Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 16, Number 12, September 2011 ISSN 1531-7714

Best Practices in Using Large, Complex Samples: The Importance of Using Appropriate Weights and Design Effect Compensation

Jason W. Osborne
Old Dominion University

Large surveys often use probability sampling in order to obtain representative samples, and these data sets are valuable tools for researchers in all areas of science. Yet many researchers are not formally prepared to appropriately utilize these resources. Indeed, users of one popular dataset were generally found *not* to have modeled the analyses to take account of the complex sample (Johnson & Elliott, 1998) even when publishing in highly-regarded journals. It is well known that failure to appropriately model the complex sample can substantially bias the results of the analysis. Examples presented in this paper highlight the risk of error of inference and mis-estimation of parameters from failure to analyze these data sets appropriately.

Large, governmental or international data sets are important resources for researchers in the social sciences. They present researchers with the opportunity to examine trends and hypotheses within nationally (or internationally) representative data sets that are difficult to acquire without the resources of a large research institution or governmental agency.

However, there are challenges to using these types of data sets. For example, individual researchers must take the data as given—in other words, we have no control over the types of questions asked, how they are asked, to whom they are asked, and when they are asked. The variables are often not ideally suited to answering the particular questions you, as an individual researcher might wish to ask.

Despite their potential shortcomings, these valuable resources are often freely available to researchers (at least in public release formats that have had potentially identifying information removed). There is, however, one cost worth discussing: the expectation that researchers will utilize best practices in using these samples. Specifically, researchers must take the time to understand the sampling methodology used and appropriately utilize weighting and design effects, which to a novice can be potentially confusing and intimidating. There is mixed evidence on researchers' utilization of appropriate methodology (e.g., Johnson & Elliott, 1998), which highlights the need for more conversation around this important issue. The goal of this brief paper is to introduce some of the issues around using complex samples and explore the possible consequences (e.g., Type I errors) of failure to appropriately model the complex sampling methodology.

What types of studies use complex sampling?

Many of the most interesting social science and health sciences databases available to researchers use complex sampling. For example, data from the National Center for Educational Statistics in the USA (e.g., National Education Longitudinal Study of 1988 (NELS88), Third International Mathematics and Science Study (TIMSS), etc.)[i], Centers for Disease Control and Prevention (e.g., National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES)),[ii] and the Bureau of Justice Statistics (e.g., National Crime Victimization Survey (NCVS).[iii] Almost any survey seeking a representative sample from a large population will probably have a complex multi-stage probability sampling methodology, as it is relatively efficient and allows for estimation of representative samples.

Why does complex sampling matter?

In most of the examples cited above, the samples are not simple random samples, but rather complex samples with multiple goals. For example, in NELS 88, students in certain underrepresented racial groups and in private schools were *oversampled* (i.e., more respondents selected than would

typically be the case for a representative sample), meaning that the sample is not, in its initial form, necessarily representative (Ingels, 1994; Johnson & Elliott, 1998). Furthermore, in any survey such as the ones discussed above there is a certain amount of non-response that may or may not be random, making unweighted samples potentially still less representative.

Finally, in multistage probability sampling, in contrast to simple random sampling, complex sampling often utilizes cluster sampling (especially where personal interviews are required), where clusters of individuals within primary sampling units are selected for convenience (e.g., in the Education Longitudinal Study of 2002, approximately 20,000 students were sampled from 752 schools, rather than simply random sampling from the approximately 27,000 schools that met criteria within the United States (Bozick, Lauff, & Wirt, 2007)). Thus, students within clusters are more similar than students randomly sampled from the population as a whole. This effectively reduces the information contained in each degree of freedom. Called "design effects" (Kish, 1965 is often credited with introducing this concept) these effects of sampling must also be accounted for or the researcher risks not only mis-estimating effects, but making Type I errors because this common modern sampling strategy can lead to violation of traditional assumptions of independence of observations. Specifically, without correcting for design effects, standard errors are often underestimated, leading to significance tests that are inappropriately sensitive (e.g., Johnson & Elliott, 1998; Lehtonen & Pahkinen, 2004).

Note that complex, multi-stage probability sampling is *not* the same as multi-level (i.e., nested, or hierarchical) analyses or data. Both are important, modern techniques that correctly deal with different violations of assumptions or issues. Multi-stage probability sampling has to do with the sampling methodology employed, which violates our assumption that samples are drawn randomly from populations of interest.[1] In particular, using this methodology violates our assumption that each data point represents an equal amount of information, and in the case of cluster sampling, can also violate assumptions of independence of observations. Some sub-populations of interest might be over-sampled, and others might be under-sampled, and thus from a conceptual point of view, each data point represents a different portion of the overall population, which can be corrected by methods discussed below.

