Crime & Delinquency
2014, Vol. 60(2) 306–325

DOI: 10.1177/0011128714522114
cad.sagepub.com
SAGE

Adjusting for Design Effects in Disproportionate Stratified Sampling Designs Through Weighting

Paul E. Tracy[1] and Danielle Marie Carkin[1]

Abstract

This article validates the necessity of adjusting for the design effects in disproportionate stratified sampling designs through the use of sample weights. Using data from the 1958 Birth Cohort study, we demonstrate that complex sampling designs introduce sampling error and even sampling bias into sample data. Such sample data are a poor representation of population parameters. These design effects can be addressed through the application of sample weights.

Keywords

sampling bias, sampling error, sampling weights, stratified sampling

Introduction

Over the past year, numerous authors have submitted manuscripts to *Crime & Delinquency* based on secondary analysis of two noteworthy data sets: (a) National Longitudinal Study of Adolescent Health (Add Health) and

[1]University of Massachusetts Lowell, USA

Corresponding Author:
Paul E. Tracy, School of Criminology and Justice Studies, University of Massachusetts Lowell, 113 Wilder Street, Lowell, MA 01854-3060, USA.
Email: Dr.Paul.Tracy@gmail.com

(b) National Longitudinal Study of Youth 1979 (NLSY79). Each of these studies uses complex sampling designs with over-sampling of sub-populations. The purpose of this article is to address the inferential and statistical modeling problems that can arise when researchers fail to appreciate the implications of sampling design, and especially, when disproportionate sampling (i.e., over-sampling of certain population sub-groups) has been employed. In such instances, sample data must be weighted to remedy the design effects and/or possible selection effects due to disproportionate sampling.

According to Henry, one of the most common errors that occurs in research that uses disproportional sampling is that the analyst fails to weight the sample data to adjust population estimates owing to the design bias introduced into the study. This can often happen when analysts conducting secondary research are unfamiliar with the original sample design (Henry, 1990). Similarly, Fowler (1984) has noted that the effects of sampling design on sampling errors are often unappreciated; it is not uncommon to see reports of confidence intervals that assume simple random sampling when the design was clustered. Indeed, any statistical measure that relies on standard errors would be distorted, and perhaps substantially, by failing to appreciate the effects of stratified sampling (especially with disproportionate selection).

There is a basic feature of most social science research that is universal, or nearly so—The vast majority of data are collected from samples. The population may be unknown, or just inaccessible, or perhaps it is so large that it is either impractical or too expensive for the population to be enumerated and accessed. Thus, except for the rare occurrence when population data (or pseudo-population data as in a birth cohort) are used, researchers must confront the issue of sampling. Sampling design is thus the sine qua non of practically all criminological research. Despite the importance of sampling design, researchers often pay less attention to sampling issues and the implications thereof for external validity and statistically reliable population inferences, and devote considerably more attention instead to theoretical issues, operationalization of concepts, measurement, or the development and testing of more and more sophisticated quantitative analysis routines and statistical modeling. It is seldom the case that researchers are interested in sample findings for their own sake, but rather, the goal is to generalize the findings to a broader and more theoretically meaningful population.

Henry (1990) has noted that many researchers will refer to their samples as "representative" even though this is merely a subjective judgment rendered in the absence of objective criteria to sustain such judgment (p. 11). Clearly, external validity is an important issue that bears considerable attention (Campbell & Stanley, 1966; Cook & Campbell, 1979). Despite the

importance of sampling, researchers seemingly pay less attention to sampling design issues and their implications for statistical analysis, and ultimately, the internal and external validity of the findings.

The implications of sampling design for meaningful analysis of data and interpretation of findings take on manifold significance when researchers go beyond simple random sampling and employ complex sampling designs such as stratified or multistage designs. Increasingly, researchers have turned to complex sampling designs to ensure that theory will be optimally tested by providing a sufficient inclusion of substantively meaningful segments of the population, especially when certain demographic correlates are known to be strongly associated with the criterion measure. This is especially the situation in criminology where various sociodemographic factors are clearly related to the crime measures. Criminological research thus must attend to samples that reflect diversity across such measures as sex, race/ethnicity, age, social class, and so on. As a consequence, a great deal of primary criminological research is now utilizing stratified or multistage sampling designs, and many researchers are conducting secondary analyses of these data sets. A serious problem arises when a secondary analyst fails to appreciate that complex sampling designs bring about serious statistical issues that must be addressed when using samples to estimate population parameters.

