, and N is the population size. To calculate the normalized weight in SPSS using the sample and population sizes, first compute the sum of the raw weight variable (i.e., the population size) and the number of valid cases (i.e., the sample size) using any of the appropriate descriptive statistics options in SPSS (e.g., frequencies, descriptive, or explore). This yields a population size of 1,033,966 and a sample size of 25,780. A normalized weight can be computed easily in SPSS (Figure 2). The normalized weight can then be applied in the analysis using "weight cases" from the data option in the toolbar (Figure 3).

Design Effect Adjusted Weight

Although the normalized weight takes into account disproportionate sampling, it does so assuming a simple random sample. When a model-based approach is not used, the design effect must be included in the analyses to account for the clustered design and potential homogeneities that exist within the clusters. A design effect adjusted weight can be calculated by dividing the normalized weight by the design effect of the outcome variable,

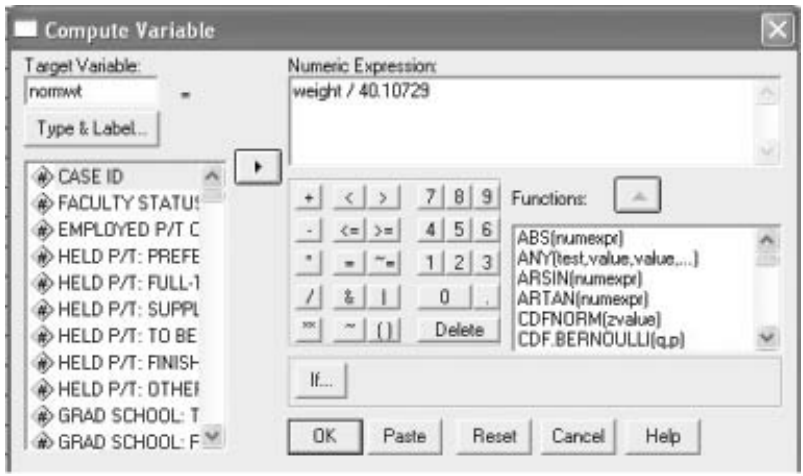

FIGURE 1. Computing normalized weight using the mean weight.

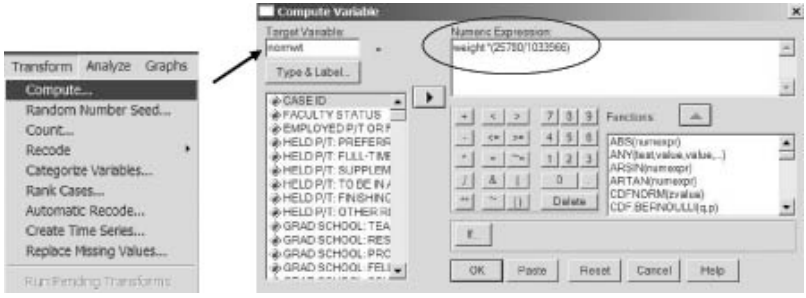

FIGURE 2. Computing the normalized weight using the population and sample size.

$$w_{\mathrm{DEFF}.N} = \frac{w_N}{\mathrm{DEFF}}.$$

In this example, the most appropriate design effect was located in the NSOPF:93 methodology report (Selfa et al., 1997, p. 45). Not all variables that may be of interest will have design effects provided in the methodology report. In this example, for instance, the variable in our public dataset is publication productivity over the past 2 years, whereas our design effect variable is more strictly defined as publication productivity over the past 2 years *in nonrefereed jour-*

nals (DEFF = 3.48). 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 over subgroups of the independent variable is appropriate to use (Huang, Salvucci, Peng, & Owings, 1996). Computing an adjusted weight using the normalized weight in SPSS is illustrated in Figure 4. The design effect adjusted weight can then be applied in the analysis using "weight cases" from the data option in the toolbar (Figure 2).

Fluctuation of Parameter Estimates in Varying Applications of Weights

Comparison of Means and Standard Errors

An example of how means and standard errors from NSOPF:93 variables (see Table 1 for variables, values, and labels) compare when weights are not used and when using raw weights compared with normalized and design effect adjusted

FIGURE 3. Applying the normalized weight.