[1] In fact, I think most sampling methodologies violate this assumption. Inspection of top journals in any field reveal few studies that could be correctly classified as simple random samples from a population of interest.

Multi-level or nested data are data with variables measured at different levels of organization (e.g., students within classrooms within schools, or employees within corporations within sectors). This violates assumptions of independence of observations (for a brief primer on this topic, see Osborne, 2000, 2008) that can lead to mis-estimation of parameters and mis-specification of analysis models if not taken into account.

Note also that the two are not mutually exclusive. For example, many of the data sets discussed above are both nested data *and* complex samples. For example, in the NCES data, researchers often want to model teacher- or school-level effects on student performance, which creates a multi-level analysis in the context of a complex, multi-stage probability sample. But researchers can also encounter nested data sets that are not produced using probability samples, and probability samples that are not nested.

In sum, there are two issues introduced by complex sampling: a sample that employs advanced sampling techniques (or has non-response or missing data that needs to be accounted for), causing the sample to potentially deviate from representative of the population of interest, and a sample that violates assumptions of independence of observations, potentially leading to significant mis-estimation of significance levels in inferential statistical tests.

What are best practices in accounting for complex sampling?

In most samples of this nature the data provider includes information in the data set (and in the user documentation) to facilitate appropriate use of the data. For example, weights for each individual, information about design effects (DEFFs) for the overall sample and different subpopulations, and information on which primary sampling unit and cluster each individual belongs to.

More information on these topics is available in most user manuals for those interested in the technical details of how each of these pieces of information are calculated and used.[iv]

Most modern statistical packages can easily apply sample weights to a data set. Applying the appropriate weight creates a sample that is representative of the population of interest (e.g., 8th graders in the US who remained in school through 12th grade, to continue the previous example from NELS88). The problem is that application of weights dramatically increases the sample size to approximately the size of the population. For example, in NELS88, for example, a sample of approximately 25,000 becomes the population of over 3,000,000 students), dramatically (and illegitimately) inflating the degrees of freedom used in inferential statistics). Previous best practices included scaling the weights, so that the weighted

sample has the same weighted number of participants as the original, unweighted sample. I did this in some of my early research (Osborne, 1995, 1997) thanks to the mentoring of Robert Nichols, one of the faculty I worked with. But scaling the weights doesn't take into account the design effects, which should further reduce the degrees of freedom available for the statistical tests.

Not all statistical software provides for accurate modeling of complex samples (e.g., with SPSS an add-on module is required; in SAS, STATA, and SUDAAN, complex sampling appears to be incorporated, and there is also freely available software such as AM[v] that correctly deals with this issue. However many software packages that incorporate complex sampling do not allow for advanced analyses such as structural equation modeling, hierarchical linear modeling, etc.).[2] For those without access to software that models complex samples accurately (again, as was the case long ago when I first started working with large data sets) one way to approximate best practices in complex sampling would be to further scale the weights to take into account design effects (e.g., if the DEFF = 1.80 for whatever sample or sub-sample a researcher is interested in studying, that researcher would divide all weights by 1.80).

However, the most desirable way of dealing with this issue is using software that has the capability to directly model the weight, primary sampling unit, and cluster directly, which best accounts for the effects of the complex sampling (e.g., Bozick et al., 2007; Ingels, 1994; Johnson & Elliott, 1998). In most cases, a simple set of commands informs the statistical software what weight you desire to use, what variable contains the PSU information, and what variable contains the cluster information, and the analyses are adjusted from that point on, automatically.

Does it really make a difference in the results?

Some authors have argued that, particularly for complex analyses like multiple regression, it is acceptable to use unweighted data (e.g., Johnson & Elliott, 1998). This advice is in direct opposition to the sampling and methodology experts who create many of these data sets and sampling frames, and is also in opposition to what makes conceptual sense. Thus, in order to explore whether this really does have the potential to make a substantial difference in the results of an analysis, I performed several analyses below under four different conditions that might reflect various strategies researchers would take to using this sort of data: (a) unweighted (taking the sample as is), (b) weighted only (population estimate), (c) weighted, using weights scaled to maintain original sample size and scale weights to account for DEFF (best approximation),and (d) using appropriate complex sampling analyses via AM software, which is designed to accurately account for complex sampling in analyses.