Stratified Random Sampling

Stratified random sampling is an alternative sampling design to simple random or systematic random sampling procedures that are more commonly used. Stratified sampling involves a process of classifying the elements in the target population, so that a more representative sample may be achieved. The sampling frame is first divided into categories or strata of relatively homogeneous sub-populations that have theoretical or substantive meaning in the research. The researcher then draws independent random samples from each of the strata. Each of the sample strata itself thus represents a sampling frame and the selection of sample elements proceeds the same as in simple random sampling or systematic random sampling.

Stratified sampling, however, introduces a second-order issue into the process of drawing the sample. The selection process can be *proportionate stratification* in which a uniform sampling fraction is used such that if cases are selected for the sample in the same fraction as their representation in the population (e.g., if the four population strata show 20%, 25%, 40%, and 15%, then the cases across the four sample strata would comprise the same percentages). Alternatively, there are valid reasons why a *disproportionate stratification* design would be used to over-sample cases from one or more strata

(Cochran, 1977; Kish, 1992). Stratified random sampling has distinct research design and statistical advantages.

First, in terms of design concerns, a researcher should consider using a stratified sampling design when there may be a strong theoretical rationale that specific characteristics of the population may have an effect on the dependent variables being studied. In criminology, for example, a researcher would surely have a problem if certain characteristics such as sex, race/ethnicity, age, and social class were not reflected in the sample. A stratified sample ensures representation of these important correlates of the phenomenon of interest. Furthermore, unlike the chance occurrence in a simple random sample that there are enough cases across the categories of these important factors, a stratified sample gives the researcher the opportunity to control the exact number of these cases (Kalton, 1983). Moreover, the disproportionate stratified sampling technique ensures that sufficient sample cases are available for analysis when the population strata have comparatively low prevalence but yet are important for substantive purposes (Henry, 1990).

Fowler (1984) has provided a simple example that elucidates the issue convincingly. If a researcher used a simple or systematic sampling design, then a population sub-group that constituted 10% of the population would be expected to comprise by chance about 10% of the sample (subject to sampling error of course). If the researcher could only afford the time and financial resources to draw a survey sample of 1,000 cases, then the random sample would include 100 such cases from the population sub-group. If, however, the researcher knew from prior researcher that these 100 cases would likely be diluted as cross classification with other variables was used in the analysis, and that a sub-sample of 150 of the targeted cases was actually needed, then the total sample would need to be 50% bigger (i.e., 1,500 cases for the targeted 10% population group to show up in the sample 150 times). These additional 500 cases, which cost extra time and money, are only being sampled to yield the extra 50 cases of the target sub-group. There is a much easier way—Disproportionate stratified sampling will accomplish an over-sampling of the target sub-groups while still maintaining the original sample size of 1,000 cases. This particular use of disproportional sampling is advisable when sub-population analysis is needed, the sub-population sample size that results from simple random or proportional selection is too small, and the standard errors are consequently too high. Rather than increase the overall sample size, the size of individual strata can be adjusted.

Second, another design issue concerns the fact that in some research designs, sample estimates of the population are not just desirable, they are mandatory. This occurs when the sub-populations represent *domains of study* themselves. These domains of study have become very prevalent in criminology. It is now

well recognized that a researcher often must disaggregate sample data into relevant sub-categories to test separate statistical models, as, for example, with males and females, owing to the fact that sex is strongly related to the prevalence, incidence, and severity of crime. Failure to disaggregate and use pooled data will likely result in the male data overwhelming female data, thus precluding the development and testing of the relationships. Kalton (1983) has noted that when a population sub-group is small but still represents a valuable domain of study, then simple random sampling or even proportionate stratified sampling will yield too small a sample to produce sufficiently precise estimators. In this situation, the problem can be addressed through disproportionate stratification.