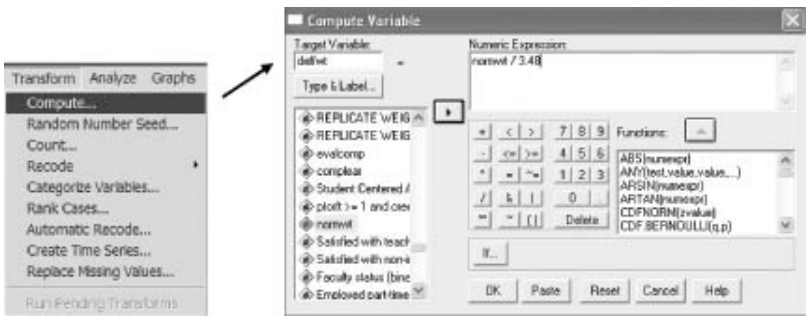

FIGURE 4. Computing the design effect adjusted weight.

weights is provided in Table 2. Using the design effect adjusted weights ensures that both disproportionate sampling and cluster sampling have been accounted for and thus produces the most accurate estimates. The sample size is reflected in the unweighted and normalized weighted models, and the population size is evident in the raw weighted model. The design effect adjusted weight reflects neither the sample nor the population size; however, this is not of concern to the researcher. The importance of the design effect adjusted weight is that, with its application to the data, the results reflect the underlying population (usually a nationally representative sample) regardless of the sample size reflected from the analysis.

As seen here, the means and standard errors of the unweighted variables differ when compared with the weighted means. Gender is a dummy-coded variable with 0 = male and 1 = female. The mean for gender (i.e., the proportion of women in the sample) is lower when no weight is applied, reflecting oversampled women in the NSOPF:93 design. For all variables, the means are stable regardless of which weight is applied. Because the population size is used in the

TABLE 1. National Study of Postsecondary Faculty (NSOPF:93) Variables

Variable label	Variable	Value
Gender	F51	1 = male; 0 = female
Total household income ($)	E49	1 = 0–39,999 2 = 40,000–54,999 3 = 55,000–69,999 4 = 70,000–84,999 5 = 85,000–99,999 6 = 100K and up
Satisfied with job overall	D40I	1 = Very dissatisfied 2 = Somewhat dissatisfied 3 = Somewhat satisfied 4 = Very satisfied
Number of for-credit classes taught	C22A	0 = 0 1 = 1–2 2 = 3–4 3 = 5 and over
Productivity: 2 years, publications	X13B20	0 = 0 1 = 1 2 = 2 3 = 3 4 = 4 5 = 5–9 6 = 10 and above

TABLE 2. Unweighted, Raw, Normalized, and Design Effect Adjusted Weighted Estimates

		Unweighted			Raw			Normalized			Design effect adjusted		
Variable label	Range	*N*	*M*	*SE*	*N*	*M*	*SE*	*N*	*M*	*SE*	*N*	*M*	*SE*
Gender	1	25,780	.57	.003	1,033,966	.61	.000	25,780	.61	.003	7,408	.61	.006
Total household income ($)	5	25,780	3.27	.011	1,033,966	3.37	.002	25,780	3.37	.011	7,408	3.37	.021
Satisfied with job overall	3	25,780	3.17	.005	1,033,966	3.20	.001	25,780	3.20	.005	7,408	3.20	.009
Number of for-credit classes taught	3	25,780	1.35	.005	1,033,966	1.17	.001	25,780	1.17	.005	7,408	1.17	.010
Productivity: 2 years, publications	6	25,780	1.76	.014	1,033,966	1.85	.002	25,780	1.85	.015	7,408	1.85	.027

derivation, the standard errors for the raw weighted sample are nearly nonexistent. The standard errors are largest when using design effect adjusted weights, approximately twice as large compared with the normalized weighted observations. This is expected as the standard errors are adjusted for the homogeneous clusters present in the original sample design by using the design effect adjusted weight.

Comparison of Independent t *Tests*

The impact on estimates when weights and design effects are used or are not used can also be illustrated in hypothesis tests. For example, I review a simple independent *t* test using gender as the independent variable and number of publications in the past 2 years as the dependent variable. As seen in Table 3, although the test is statistically significant in each case, the *t* test statistic varies dramatically, as expected. The results using the design effect adjusted weights reflect the widest confidence interval distance of the mean difference (.208) compared with the normalized weighted (.111), unweighted (.107), and raw weighted (.01613) results. This is anticipated, given the larger standard error when the cluster design is accommodated. Practical significance, as reflected in η^2, is largest for the design effect adjusted weighted sample and nearly twice as large as either the unweighted or normalized weighted samples. Although interpretation of η^2 is a small effect regardless of which model is reviewed (weighted or unweighted), the larger effect size for the sample using the design effect adjusted weight is important to consider. When similarities among clusters are accommodated by incorporation of the design effect, practical significance is illuminated in the model.