METHODS

In order to examine the effects of utilization of best practices in modeling complex samples, the original 10th grade (G10COHRT=1) cohort from the Education Longitudinal Study of 2002 (along with the first follow-up) public release data was analyzed. Only students who were part of the original cohort (G10COHRT=1) and who had weight over 0.00 on F1PNLWT (the weight for using both 10th and 12th grade data collection time points) were retained so that the identical sample is utilized throughout all analyses.

Condition

Unweighted. In this condition, the original sample (meeting condition G10COHRT=1 and F1PNLWT>0.00) was retained with no weighting or accommodation for complex sampling. This resulted in a sample of N=14,654.

Weighted. In this condition, F1PNLWT was applied to the sample of 14,654 who met the inclusion criteria for the study. Application of F1PNLWT inflated the sample size to 3,388,462. This condition is a likely outcome when researchers with only passing familiarity with the nuances of weighting complex samples attempt to use a complex sample.

Scaled weights. In this condition, F1PNLWT was divided by 231.232 (the ratio of the inflated sample size with weights applied and the unweighted sample: 3,388,462/14,654), bringing the sample size back to approximately the original sample size but retaining the representativeness of the population. Further, the weights were scaled by the design effect (1.88 for examples using only males yielding a final sample of 3923 males, or 2.33 for examples using all subjects, yielding a final sample of 6,289)[3] to approximate use of best practices. This condition is a likely outcome when a researcher is sophisticated enough to understand the importance of correcting for these issues but does not have access to software that appropriately models the complex sampling (or is using advanced analytical techniques such as

[2] I was unable to determine if the R statistical software incorporates complex sample handling, but encourage readers to explore R as an option as it often has advanced techniques incorporated prior to commercial programs. R is freely available on Unix, Windows, and Macintosh platforms at http://cran.r-project.org/

[3] These DEFF estimates are usually easily found in the user manuals for these data sets. Researchers need to decide what aspects of the sample they are interested in using, and utilize the appropriate DEFF estimate for that aspect of the sample, as I did. Not all DEFF are the same for all subgroups.

structural equation modeling that does not incorporate complex sampling methodology at this time).

Appropriately modeled. In this case, AM software was utilized to appropriately model the weight, PSU, and cluster information provided in the data to account for all issues mentioned above. This is considered the "gold standard" for purposes of this analysis.

RESULTS

Four different analyses were compared to explore the potential effects of failing to use best practices in modeling complex samples

Large effect in OLS regression.

In this example, 12th grade mathematics IRT achievement score (*F1TXM1IR)* is predicted from base year reading IRT achievement score (*BYTXRIRR*) controlling for socioeconomic status (F1SES2). The results of this analysis

Table 1: Large effect: OLS regression predicting F1 Math achievement from BY Reading Ach

Analysis	Group	b	SE	*t (df)*	*p* <	Beta
SPSS- no weighting	WhiteM	1.009	.019	14.42 (3858)	.0001	.647
	AfAmM	0.959	.040	23.91 (807)	.0001	.638
SPSS- weight only	WhiteM	1.027	.001	872.25 (927909)	.0001	.658
	AfAmM	0.951	.003	379.40 (201334)	.0001	.642
SPSS- weights scaled for N, DEFF	WhiteM	1.027	.025	41.806 (2132)	.0001	.658
	AfAmM	0.951	.052	18.138 (460)	.0001	.642
AM weight, PSU, Strata modeled	WhiteM	1.027	.023	45.35 (362)	.0001	
	AfAmM	0.951	.049	19.41 (232)	.0001	

Note: males only; BYTXRIRR predicting F1TXM1IR controlling for F1SES2; identical sample. In all analyses- i.e. G10COHRT=1, F1PNLWT>0. Lower right cell empty as AM does not provide standardized regression coefficients.

across all four conditions are presented in Table 1.

As Table 1 shows, with a strong effect (e.g., $\beta > 0.60$) there is not a substantial difference in the effect regardless of whether the complex sampling design is accounted for or not. However, note that the standard errors vary dramatically across condition, with the weighted only condition being mis-estimated by a factor of 16 times or more. Note also that the scaled weights condition closely approximates the appropriately modeled condition. However, as following analyses will show, this is possibly the exception, rather than the rule.

Modest effect in binary logistic regression.

