Incorporating complex sample design effects when only final survey weights are available

Brady T. West
Survey Research Center
Institute for Social Research
University of Michigan, Ann Arbor
Ann Arbor, MI
bwest@umich.edu

Sean Esteban McCabe
Institute for Research on Women and Gender
Substance Abuse Research Center
University of Michigan, Ann Arbor
Ann Arbor, MI
plius@umich.edu

Abstract. In this article, we consider the situation that arises when a survey data producer has collected data from a sample with a complex design (possibly featuring stratification of the population, cluster sampling, and unequal probabilities of selection) and for various reasons only provides secondary analysts of those survey data with a final survey weight for each respondent and "average" design effects for survey estimates computed from the data. In general, these "average" design effects, presumably computed by the data producer in a way that fully accounts for all the complex sampling features, already incorporate possible increases in sampling variance due to the use of the survey weights in estimation. The secondary analyst of the survey data—who uses the provided information to compute weighted estimates; computes design-based standard errors reflecting variance in the weights (by using Taylor series linearization, for example); and inflates the estimated variances using the "average" design effects provided—is applying a "double" adjustment to the standard errors for the effect of weighting on the variance estimates, leading to overly conservative inferences. We propose a simple method to prevent this problem and provide a Stata program for applying appropriate adjustments to variance estimates in this situation. We illustrate two applications of the method with survey data from the Monitoring the Future study and conclude with suggested directions for future research in this area.

Keywords: st0001, deft2corr, survey design effects, survey weights, average design effects

1 Background

Standard practice in the design-based analysis of complex sample survey data requires data analysts to identify variables containing final survey weights, possibly compensating for unequal probabilities of selection, nonresponse adjustments, and poststratification adjustments. Data analysts must also identify 1) variables identifying first-stage sampling error strata and first-stage sampling error computation units ("ultimate clusters") or 2) variables containing replicate survey weights, enabling the use of replicated variance estimation procedures (Heeringa, West, and Berglund 2010, chap. 4). In both cases, the final survey weights enable computation of unbiased estimates of descriptive parameters and regression parameters in finite populations. In the former case, one estimates the sampling variance of a parameter estimate by using a first-order Taylor series

approximation—approximating the parameter estimate as a linear function of weighted sample totals and then computing the variance of this approximation. This is known as Taylor series linearization (TSL). This variance estimation method introduces a slight positive bias in variance estimates and slightly conservative inferences about the population of interest; see Wolter (2007) for more details. Replicated variance estimation methods, such as jackknife repeated replication (JRR) and balanced repeated replication (BRR), are also possible when these design codes are available. In the latter case, when only the final survey weight and replicate survey weights are available, JRR and BRR can be used to estimate variances. Asymptotically, TSL, JRR, and BRR converge to very similar variance estimates for most parameter estimates (Rao and Wu 1985).

Stata software provides data analysts with several easy-to-use tools implementing these analysis procedures, including `svyset` (for identifying complex design features) and `svydescribe` (for simple descriptive analyses of the identified design features). Stata also allows analysts to insert the `svy` modifier before a wide variety of descriptive (for example, `mean`) and model-based (for example, `regress`) commands to implement appropriate design-based analyses. Unfortunately, not all public-use survey datasets contain all the necessary variables enabling the alternative variance estimation procedures. Stratum and sampling error computation unit codes are often excluded from public-use survey datasets to maintain respondent confidentiality and limit disclosure risk (Lu and Sitter 2008). Some datasets may also exclude replicate weights, providing data analysts with only the final survey weight (for example, see Inter-University Consortium for Political and Social Research [2008]). In this case, the data producer needs to provide the data user with design effects for a wide variety of key parameters (for example, see Johnston et al. [2011, 532–552]), which the data user can then use to adjust variance estimates and associated confidence intervals, correctly accounting for complex sampling features (for example, see Thomas, Heck, and Bauer [2005]).