This example is simplistic, but it illuminates the potential for increased Type I

TABLE 3. Comparison of Independent *t* Test Results With Weighting Options

Weighting option	*t*	*df*	*SE* of difference	*p*	95% CI of difference	η^2
Unweighted	25.410	25420.693	.027	.000	.658, .765	.02443
Raw	180.581	942515.1	.004	.000	.799, .817	.03057
Normalized	28.513	23497.56	.028	.000	.753, .864	.03057
Design effect adjusted	15.283	6750.508	.053	.000	.704, .912	.03057

Note. Equal variances are not assumed. CI = confidence interval.

errors in hypothesis testing when ignoring the complex sampling design. It may seem illogical to conduct tests of inference using the raw weight (i.e., the population), but a novice researcher may mistake the application of the raw weight to the analysis as appropriate compensation for the sampling design and oversampled groups.

Although not shown here, adjusting the *t* test statistic of the normalized weighted analyses by DEFT (Strategy 2) yields nearly identical conclusions as when the design effect adjusted weight is applied to the sample.

Comparison of Weights on Multiple Linear Regression

As shown with the *t* test, similar weighted versus unweighted results are seen from a simplistic multiple linear regression analysis. In this example, two composite satisfaction variables are computed based on the results of a principal components factor analysis: Predictor 1—satisfaction with teaching (sum of five variables, such as satisfaction with quality of graduate students, satisfaction with authority to decide on courses taught, and other), and Predictor 2—satisfaction with job structure (sum of eight variables, such as satisfaction with workload, satisfaction with advancement opportunities, satisfaction with salary, and other). The two satisfaction variables, collectively explaining approximately 55% of the total variance, serve as predictor variables with number of publications in the past 2 years as the criterion variable.

The correlation matrix (Table 4) indicates little difference between the unweighted and weighted models in the bivariate correlation between productivity in publications and teaching satisfaction and no differences between satisfaction with teaching and satisfaction with job structure. However, the bivariate correlation between number of publications and satisfaction with job structure indicates a nonsignificant negative relationship when unweighted ($r = -.002$) and a positive relationship when weighted ($r = .020$). The positive bivariate relationship is significant for the raw weighted sample but not significant for the design effect adjusted weighted sample. Had a researcher used either the raw or normalized weight and not accommodated for intracluster correlation, an erroneous decision would have been made in interpreting the correlation between number of publications and satisfaction with job structure. Specifically, a different statistical decision would have been reached, including possibly committing a Type I error.

The multiple linear regression (Table 5) *F* test indicates an overall significant regression model, regardless of using or not using a weight and regardless of which weight was used. However, the *F* test statistic is substantially smaller when the design effect adjusted weight is used. In the regression model, the regression coefficients are stable regardless of the weight used. For Predictor 1 (satisfaction with teaching), the estimated regression coefficients are smaller (approximately 20%) when weights are applied compared with when weights are

TABLE 4. Comparison of Correlation Results With Weighting Options

Variable	X_1				X_2			
	Unweighted	Raw	Normalized	Design effect adjusted	Unweighted	Raw	Normalized	Design effect adjusted
Number of publications past 2 years (X_1)	—	—						
Satisfaction with teaching (X_2)	.058*	.052	.052	.052	—	—		
Satisfaction with job structure (X_3)	–.002	.020**	.020**	.020	.066	.066	.066	.066

*$p < .01$. **$p < .05$. ***$p < .001$.

TABLE 5. Comparison of Multiple Linear Regression Results With Weighting Options in SPSS

Statistic	Intercept	Satisfaction with teaching	Satisfaction with job structure
B			
Unweighted	1.732	.009	–.003
Raw weight	1.608	.007	.008
Normalized weight	1.608	.007	.008
Design effect adjusted weight	1.608	.007	.008
SE			
Unweighted	.067	.001	.003
Raw weight	.011	.000	.000
Normalized weight	.069	.001	.003
Design effect adjusted weight	.129	.002	.006
β			
Unweighted	—	.059	.006
Raw weight	—	.051	.017
Normalized weight	—	.051	.017
Design effect adjusted weight	—	.051	.017
t			
Unweighted	25.907	9.413	–.928
Raw weight	147.299	51.524	16.984
Normalized weight	23.258	8.135	2.682
Design effect adjusted weight	12.466	4.360	1.437
p			
Unweighted	.000**	.000**	.354
Raw weight	.000**	.000**	.000**
Normalized weight	.000**	.000**	.007*
Design effect adjusted weight	.000**	.000**	.151