To test the effects of condition on a more modest effect, African American males were selected for a logistic regression predicting dropout (F1DOSTAT; 0=never, 1=dropped out), from the importance of having children, controlling for standardized reading test scores in 10th grade.

Table 2: Modest effect: Logistic regression predicting dropout from Importance having Children

Analysis	b	SE	Wald	p <	EXP(b)
SPSS- no weighting	-0.09	0.146	5.59	.018	0.709
SPSS- weight only	-0.346	0.008	1805.85	.0001	0.708
SPSS- weights scaled for N, DEFF	-0.344	0.170	4.154	.042	0.708
AM weight, PSU, Strata modeled	-0.346	0.177	3.806	.052	

Note: African American males only; F1DOSTAT never vs. DO only; controlling for BYTXRSTD. AM did not give odds ratios, and the lower right cell is empty.

The results of these analyses are presented in Table 2.

The results indicate that the conclusions across all four analyses are similar—that as the importance of having children increases, the odds of dropping out decrease among African American males. However, there are several important differences across the conditions. First, the standard error of *b* varies dramatically across the four analyses, again mis-estimating the SE by up to 22 times. Second, the results from the scaled weights analyses and the appropriately modeled analysis were most similar. Finally, this analysis is an example of a potential Type I error: using the original sample with no weights or non-scaled weights produces a clear rejection of the null hypothesis, while the

appropriately weighted analysis might not if one uses a rigid $p < .05$ cutoff criterion for rejection of the null hypothesis.[4]

Null effect in ANOVA.

To test the effects of condition on an analysis where the null hypothesis should be retained (no effect), an ANOVA was performed examining sex differences (F1SEX) in the importance of strong friendships (F1S40D). Using the "gold standard" of modeling the complex sample effects via AM, as Table 3 indicates, there should be no differences across groups.

The results in Table 3 are a good example of the risks associated with failing to appropriately model or approximate complex sampling weights and design effects. A researcher using only the original weights would conclude there are sex differences in the importance of strong friendships amongst high school students when in fact there are probably not. Again, SEs are substantially mis-estimated (again by a factor of 20 or so). Finally, there is again similarity between the third (scaled weights) and fourth condition (AM analysis) indicating that the approximation in this case yields similar results to the AM analysis.

Table 3: Null effect: sex differences in importance of strong friendships (F1S40D)

Analysis	Group	Mean	SE mean	t (df)	p <
SPSS- no weighting	Male Female	2.827 2.838	.0050 .0048	-1.67 (14539)	.095
SPSS- weight only	Male Female	2.822 2.833	.0003 .0003	-25.53 (3360675)	.0001
SPSS- weights scaled for N, DEFF	Male Female	2.822 2.833	.0077 .0075	-1.100 (6236)	.27
AM weight, PSU, Strata modeled	Male Female	2.822 2.833	.0060 .0060	-1.366 (386)	.17

[4] I do not espouse rigid cutoffs in quantitative analysis, as a probability of <.052 is not significantly different from a probability of <.049, but very different decisions would be made as a result. There is a well-developed literature around the failings of null hypothesis statistical testing that I encourage to you explore if you are interested in this issue.

Null effect in OLS regression.

In the final example, a multiple regression analysis predicted cumulative 9th-12th grade GPA (F1RGPP2) from school poverty (% students with free or reduced lunch; BY10FLP) controlling for dummy-coded race (based on F1RACE), and whether the school was public or private (BYSCTRL).

Table 4: Null effect: predicting student GPA from school poverty, controlling for race, school sector

Analysis	b	SE	t (df)	p <
SPSS- no weighting	-0.21	0.069	-2.98 (5916)	.003
SPSS- weight only	-0.01	0.005	-2.09 (1124550)	.04
SPSS- weights scaled for N, DEFF	-0.01	0.11	-0.09 (2078)	.93
AM weight, PSU, Strata modeled	-0.01	0.17	-0.058 (228)	.95

As Table 4 shows, in this case there is a stark contrast between appropriately modeled complex sampling and less ideal analyses. In this example, researchers using the unweighted sample or a weighted sample would make a Type I error, rejecting the null hypothesis and concluding there is a significant (albeit weak) relationship between school poverty and student GPA once other background variables were covaried. The last two conditions (scaled weights and AM modeling) produced similar and contrary results— that there is no relationship between these two variables when the sampling frame is approximated or modeled appropriately.