In general, following the notation used by Park et al. (2003), the design effect for a given parameter estimate $\widehat{\theta}$ is defined as follows:

$$\text{Deft}^2\left(\widehat{\theta}\right) = \frac{\text{Var}(\widehat{\theta})_{\text{CD}}}{\text{Var}(\widehat{\theta'})_{\text{SRSWR}}} = \text{Deft}^2_{\text{strat}} \times \text{Deft}^2_{\text{clust}} \times \text{Deft}^2_{\text{weights}} \qquad (1)$$
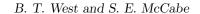
In (1), $\widehat{\theta'}$ refers to an estimate of the parameter ignoring the weights and assuming simple random sampling with replacement (SRSWR). The total design effect $\text{Deft}^2(\widehat{\theta})$, which is estimated in practice, accounts for the multiplicative change in the variance of an estimate under SRSWR due to complex sample design (CD) features. These include without-replacement selection, stratified cluster sampling (generally resulting in an increase in the variance due to cluster sampling, that is, $\text{Deft}^2_{\text{clust}} > 1$, and a decrease in the variance due to stratified sampling, that is, $\text{Deft}^2_{\text{strat}} < 1$), and the use of weights in estimation (generally resulting in an increase in the variance, that is, $\text{Deft}^2_{\text{weights}} > 1$; see Heeringa, West, and Berglund [2010, sec. 2.5]).

Previous work has shown that the total design effect for an estimated mean is an approximate function of complex interactions between the various sample design fea-

tures, the relationship of the variable of interest with the sampling weights, and the distribution of the variable of interest (Park and Lee 2004). For practical purposes, we write the total design effect in (1) as the simple product of three separate design effects due to each complex sampling feature, as suggested by Park et al. (2003). Importantly, this result only holds if the survey variable of interest and the survey weights are independent (Park and Lee 2004). Under this assumption, if a data producer only provides a data user with a final survey weight and average total design effects, the data user can compute weighted estimates and linearized variance estimates based on the weights (introducing $\text{Deft}^2_{\text{weights}}$) and then further adjust the estimated variances to incorporate additional design effects due to stratified cluster sampling.

Because total design effects include effects on the variance due to all the elements of complex sampling (weighting, stratification, and clustering), one does not want to "doubly adjust" for the effects of weighting if a linearized variance estimator incorporating the weights has already been used. For example, suppose that a data producer provides the public with a survey weight in a data file, possibly incorporating compensations for unequal probability of selection, differential nonresponse, and poststratification. The data producer also provides the data user with average design effects for many estimates of interest, several of which suggest that $\text{Deft}^2(\hat{\theta}) = 2.0$. Importantly, this average design effect of 2.0 already includes multiplicative increases in the variance due to weighting (as computed by the data producer), as shown in (1). The data user then computes weighted estimates and linearized variance estimates for those weighted estimates (which incorporate increases in variance due to variability in the weights). The data user proceeds to multiply the estimated variance by the average design effect of 2.0 (which already includes the increase in variance due to weighting), as instructed by the data producer. The net result is an unnecessary inflation of the variance of the estimate and overly conservative inferences.

A simple adjustment procedure based on the approximation in (1) can be used if the data user is provided only with a final survey weight and an average total design effect for a number of key statistics. After estimating a parameter and using TSL to estimate the variance of the parameter estimate, one can estimate the design effect due to the use of weights (which we denote as $\text{Deft}^2_{\text{weights}}$) by using the `estat effects` command in Stata. The average design effect provided by the data producer $(\overline{\text{Deft}^2})$ can then be divided by the design effect due to weighting to "extract" the approximate portion of the overall average design effect due to stratified cluster sampling:

$$\frac{\overline{\text{Deft}^2}}{\text{Deft}^2_{\text{weights}}} \approx \text{Deft}^2_{\text{strat}} \times \text{Deft}^2_{\text{clust}}$$

This extracted portion of the overall design effect can then be used to adjust estimated variances and corresponding test statistics and confidence intervals. If degrees of freedom based on the complex sample are also provided by the data producer, these can be incorporated into the adjustments as well. If not, large-sample critical values based on the standard normal distribution can be used to compute p-values for standard test statistics and to construct confidence intervals.