Weighting option	R^2	Adjusted R^2	F	ANOVA p
Criterion variable: number of publications in past 2 years				
Unweighted	.003	.003	(2, 25777) = 44.351	.000**
Raw	.003	.003	(2, 1033963) = 1536.386	.000**
Normalized	.003	.003	(2, 25777) = 38.303	.000**
Design effect adjusted	.003	.003	(2, 7405) = 11.003	.000**

*$p < .01$. **$p < .05$. ***$p < .001$.

not applied, and β is approximately 15% smaller when weights are applied. For Predictor 2 (satisfaction with job structure), the estimated regression coefficients are larger when weights are applied (over 2.5 times as large), and β is nearly 3 times larger when weights are applied. The standard error parameter estimates are largest for the design effect adjusted weighted model, as expected. The *t* test statistics are smallest for the design effect adjusted weighted sample, comparatively half the size of the unweighted and normalized weighted samples. As expected, the raw weighted sample produces the largest *t* statistic given the population size associated with the raw weight. The striking differences in the *t* statistic values are then evident in variable significance. Significance is stable for the intercept and Predictor 1 (satisfaction with teaching) regardless of using or not using a weight. However, Predictor 2 (satisfaction with job structure) is significant only when the raw and normalized weights are applied. Again, had the researcher believed that weighting the sample without consideration of intracluster correlation was sufficient, a different statistical decision would have been reached, including the possibility of a Type I error.

Weights and Design Effects Applied in AM Using ECLS-K

The public-use NSOPF:93 dataset does not contain strata or cluster variables; therefore, analyzing the data using specialized software such as AM is not advantageous for that particular dataset. To illustrate the use of specialized software for dealing with complex sample designs using AM, I used the NCES ECLS-K public-use dataset. The ECLS-K is one of two cohorts (the other being a birth cohort) that make up the ECLS longitudinal study. The ECLS-K provides descriptive data on children at entry and transition into school and progress through Grade 5. Family, school, and community variables, along with individual variables, are available. Using a multistage probability sample design to select a nationally representative sample of children attending kindergarten in 1998–1999, the primary sampling units (PSUs) were geographic areas (counties or groups of counties). Schools were the second-stage units within PSUs sampled, and students within schools were the third-stage units. In the base year, Asian/Pacific Islanders were oversampled. A subsample of ECLS-K PSUs was selected for the fall first-grade data collection (U.S. Department of Education, 2002). In this example, only students who had scores on the first three reading assessments (reading IRT scale score in fall kindergarten, C1RRSCAL; reading IRT scale score in spring kindergarten, C2RRSCAL; and reading IRT scale score in fall first grade, C3RRSCAL) were included in the subset analyzed.

AM software can be downloaded free of charge from http://www.am.air.org. The first step in using AM software is to import data. Bringing SPSS data files into AM is an easy process using the import function (Figure 5). Once the data are brought into AM, designating the weight is performed by right clicking on

the weight variable and selecting "edit metadata." This will bring up a dialog box that allows the researcher to assign the variable the role of weight (Figure 6). The strata and cluster (PSU) variables in the dataset must also be defined using this same process.

A number of basic statistics can be generated using AM software (Figure 7). In this example, I conduct a multiple linear regression using reading IRT scale scores from fall and spring kindergarten as the predictor variables and reading IRT scale score from fall first grade as the criterion variable. The independent and dependent variables are dragged from the left column into their respective dialog boxes on the right. Users also have the option of selecting how the output is generated (Web browser, spreadsheet, or text file; Figure 8). Once all the variables and options are defined, click "OK" to generate the output.

FIGURE 5. Importing SPSS data into AM.