Third, stratified sampling designs also yield desirable statistical properties. The main statistical benefit from stratifying rather than taking a simple random sample is that a smaller number of cases can be drawn to achieve the same degree of accuracy (error rate). This is an important consideration because the smaller sample size required by stratified random sampling can reduce costs, sometimes substantially. Any random sampling design, by whatever method, is guided by concerns about sampling error. Sampling error can be reduced through drawing a larger sample or using homogeneous groupings or "strata" of a heterogeneous population. Henry (1990) has noted that stratification reduces error because there is always a gain from stratification in terms of precision or reduction of the standard error. According to Henry, the magnitude of the gain depends on two factors: (a) the variability between strata such that the greater the difference between the means of the strata and the overall means, the greater the gain and (b) because of the increased homogeneity within a stratum, the greater the gain. Kish (1965) in his classic work on sampling has explained this situation in terms of the variance of the dependent variable explained by the stratifying variable—The greater the explanatory power of the stratifying variable, the greater the gain from stratification. Henry has further explained that this benefit from disproportional stratification results from lowering the sampling variability in one stratum, where the standard deviation is relatively high, by increasing the number of sampling units allocated to that stratum thus minimizing within-groups variation and maximizing between-groups differences.

The Weighting Issue

As noted above, sample data from stratified sampling designs must be weighted for results to be valid. We provide here the cautions from the documentation of the Add Health and NLSY data sets.

Add Health. The Add Health study uses a highly complex cluster sampling design with disproportionate sampling and over-sampling of certain sub-populations (Harris, 2012). Chantala (2006) has noted that "the sampling plan used to collect the Add Health data has resulted in the Add Health sample differing from the target population of U.S. adolescents in ways that can influence analysis results" (p. 3). Chantala (2006) further explained,

> Unless appropriate adjustments are made for sample selection and participation, estimates from analyses using the Add Health data can be biased when any factor that influenced being a participant in the Add Health Study also influences the outcome of interest. For example, black adolescents whose parents were college graduates are one of the many over-sampled groups. Thus, parental education is a factor that affected participation of black youth in the Add Health study and can influence family income. Unless the analysis technique uses appropriate statistical methods to adjust for over sampling, estimates of the income of blacks will be biased. Any analysis that includes family income or other variables related to family income can also produce biased estimates unless proper adjustments are made for over sampling. (p. 4)

Chantala (2006) also cautioned that to obtain unbiased estimates, "it is important to account for the sampling design by using analytical methods designed to handle clustered data collected with unequal probability of selection" and that "failure to account for the sampling design usually leads to under-estimating standard errors and false-positive statistical test results" (pp. 4-5). More recently, Chantala and Tabor (2010) have indicated,

> The Add Health Study is a nationally representative, probability-based survey of adolescents in grades 7 through 12 conducted between 1994 and 1996. The sample design used to collect the data has introduced a complexity to analysis. ***Failing to account for this complexity may result in biased parameter estimates and incorrect variance estimates.*** Hence, you must correct for design effects and unequal probability of selection to ensure that your results are nationally representative with unbiased estimates. (p. i)

NLSY79. Similarly, the NLSY79 study is composed of a sample of 12,686 individuals, born between 1957 and 1964, who were aged 14 to 22 years when first interviewed in 1979. The total sample is comprised of the following three sub-samples: (a) a cross-sectional sample (n = 6,111) designed to represent the non-institutionalized civilian segment of young people living in the United States in 1979, (b) a military supplemental sample (n = 1,280), and (c) a supplemental sample (n = 5,295) designed to over-sample civilian Hispanic or Latino, African American, and economically disadvantaged, non-Black/non-Hispanic youths (Frankel, McWilliams, & Spencer, 1983). As was

the case with the Add Health study, the researchers (Moore, Pedlow, Krishnamurty, & Wolter, 2000) responsible for the design of the NLSY79 and NLSY97 studies have expressly noted the need to use sampling weights:

> Data from large-scale national samples typically need to be weighted to achieve an unbiased estimator of the population total. The weights are needed for four main reasons. First, the weights compensate for differences in the selection probabilities of individual cases, which often arise by design, as in the NLSY97/PAY97, where different overall sampling rates were required for Hispanics, non-Hispanic blacks, and others within the eligible age ranges. Second, weighting compensates for subgroup differences in participation rates; even if the sample as selected were representative of the larger population, differences in participation rates can compromise the representativeness of the sample . . . Third, weights compensate for random fluctuations from known population totals due to sampling. For instance, if one sex were overrepresented in the NLSY97 sample purely by chance, it would be possible to use data from the Decennial Census or the Current Population Survey to adjust for this departure from the population distribution. And fourth, adjusting the data to known population totals can help reduce the impact of survey undercoverage. (p. 32)

Despite the existence of substantial sampling literature on the necessity of weighting sample data, and despite the user documentation to the same effect, many researchers are seemingly unaware of the problems that can arise when the Add Health and NLSY data are not weighted. Below, we provide empirical results that demonstrate the significance of these problems.

Data and Method

1958 Birth Cohort Study

The 1958 Birth Cohort Study is an ideal data set to explore the issue of design effects and their remediation through weighting. The interview follow-up of the 1958 Birth Cohort Study concerned an investigation of the type, strength, duration, and sequencing of the relationships among selected developmental issues through the life course and alternative measures of criminality (i.e., both official and self-reported dimensions of juvenile delinquency and adult crime). The question of sampling design is always problematic and involves the kind of random sampling to be used (i.e., simple, systematic, stratified, or multistage) and a determination of the size of the sample. Sampling design was especially problematic in the 1958 cohort follow-up study owing to the following issues. First, the analysis of the delinquency careers in the full birth cohort indicated that certain demographic factors (sex, race/ethnicity, and socioeconomic status [SES, see Tracy, 1981) were definitively associated

with the criterion measures (Tracy, Wolfgang, & Figlio, 1990). Second, we were interested in the relationship of juvenile delinquency careers to subsequent adult criminal careers, and the correspondence between official measures of delinquency and crime to the hidden dimensions of criminal behavior that was never captured in the official data (Tracy & Kempf-Leonard, 1996).

Thus, to achieve these objectives, we needed a sampling design that guaranteed that cases would be sampled across these important characteristics of the cohort, so we chose to use a stratified random sampling scheme which yielded 26 sample strata produced from the combinations of sex, race/ethnicity, SES, number of juvenile offenses, and number of juvenile status offenses. Because we needed to ensure that high-risk cases were available in the sample in sufficient numbers to permit meaningful analyses, we implemented a disproportionate sample selection method. Thus, because the 1958 birth cohort is large, and because it is prohibitively expensive to conduct lengthy personal interviews with very large samples, we decided to employ a particular sampling scheme that would capture the most relevant background and juvenile offense characteristics of the cohort and yield a sample size sufficient for substantive analysis after the usual attrition in the sample owing to non-response.

Sample methodology. Table 1 displays the sample layout, the strata-specific cohort and sample sizes, and the selection probabilities. It should be noted that initially the sample was divided into two groups, males and females, but there were ultimately fewer sample strata for females (6 strata) compared with males (20 strata), owing to the fewer number of female offenders, and especially the two types of recidivists (non-chronic and chronic) among females in the cohort.

The 20 strata for males are as follows:

1. Race/ethnicity produces two groups of non-Whites and Whites.
2. SES (low and high) results in four strata.
3. Four-level juvenile delinquency status: no offenses, one offense, two to four offenses (non-chronic recidivist), and five or more offenses (chronic recidivist), yields 16 strata.
4. A fourth factor "status offenses" served to further classify the non-offenders from the delinquency status strata into two more strata: no offenses and no status offenses; and no offenses but one or more status offenses which yielded a total of 20 male strata.

The first three strata measures produced a set of 16 strata of race by SES by delinquency status. The four-level delinquency status variable measured

Table 1. Layout of Cohort and Sample Strata.