The following Stata ado-file defines a simple command that takes five inputs: 1) a weighted parameter estimate; 2) the linearized estimate of the standard error; 3) an average design effect; 4) a design effect due to weighting; and 5) an indicator of whether exponentiated forms of parameter estimates are desired. The ado-file enables this type of design effect adjustment.

```
program define deft2corr
    di ""
    if `5´ == 1 {
        di "Exponentiated Estimate: " exp(`1´)
        di "95% CI LL: " exp(`1´ - 1.96*`2´*(sqrt(`3´)/sqrt(`4´)))
        di "95% CI LL: " exp(`1´ + 1.96*`2´*(sqrt(`3´)/sqrt(`4´)))
    }
    if `5´ == 0 {
        di "Estimate: " `1´
        di "95% CI LL: " `1´ - 1.96*`2´*(sqrt(`3´)/sqrt(`4´))
        di "95% CI LL: " `1´ + 1.96*`2´*(sqrt(`3´)/sqrt(`4´))
    }
    di "Z statistic: " `1´ / (`2´*(sqrt(`3´)/sqrt(`4´)))
    if `1´ / (`2´*(sqrt(`3´)/sqrt(`4´))) > 0 di "p-value: " 2*(1-
        normal(`1´ / (`2´*(sqrt(`3´)/sqrt(`4´)))))
    if `1´ / (`2´*(sqrt(`3´)/sqrt(`4´))) <= 0 di "p-value: "
        2*(normal(`1´ / (`2´*(sqrt(`3´)/sqrt(`4´)))))
end
```

We illustrate the use of the `deft2corr` command by analyzing data from the Monitoring the Future (MTF) study (2007–2009).

2   Illustration

Nonmedical use of prescription opioids is a growing public health problem in the United States. Previous research focusing on young adults has found that more than 1 in every 10 lifetime nonmedical users of prescription opioids report intranasal administration (snorting) (McCabe et al. 2007). Furthermore, approximately 67% of intranasal users screened positive for drug abuse in the past year compared with approximately 6% of nonusers and 26% of nonmedical users who reported oral administration only. The use of prescription opioids via intranasal and other nonoral routes of administration is an extremely dangerous behavior that has been linked to a number of adverse physical consequences (Jewers et al. 2005; Watson et al. 2004; Yewell et al. 2002). Additionally, the rate of delivery of a drug to the brain directly correlates to the abuse potential of the drug, and intranasal, along with other nonoral routes of administration, delivers a drug to the brain at a much faster rate than oral administration (Kollins 2003; Roset et al. 2001).

In this example, we analyze survey data from the MTF study (2007–2009) and focus on two research objectives for 12th grade students in the United States. The first objective is to estimate the proportion of nonmedical users of prescription opioids among high school seniors (modal age 18) using intranasal administration. The second objective is to use logistic regression modeling to estimate differences in the odds of intranasal administration between nonmedical users only, nonmedical users who began using non-

medically prior to medical use, and nonmedical users who began using medically before nonmedical use (adjusting for race and ethnicity, year, geographic region of the school, and metropolitan statistical area). The MTF study provides data users only with a final survey weight (the variable V5; Inter-University Consortium for Political and Social Research [2008]) and includes appendices of total design effects for a variety of estimates, enabling the computation of average total design effects (Johnston et al. 2011, app. C). For this illustration, we use an average MTF total design effect of 2.0 (Johnston et al. 2011).

Prior to running the analyses, we examined the critical assumption underlying the result for the total design effect in (1) and found that the correlation of the survey weight variable (V5) and the nasal administration indicator (V1615) was negligible ($r = 0.01$, $p = 0.82$). The following Stata commands implement the analyses and illustrate the use of the `deft2corr` command:

```
* Set up survey weight and linearized variance estimation (default).
svyset [pweight = V5]

* Estimate the proportion of nonmedical users (subpop) using nasal
* admin (V1615).
svy, subpop(if usehistory == 3 | usehistory == 4 | usehistory == 5): prop V1615

* Request DEFT2 due to weighting (only).
estat effects

* Apply deft2corr, inputting the estimate, linearized standard error, average
* total DEFT2, and DEFT2 component due to weights.
deft2corr 0.3627 0.0257 2 1.3528 0

* Fit logistic regression model for odds of nasal admin.
svy, subpop(if usehistory == 3 | usehistory == 4 | usehistory == 5): ///
 logit V1615 ib3.usehistory i.V1151 i.V1 i.V13 i.V17

* Request DEFT2 estimates due to weighting (only).
estat effects

* Apply deft2corr for each parameter estimate, requesting
* exponentiated estimates (adjusted odds ratios).
deft2corr 1.7938 0.5310 2 1.4987 1
deft2corr 2.0390 0.5172 2 1.5098 1
```