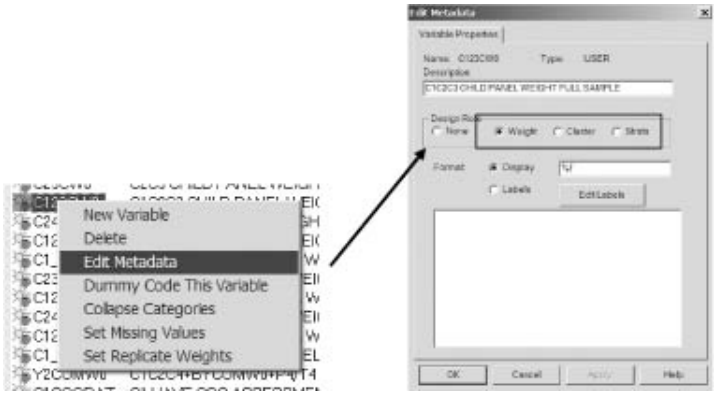

FIGURE 6. Designating the role of weight, strata, and cluster variables.

As anticipated, ignoring the complex sample design in AM software compared with defining the weight, cluster, and strata produces underestimated standard errors and inflated test statistic values (Table 6). Differences in slope and regression estimates are also present. When weights are ignored, the fall kindergarten reading score predictor has a smaller estimate, and the intercept and spring

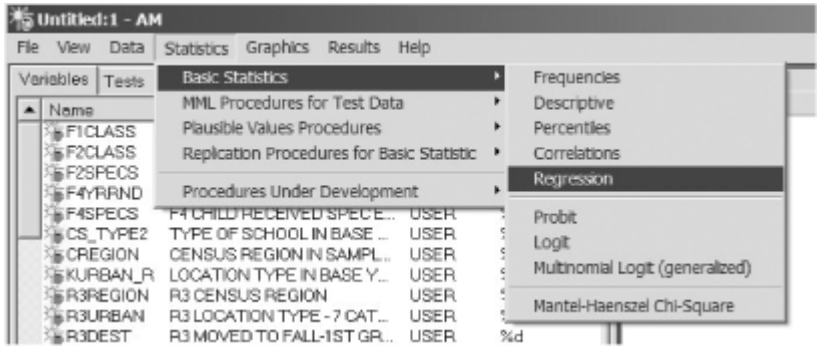

FIGURE 7. Generating basic statistics with AM.

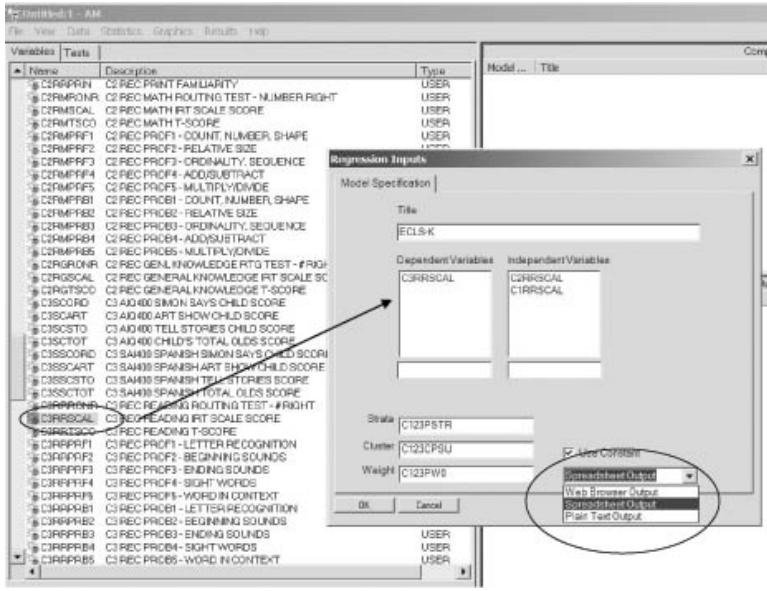

FIGURE 8. Model specification.

TABLE 6. Comparison of Early Childhood Longitudinal Study—Kindergarten Class of 1998–1999 (ECLS-K) Multiple Linear Regression Results With Weighting Options Using AM Software

	B		*SE*		*t*		*p*	
Variable	Unweighted	Adjusted	Unweighted	Adjusted	Unweighted	Adjusted	Unweighted	Adjusted
Intercept	4/264	4.214	.258	.453	16.513	9.307	.000***	.000***
Reading IRT scale score (fall kindergarten)	.200	.216	.018	.026	11.240	8.329	.000***	.000***
Reading IRT scale score (spring kindergarten)	.891	.884	.014	.018	63.188	49.902	.000***	.000***
	Unweighted		Adjusted					
R^2	.823		.817					
Adjusted Wald F[a]	(2, 4334) = 8621.64		(2, 84) = 3867.21					

Note. Criterion variable = reading IRT scale score, fall first grade.
[a]Denominator degrees of freedom in the adjusted Wald *F* test is calculated as the number of clusters minus the number of strata (unless a stratum has a single primary sampling unit) minus the number of regressors (J. Cohen, personal communication, December 16, 2003).
*$p < .01$. **$p < .05$. ***$p < .001$.

kindergarten reading score predictor have a larger estimate compared with the estimates produced when the weight, cluster, and strata are defined.