Strata No.	Sex	Race	SES	Delinquent offenses	Status offenses	Cohort size	Drawn sample	Selection probabilities
1	F	W	—	0	0	6,027	78	.01294
2	F	W	—	1	1+	527	78	.14800
3	F	W	—	2+	—	83	78	.93975
White female strata						6,637	234	.03525
4	F	NW	—	0	—	6,001	78	.01299
5	F	NW	—	1	1+	1,143	78	.06824
6	F	NW	—	2+	—	219	78	.35616
Non-White female strata						7,363	234	.03178
Total female strata						14,000	468	.03342
7	M	W	LO	0	0	881	78	.08853
8	M	W	LO	0	1+	51	51	1.00000
9	M	W	LO	1	—	176	78	.44318
10	M	W	LO	2-4	—	141	78	.55319
11	M	W	LO	5+	—	69	69	1.00000
White male low SES strata						1,318	354	.26858
12	M	W	HI	0	0	3,923	78	.01988
13	M	W	HI	0	1+	102	78	.76470
14	M	W	HI	1	—	500	78	.15600
15	M	W	HI	2-4	—	276	78	.28260
16	M	W	HI	5+	—	97	78	.80412
White male high SES strata						4,898	390	.07962
White male strata						6,216	744	.11969
17	M	NW	LO	0	0	2,830	78	.02756
18	M	NW	LO	0	1+	257	78	.30350
19	M	NW	LO	1	—	777	78	.10038
20	M	NW	LO	2-4	—	768	78	.10156
21	M	NW	LO	5+	—	464	78	.16810
Non-White male low SES strata						5,096	390	.07653
22	M	NW	HI	0	0	1,211	78	.06440
23	M	NW	HI	0	1+	94	78	.82978
24	M	NW	HI	1	—	244	78	.31967
25	M	NW	HI	2-4	—	198	78	.39393
26	M	NW	HI	5+	—	101	78	.77227
Non-White male high SES strata						1,848	390	.21103
Non-White male strata						6,944	780	.11232
Total male strata						13,160	1,524	.11580

Note. SES = socioeconomic status.

whether the subject had never committed a delinquent offense or had committed such an offense once, two to four times, or five or more times, and offense here refers to legitimate crimes and not status offenses, such as running away, truancy, or curfew violations. We measured delinquency status this way because we wanted to capture the issue of serious versus trivial violations of the law. Thus, a non-offender in this variable could have never committed an offense of any kind, *or* he could have committed one or more pure status offenses.

We had intended to have equal sizes in all sample strata, but for some strata we reached complete enumeration before reaching the desired cell size of 78 cases. Thus, there are two male strata with less than the required number of 78 cases. Table 3 shows the layout of the 20 strata for the male portion of the sample. There are a total of 1,524 males in the drawn sample, with 78 cases in 18 cells and two strata with fewer cases—51 cases in the White, low SES, non-offender, one or more status offenses stratum, and 69 cases in the White, low SES, five or more offenses stratum.

The six strata for females are as follows:

1. Stratifying on race/ethnicity produces two groups: non-Whites and Whites.
2. Introducing a revised measure of delinquency status which contains only three categories for females: no offenses, one offense, and two or more offenses, results in a total of six strata.
3. Like the case for males, we introduced the "status offense" measure, but for females, this resulted in no additional strata as we found no females who committed only one offense that was not a status offense.

For females, therefore, the three main strata for the two race groups are as follows: no offenses of any kind, one offense or more than one status offenses, and two or more offenses.

Results

Sampling Error: Design Effects

Whenever a sample is drawn, by definition, only that part of the population that is included in the sample is measured, and the sample cases are used to represent the entire population. Hence, there must always be some random error in the data, resulting from those members of the population who were not measured. This error will naturally be reduced as the sample size is increased, so that, if a census is performed (a 100% sample is a census), by

definition, there will be no sampling error. In sampling contexts, sampling error gives us some idea of the precision of our statistical estimate. A low sampling error means that we had relatively less variability or range in the sampling distribution. So how do we calculate sampling error? We base our calculation on the standard deviation of our sample. The greater the sample standard deviation, the greater the standard error (and the sampling error). The standard error is also related to the sample size. The greater the sample size, the smaller the standard error. Why? Because the greater the sample size, the closer the sample is to the actual population itself. If you draw a sample that consists of 50% of the population, this sample will have less error than a 25% sample. The population estimates derived from the 50% sample will be less susceptible to error than would be the case for the 25% sample. The larger the sample, the better the sample estimates will approximate the population parameters.