DISCUSSION

While this might seem an esoteric topic to many researchers in the social or health sciences, there is a wealth of compelling data freely available to researchers, and some researchers have found evidence that researchers do not always model the sampling frame appropriately (Johnson & Elliott, 1998). In brief, most modern statistical software can take complex sampling into account, either through using weights scaled for N and DEFF, or through using information such as primary and secondary sampling units (often called clusters) directly in the software. There is also, as mentioned above, free software that correctly models

complex samples, although it does have a small learning curve. Thus, there is little excuse for failing to take sampling into account when using these datasets.

In three of the four examples included above, there are potentially serious errors at risk if a researcher fails to take the sampling effects into account. In two of the four analyses, researchers would clearly make a Type I error, while in the logistic regression example it is less clear but still troubling.

Further, most of the analyses highlight how unweighted samples can mis-estimate not only parameter estimates, but also standard errors. This is because the unweighted sample is *not* representative of the population as a whole, and contains many eccentricities such as oversampling of populations of interest and perhaps nonrandom dropout patterns. Weighting provides a better parameter estimate, but unless further measures are taken, serious errors can occur in hypothesis testing and drawing of conclusions. Thus, while it requires extra effort to appropriately model the complex samples in these data sets, it is a necessary step to have confidence in the results arising from the analyses.

REFERENCES

Bozick, R., Lauff, E., & Wirt, J. (2007). Education Longitudinal Study of 2002 (ELS: 2002): A first look at the initial postsecondary experiences of the sophomore class of 2002 (NCES 2008-308). *Washington, DC: National Center for Education Statistics, Institute of Education Sciences, US Department of Education. Retrieved June, 10*, 2008.

Ingels, S. (1994). *National Education Longitudinal Study of 1988: second follow-up: student component data file user's manual*: US Dept. of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Johnson, D. R., & Elliott, L. A. (1998). Sampling Design Effects: Do They Affect the Analyses of Data from the National Survey of Families and Households? *Journal of Marriage and Family, 60*(4), 993-1001.

Kish, L. (1965). Selection techniques for rare traits. *Genetics and the epidemiology of chronic diseases*.

Lehtonen, R., & Pahkinen, E. (2004). *Practical methods for design and analysis of complex surveys*: Wiley.

Osborne, J. W. (1995). Academics, self-esteem, and race: A look at the assumptions underlying the Disidentification hypothesis. *Personality and Social Psychology Bulletin, 21*(5), 449-455.

Osborne, J. W. (1997). Race and academic disidentification. *Journal of Educational Psychology, 89*(4), 728-735.

Osborne, J. W. (2000). Advantages of Hierarchical Linear Modeling. *Practical Assessment, Research & Evaluation, 7*(1).

Osborne, J. W. (2008). A brief introduction to Hierarchical Linear Modeling. In J. W. Osborne (Ed.), *Best Practices in Quantitative Methods*. Thousand Oaks, CA: Sage.

Endnotes

i. Available through the NCES website (http://nces.ed.gov/) or the ICPSR web site (http://www.icpsr.umich.edu)
ii. Available through the CDC website (http://www.cdc.gov/nchs/index.htm
iii. Available through the BJS website (http://bjs.ojp.usdoj.gov/index.cfm?ty=dctp&tid=3)
iv. In many data sets there are multiple options for weights. For example, in NELS 88, a survey of 8th grade students who were then followed for many years, there is a weight only for individuals interested in using the first (BY) data collection. There is a similar weight for each other data collection point (F1, F2, F3, etc.). Yet not all students present in BY are also present in F1 and F2, so if I want to perform an analysis following students from 8th grade to 10th and 12th grade, there is also a weight (called a *panel* weight) for longitudinal analyses. This highlights the importance of being thoroughly familiar with the details of the user manual before using data from one of these studies.
v. Available from http://am.air.org/

Citation:

Osborne, Jason (2011). Best Practices in Using Large, Complex Samples: The Importance of Using Appropriate Weights and Design Effect Compensation. *Practical Assessment, Research & Evaluation*, 16(12). Available online: http://pareonline.net/getvn.asp?v=16&n=12.

Acknowledgement

The author thanks John Wirt from the National Center for Educational Statistics and David Miller from the American Institute for Research for their mentorship on this issue. Further acknowledgements to a mentor from too long ago at the University of Buffalo, Robert Nichols, who helped him learn to love working with complex data sets.

Author:

Jason W. Osborne
Educational Foundations and Leadership,
Darden College of Education
Old Dominion University
Norfolk, VA, 23529
jxosborn [at] odu.edu