Table 1 presents results from running the commands above (columns 2 and 5), along with the resulting 95% confidence intervals when only using the weights without any adjustment for stratification and clustering (column 3) and the resulting 95% confidence intervals when applying the average total design effect on top of the linearized standard error already incorporating the weights (the "double adjustment", column 4).

Table 1. Comparisons of design-based 95% confidence intervals for selected parameters using alternative design effect adjustments (source: MTF 2007–2009)

| Parameter | Weighted estimate | 95% confidence intervals: | | |
		Without adjustment for stratification and clustering	With "double adjustment" for weighting	With correct adjustment
Proportion: Nasal administration	0.3627	[0.3124, 0.4130]	[0.2915, 0.4339]	[0.3015, 0.4239]
AOR: NMPM versus MPNM	6.0120	[2.1230, 17.0245]**	[1.3798, 26.1973]*	[1.8067, 20.0071]**
AOR: NMO versus MPNM	7.6826	[2.7874, 21.1746]***	[1.8320, 32.2206]**	[2.3923, 24.6734]***

Notes: AOR = Adjusted odds ratio.
NMPM = Nonmedical use prior to medical use.
MPNM = Medical use prior to nonmedical use.
NMO = Nonmedical use only.
t test of null hypothesis that regression parameter is equal to 0: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

An estimated 36.27% of nonmedical users of prescription opioids in the 12th grade (modal age 18) during 2007–2009 administered the opioids intranasally. Nonmedical users who used nonmedically prior to any medical use and nonmedical users only (with no medical use ever) had 6.01 and 7.68 times greater odds of intranasal administration than nonmedical users who used medically first, holding the other covariates fixed.

The results in table 1 show how the correct design effect adjustment (last column) produces 95% confidence intervals for the parameters of interest with lower and upper limits that lie between the overly conservative limits computed using the "double" adjustment for the effect of weighting and the overly liberal limits computed without adjusting for the effects of stratification and clustering. In the cases of the two adjusted odds ratios, the level of significance actually varies depending on the adjustment used. This suggests that the overly conservative "double" adjustment could impact tests of significance in other cases and contexts.

Notably, the average total design effect provided by a data producer (2.0 in this case) is essentially arbitrary and can have a big impact on inferences. For this reason, when using these adjustment methods, data users must obtain an appropriate average total design effect for the subpopulation and the estimates in which they are interested. We acknowledge that more complex types of corrections may be needed if the provided survey weights are informative or strongly related to the survey variables of interest. Future research on this scenario needs to consider whether the results of Park and Lee (2004) can be applied to enable data users who are only provided with survey weights and average design effects to correctly compute estimated standard errors that fully reflect all the complex features of a given sample design.

3   Acknowledgments

4   References

Heeringa, S. G., B. T. West, and P. A. Berglund. 2010. *Applied Survey Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Inter-University Consortium for Political and Social Research. 2008. Monitoring the future: A continuing study of American youth (12th-grade survey), 2007. http://www.icpsr.umich.edu/icpsrweb/SAMHDA/studies/22480.

Jewers, W. M., Y. B. Rawal, C. M. Allen, J. R. Kalmar, E. Fox, G. E. Chacon, and P. P. Sedghizadeh. 2005. Palatal perforation associated with intranasal prescription narcotic abuse. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology* 99: 594–597.

Johnston, L. D., P. M. O'Malley, J. G. Bachman, and J. E. Schulenberg. 2011. *Monitoring the Future National Survey Results on Drug Use, 1975–2010. Volume I: Secondary School Students*. Ann Arbor, MI: University of Michigan.