Tables 2 and 3 provide comparisons between the full cohort (n = 27,160) and the drawn sample (n = 1,992) concerning how well the sample *reflects or represents* the full cohort (the amount of sampling error). Table 3 provides a comparison of the full birth cohort with the drawn sample concerning the prevalence (i.e., proportion) of adult criminals. The results show that the sample estimate of adult criminals is a poor reflection of the true score in the full cohort. In the full cohort, there were 3,617 adult criminals (13.3%) as compared with 610 (30.6%) in the drawn sample. The drawn sample thus over-represents adult criminals by 17.3%. The last column of Table 2 shows the strata-specific sampling error. For most strata, the sampling error is small, but there are a few strata for which the difference between the cohort and the drawn sample is quite large (e.g., Strata 9, 10, and 18).

Similarly, Table 3 provides a comparison of the full birth cohort with the drawn sample concerning the incidence (i.e., mean number) of adult crimes. The results show that the sample estimate of adult crime incidence is a poor representation of the actual score in the full cohort. In the full cohort, there were 3,617 adult criminals who committed on average 2.5 crimes as compared with 610 criminals in the drawn sample who committed an average of 3.17 crimes. The drawn sample thus over-represents adult criminality by 0.67 offenses per criminal. The last column of Table 3 shows the strata-specific sampling error. For most strata, the sampling error is very small but there are a few strata for which the difference between the cohort and the drawn sample is large and quite problematic (e.g., Strata 1, 4, 5, 6, 12, and 20).

The drawn sample was a random sample of the full cohort. How could there have been so much sampling error? If we did not know the population scores (the cohort), then the drawn sample would yield very poor estimates of both the number and percentage of adult criminals in the cohort and the extent

Table 2. Sampling Error: Congruence of Cohort and Drawn Sample on Criterion Measure (Prevalence of Adult Criminals).

Strata		Cohort			Drawn sample			
No.	Category	Total cases	Adult offenders	% adult offenders	Sample size	Adult offenders	% offenders	Sampling error (%)
1	FW 0/0	6,027	73	1.21	78	0	0.00	−1.21
2	FW 1/1+	527	40	7.59	78	9	11.54	3.95
3	FW 2+	83	13	15.66	78	12	15.38	−0.28
4	FNW 0/0	6,001	231	3.85	78	5	6.41	2.56
5	FNW 0/1+	1,143	122	10.67	78	7	8.97	−1.70
6	FNW 2+	219	61	27.85	78	22	28.21	0.35
7	MWLO 0/0	881	143	16.23	78	11	14.10	−2.13
8	MWLO 0/1+	51	12	23.53	51	12	23.53	0.00
9	MWLO 1	176	55	31.25	78	18	23.08	−8.17
10	MWLO 2-4	141	71	50.35	78	32	41.03	−9.33
11	MWLO 5+	69	50	72.46	69	50	72.46	0.00
12	MWHI 0/0	3,923	381	9.71	78	5	6.41	−3.30
13	MWHI 0/1+	102	23	22.55	78	17	21.79	−0.75
14	MWHI 1	500	129	25.80	78	22	28.21	2.41
15	MWHI 2-4	276	123	44.57	78	37	47.44	2.87
16	MWHI 5+	97	59	60.82	78	48	61.54	0.71
17	MNWLO 0/0	2,830	561	19.82	78	18	23.08	3.25
18	MNWLO 0/1+	257	79	30.74	78	30	38.46	7.72
19	MNWLO 1	777	260	33.46	78	23	29.49	−3.97
20	MNWLO 2-4	768	368	47.92	78	36	46.15	−1.76
21	MNWLO 5+	464	308	66.38	78	53	67.95	1.57
22	MNWHI 0/0	1,211	187	15.44	78	9	11.54	−3.90
23	MNWHI 0/1+	94	27	28.72	78	22	28.21	−0.52
24	MNWHI 1	244	72	29.51	78	23	29.49	−0.02
25	MNWHI 2-4	198	104	52.53	78	41	52.56	0.04
26	MNWHI 5+	101	65	64.36	78	48	61.54	−2.82
All		27,160	3,617	13.32	1,992	610	30.62	17.31

of the criminality of these cohort criminals. Our sample data do not *represent* the cohort very well at all. Any inferences we make about the cohort from the sample data would be erroneous.