Weights and Design Effects Applied in SPSS Using ECLS-K

How do estimates fare from the same model when using SPSS? Using SPSS, I created a normalized weight by dividing the panel weight (C123CW0) by the mean weight of the entire sample (224.1938). The design effect adjusted weight was derived from the ratio of the normalized weight to the design effect for first-grade reading IRT scale score (C3RRSCAL; DEFF = 5.291). As with the analyses using the AM software, only students who had scores on the first three reading assessments (reading IRT scale score in fall kindergarten, C1RRSCAL; reading IRT scale score in spring kindergarten, C2RRSCAL; and reading IRT scale score in fall first grade, C3RRSCAL) were included in the subset analyzed.

The influence of using or ignoring weights and design effects in SPSS is illustrated when comparing the various models (Table 7). The differences between estimates and test statistic values of unweighted, raw weighted, normalized weighted, and design effect adjusted weighted samples are similar to the results generated

TABLE 7. Comparison of Early Childhood Longitudinal Study—Kindergarten Class of 1998–1999 (ECLS-K) Multiple Linear Regression Results With Weighting Options Using SPSS Software

	B				*SE*			
Variable	U	Raw	N	A	U	Raw	N	A
Intercept	4.264	4.214	4.214	4.214	.260	.009	.136	.313
Reading IRT scale score (fall kindergarten)	.200	.216	.216	.216	.016	.001	.008	.019
Reading IRT scale score (spring kindergarten)	.891	.884	.884	.884	.013	.000	.007	.015
	Unweighted				Raw			
R^2	.823				.817			
ANOVA *F*	(2, 4333) = 297191.083*				(2, 3531848) = 7886407*			

Note. Criterion variable = reading IRT scale score, fall first grade. U = unweighted. N = normalized weighted. A = design effect adjusted weighted.
*$p < .01$. **$p < .05$. ***$p < .001$.

using the NSOPF:93 data in a multiple regression procedure (specifically, larger standard errors and smaller test statistic values with the design effect adjusted weighted sample). Therefore, I focus on the differences of estimates, standard errors, test statistic values, and significance of the AM versus the SPSS analyses.

ECLS-K Results: AM Versus SPSS

AM software allows the researcher to accommodate the weight, strata, and cluster variables specifically; therefore, correct standard error estimates are produced when the variables are defined in these roles. The differences in the unweighted SPSS model compared with the unweighted AM model may be due to the assumption in SPSS of homoscedastic residuals (J. Cohen, personal communication, May 18, 2004). The calculation of standard errors in AM does not have this assumption (Cohen). The noticeable differences in standard errors are, therefore, most likely caused by a violation of the homoscedastic residual assumption. Thomas and Heck (2001) found conservative standard errors were produced when using design effect adjusted weights as when using SAS PROC SURVEYREG, a recent SAS procedure that allows for adjustment of standard errors

	t				*p*			
	U	Raw	N	A	U	Raw	N	A
	16.405	464.275	31.004	13.473	.000*	.000*	.000*	.000*
	12.839	385.173	25.722	11.178	.000*	.000*	.000*	.000*
	70.881	1971.980	131.689	57.227	.000*	.000*	.000*	.000*
	Normalized				Design effect adjusted			
	.817 (2, 15751) = 35170.069*				.817 (2, 2974) = 6641.717*			

for complex samples, and this was also the case in this analysis. Design effect adjusted weights produced underestimated standard errors in SPSS compared with the correct standard errors using the weight, strata, and cluster variables in AM software. SPSS produced larger test statistic values. Although rejection of the hypotheses remains the same across all models, the larger standard errors and resulting smaller test statistic values generated when using the AM software suggest that, given a different model, the chance of committing a Type I error will increase substantially when using design effect adjusted weights in SPSS.