Kollins, S. H. 2003. Comparing the abuse potential of methylphenidate versus other stimulants: A review of available evidence and relevance to the ADHD patient. *Journal of Clinical Psychiatry* 64: 14–18.

Lu, W. W., and R. R. Sitter. 2008. Disclosure risk and replication-based variance estimation. *Statistica Sinica* 18: 1669–1687.

McCabe, S. E., J. A. Cranford, C. J. Boyd, and C. J. Teter. 2007. Motives, diversion and routes of administration associated with nonmedical use of prescription opioids. *Addictive Behaviors* 32: 562–575.

Park, I., and H. Lee. 2004. Design effects for the weighted mean and total estimators under complex survey sampling. *Survey Methodology* 30: 183–193.

Park, I., M. Winglee, J. Clark, K. Rust, A. Sedlak, and D. Morganstein. 2003. Design effects and survey planning. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 3179–3186. Alexandria, VA: American Statistical Association.

Rao, J. N. K., and C. F. J. Wu. 1985. Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association* 80: 620–630.

Roset, P. N., M. Farré, R. de la Torre, M. Mas, E. Menoyo, C. Hernández, and J. Camí. 2001. Modulation of rate of onset and intensity of drug effects reduces abuse potential in healthy males. *Drug and Alcohol Dependence* 64: 285–298.

Thomas, S. L., R. H. Heck, and K. W. Bauer. 2005. Weighting and adjusting for design effects in secondary data analyses. *New Directions for Institutional Research* 2005: 51–72.

Watson, W. A., T. L. Litovitz, W. Klein-Schwartz, G. C. Rodgers, Jr., J. Youniss, N. Reid, W. G. Rouse, R. S. Rembert, and D. Borys. 2004. 2003 annual report of the American Association of Poison Control Centers Toxic Exposure Surveillance System. *American Journal of Emergency Medicine* 22: 335–404.

Wolter, K. M. 2007. *Introduction to Variance Estimation.* 2nd ed. New York: Springer.

Yewell, J., R. Haydon, S. Archer, and J. M. Manaligod. 2002. Complications of intranasal prescription narcotic abuse. *Annals of Otology, Rhinology and Laryngology* 111: 174–177.

About the authors

Brady T. West, M.A., PhD, is a research assistant professor in the Survey Methodology Program, located within the Survey Research Center at the Institute for Social Research on the University of Michigan–Ann Arbor campus. He also serves as a statistical consultant at the University of Michigan Center for Statistical Consultation and Research. As a doctoral student, he won the Edward C. Bryant Scholarship for outstanding academic achievement in survey statistics (sponsored by Westat and the American Statistical Association). His current research interests include the implications of measurement error in auxiliary variables and survey paradata for survey estimation, survey nonresponse, interviewer variance, analysis of complex sample survey data, and multilevel regression models for clustered and longitudinal data. He is the lead author of *Linear Mixed Models: A Practical Guide using Statistical Software* (Boca Raton, FL: Chapman & Hall/CRC, 2007) and a coauthor (with Steven Heeringa and Pat Berglund) of *Applied Survey Data Analysis* (Boca Raton, FL: Chapman & Hall/CRC, 2010).

Sean Esteban McCabe, M.A., M.S.W., PhD, is a research associate professor at the University of Michigan Substance Abuse Research Center and the Institute for Research on Women and Gender. He is an internationally recognized scholar in the areas of web-based data collection, prescription drug misuse, gender, sexual orientation, and epidemiology of substance abuse. He has served as the director of the Substance Abuse Research Center and the Office of Student Conflict Resolution at the University of Michigan. He was recently the recipient of the University of Michigan Research Faculty Recognition Award, and he has been the recipient of two NIH Junior Investigator Awards. He reviews manuscripts for 24 health and substance abuse journals, and he reviews grant applications for several organizations, including the National Institutes of Health and the Department of Education. He has served as principal investigator of eight NIH-funded projects in the past five years, as well as a participating investigator and faculty mentor on a number of NIH-funded projects. He has authored over 90 peer-reviewed articles.