To remove the potential impact of multicollinearity between the independent variables (r = .817), a linear regression was generated using only spring kindergarten item response theory [IRT] reading scale scores to predict fall first-grade reading IRT scale scores. I used the appropriate weight (C23CW0) and mean weight (224.2039) to first create a normalized weight and then a design effect adjusted weight. Because the dependent variable remained the same, the same DEFF (5.291) was applied. Comparing AM with SPSS results in this simplified model, similar outcomes were produced compared with the multiple linear regression model. The estimates from the unweighted models were different. AM produced larger standard errors and smaller test statistics compared with those from the unweighted sample in SPSS. When applying design effect adjusted weights in SPSS and the appropriate weight, strata, and cluster variable in AM, standard errors in SPSS were more than 60% smaller than those produced using AM. In addition, test statistic values were larger in SPSS. Therefore, assuming accurate standard errors, estimates, and test statistics are generated using AM software, if design effect adjusted weights are applied in SPSS, the researcher should consider using a more conservative alpha level.

Recommendations

Two of the most critical components to consider when using national datasets that employ a complex design are the use of weights and variance estimation (Peng, 2000). Ignoring disproportionate sampling may result in biased parameter point estimates as a result of oversampled populations, and ignoring multistage sampling runs the risk of underestimated standard errors caused by homogeneous clusters and potentially increased Type I errors (Stapleton, 2002). Poor performance of test statistics and confidence intervals are also an increased possibility (Pfeffermann, 1993). I stand in agreement with Kalton (1989, p. 579) who stated, “my own view is that in most—but not all—circumstances it is preferable to conduct a weighted analysis and to compute standard errors appropriate for the sample design employed.” Similar statements have been echoed by other researchers (e.g., Pfeffermann; Thomas & Heck, 2001) and generalized to most social scientists and survey statisticians (Hoem, 1989).

Variance estimation can be performed by an exact method (e.g., using special-

ized software) as introduced in Strategy 1 or by an approximation method (e.g., applying the design effect in tandem with the weights) as introduced in this article in Strategies 2 and 3. The use of specialized software will yield exact variances as they are designed to accommodate the clustering effect. Therefore, when specialized software (e.g., AM, SUDAAN, WesVar) provides the appropriate methodological features to answer the research question and is accessible to the researcher, this is the most desirable solution for analyzing data from complex samples.

However, not all researchers have access to specialized software other than AM. Although AM is free to download online, it is still in the Beta testing version and not all statistical procedures are available that may be desired. It has been argued that multilevel modeling (i.e., model-based approach) is the appropriate statistical strategy for analyzing multistage samples. However, multilevel modeling is not always appropriate for complex surveys, and not all researchers may be interested in multilevel modeling (Kaplan & Elliott, 1997). The available datasets may not have the appropriate institution-level variables for the specified model (Kaplan & Elliott). For example, Kaplan and Elliott used the National Educational Longitudinal Study of 1988 (NELS:88; Huang et al., 1996) to study critical transitions experienced by students as they leave elementary school and progress to higher levels of schooling. Kaplan and Elliott selected a subset of variables from the student survey, teacher survey, and school survey, resulting in a sample size of 1,165 and a school-level sample size of 356. Although the structural equation model indicated adequate fit, the observed effects were quite small overall—leading the authors to conclude that this result was owed in part "to the fact that the NELS:88 test has too few items to be sensitive to school or classroom instructional indicators" (p. 342). It can be argued whether accurate parameter estimates based on weight and design effect corrections in a design-based approach or good overall model fit through the application of a model-based approach are more important. However, an overall good model fit that reflects inaccurate parameter estimates seems of ill use.

When a multilevel model is not appropriate or not desirable, a single-level model or design-based approach can be employed. When a single-level model is used, the researcher is faced with the same concern as when using model-based approaches in how to accurately adjust the data to ensure unbiased estimates caused by the oversampled groups (i.e., the use of weights) and is faced with an additional concern about the potential homogeneous clusters that exist as result of multistage sampling (i.e., the use of design effects). Design effect Strategies 2 and 3 presented in this study illustrate how design effects can be applied in conjunction with weights to produce more accurate standard errors. If the researcher does not want to deal with applying either DEFF or DEFT to each analysis as presented in Strategy 2, then Strategy 3 is the most viable option. Strategy 3 provides the most straightforward accommodation of homogeneous clusters because

the design effect for the dependent variable is applied directly to the weight to create a new, adjusted weight.

There are a number of forms that the weight variable can assume: raw, normalized, and design effect adjusted. Of these, the most appropriate to apply to studies that use design-based approaches when specialized software is not available is the design effect adjusted weight. Although the normalized weights reflect the sample size rather than the population size, the estimates generated using normalized weights are reflective of a simple random sample because the nesting nature of the clusters is not taken into account, and the result is underestimated standard errors. The result of using normalized weights in design-based approaches is underestimated standard errors and inaccurate test statistics, among others. As seen in the examples presented in this article, employing a strategy to accommodate *both* oversampled groups and cluster sampling (i.e., using weights *and* design effects) leads to the most accurate parameter estimates and the decreased potential of committing a Type I error. This suggestion has been recommended by other researchers (e.g., Thomas & Heck, 2001) and is recommended here. However, using a design effect adjusted weight in SPSS may produce underestimated standard errors when compared with accurate estimates produced by specialized software, such as AM. The examples presented here are simplified, however, and results may differ using other variables. On a final note, comparing weighted with unweighted models has been done (e.g., DuMouchel & Duncan, 1983) and is the suggested process by Skinner, Holt, and Smith (1989) to determine model adequacy. As shown in some of the examples presented here, inference may not change from the unweighted to weighted models. However, if a researcher goes to the trouble of analyzing weighted and unweighted data, failing to report the most accurate results (as reflected in the design effect adjusted weighted analysis) is cause for concern. Although statistical significance may not change from weighted to unweighted results, that may not be the case with measures of effect.

Researchers who use secondary data should anticipate a learning curve prior to beginning the data analysis. An investment of time in reading the methodology reports to understand the sampling design and any oversampling is important. In general, if a complex sample design is used and the study has oversampled one or more groups and a single-level analysis (or design-based approach) is applied, weights and design effects need to be incorporated in the analysis. The strategies presented will assist researchers who are new or less experienced with weights and design effects in understanding the application of these strategies to secondary data.

Future Research

There are a number of areas ideal for future research. Extant data studies that continue to explore methodological procedures when using weights and design

effects with complex samples are needed to further clarify appropriate procedures. Simulation and extant data studies that compare results of the application of weights and design effects to single-level models (i.e., design-based approaches) with results obtained from model-based approaches to determine the extent the results differ may also enlighten this line of research.

REFERENCES

AM statistical software. (n.d.). Retrieved May 5, 2004, from http://www.am.air.org/

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.* Unpublished doctoral dissertation, University of Alabama, Tuscaloosa.

Heck, R. H., & Mahoe, R. (2004, April). *An example of the impact of sample weights and centering on multilevel SEM models.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Hoem, J. M. (1989). The issue of weights in panel surveys of individual behavior. In D. Kasprzyk, G. Duncan, G. Kalton, & M. Singh. (Eds.), *Panel surveys* (pp. 539–559). New York: Wiley.

Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). New York: Springer-Verlag.

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

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.

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. (1997). A model-based approach to validating education indicators using multilevel structural equation modeling. *Journal of Educational 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: A Multidisciplinary Journal, 6*(4), 305–321.

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

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

Korn, E. L., & Graubard, B. I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician, 49*(3), 291–295.

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

Longford, N. T. (1995). *Model-based methods for analysis of data from 1990 NAEP trial state assessment* (NCES Publication No. 95-696). Washington, DC: National Center for Education Statistics.

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

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

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

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, 60*(Series B), 23–40.

Potthoff, R. F., Woodbury, M. A., & Manton, K. G. (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association, 87*(418), 383–396.

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:90/94) final technical report* (NCES Publication No. 96-153). Washington, DC: U.S. Government Printing Office.

Selfa, L. A., Suter, N., Myers, S., Koch, S., Johnson, R. A., Zahs, D. A., et al. (1997). *1993 National Study of Postsecondary Faculty (NSOPF:93) methodology report* (NCES Publication No. 97-467). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

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

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

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.

Tourangeau, K., Nord, C., Le, T., Wan, S., Bose, J., & West, J. (2002). *User's guide to the longitudinal kindergarten-first grade public-use data file* (NCES Publication No. 2002-149). Washington, DC: National Center for Education Statistics.

U.S. Department of Education. (2002). *User's manual for the ECLS-K first grade public-use data files and electronic code book* (NCES Publication No. 2002-135). Washington, DC: National Center for Education Statistics.

West, J., & Rathburn, A. (2004, April). *ECLS-K technical issues.* Presentation at the AERA Institute on Statistical Analysis for Education Policy, San Diego, CA.