This situation arises from what is called a *design effect*. We used a disproportionate stratified sampling design that had unequal probabilities of case selection. We did this to ensure that there would be sufficient cases for analysis for each of the 26 strata. Is there anything that we can do now to remedy this design effect? Yes, we need *sample weights* to produce estimates of population statistics that would have been obtained if proportionate sampling had been used. Thus, we need to adjust for differential probabilities of selection used in the sampling process. When sample units are not chosen under an Equal Probability Selection Method (EPSEM), unbiased estimates of population

Table 3. Sampling Error: Congruence of Cohort and Drawn Sample on Criterion Measure (Mean Number of Adult Offenses).

Strata		Cohort			Drawn sample			
No.	Category	Cases	*M*	*SD*	Cases	*M*	*SD*	Sampling error
1	FW 0/0	73	1.34	0.92	0	0	0.00	−1.34
2	FW 1/1+	40	1.53	0.96	9	1.22	0.44	−0.30
3	FW 2+	13	2.69	3.04	12	2.50	3.09	−0.19
4	FNW 0/0	231	1.42	1.11	5	3.40	4.34	1.98
5	FNW 0/1+	122	1.52	1.39	7	1.00	0.00	−0.52
6	FNW 2+	61	3.28	5.68	22	4.41	7.90	1.13
7	MWLO 0/0	143	2.22	2.64	11	2.82	4.77	0.59
8	MWLO 0/1+	12	1.92	1.51	12	1.92	1.51	0.00
9	MWLO 1	55	2.31	2.52	18	2.00	1.64	−0.31
10	MWLO 2-4	71	2.92	2.58	32	2.97	2.79	0.05
11	MWLO 5+	50	5.56	4.19	50	5.56	4.19	0.00
12	MWHI 0/0	381	1.66	1.31	5	1.00	2.06	−0.66
13	MWHI 0/1+	23	1.96	1.80	17	2.12	1.36	0.16
14	MWHI 1	129	1.90	1.92	22	1.68	1.45	0.22
15	MWHI 2-4	123	2.43	1.68	37	2.32	3.46	−0.11
16	MWHI 5+	59	3.39	3.34	48	3.44	1.43	0.05
17	MNWLO 0/0	561	2.22	2.15	18	1.94	1.43	−0.27
18	MNWLO 0/1+	79	2.13	1.85	30	2.27	2.18	0.14
19	MNWLO 1	260	2.42	2.22	23	2.43	1.47	0.01
20	MNWLO 2-4	368	3.02	2.39	36	2.33	1.51	−0.69
21	MNWLO 5+	308	4.39	3.29	53	4.42	2.98	0.02
22	MNWHI 0/0	187	2.01	1.93	9	2.56	2.19	0.54
23	MNWHI 0/1+	27	2.15	1.68	22	2.32	1.81	0.17
24	MNWHI 1	72	2.51	2.78	23	1.91	1.50	−0.60
25	MNWHI 2-4	104	3.21	3.88	41	3.76	5.44	0.54
26	MNWHI 5+	65	4.86	3.80	48	4.77	3.85	−0.09
All		3,617	2.5	2.58	610	3.17	3.48	0.67

parameters can be produced by inflating the sample cases by the reciprocal of the probability of selection to produce what are called *base weights*.

$$\text{Base Weights}: W_i = 1/P_i\left(n_i / N_i\right) \text{or } W_i = N_i / n_i,$$

where $P(i)$ is the strata-specific selection probability, i indexes the strata from 1 to 26, N_i is the total population in stratum i, and n_i are sample cases in stratum i.

Table 4. Sample Design Weights.

	Strata				
No.	Category	Cohort frequency	Drawn sample	Selection probabilities	Design weight[a]
1	FW 0/0	6027	78	.0129	77.2692
2	FW 1/1+	527	78	.1480	6.7564
3	FW 2+	83	78	.9398	1.0641
4	FNW 0/0	6001	78	.0130	76.9359
5	FNW 0/1+	1143	78	.0682	14.6538
6	FNW 2+	219	78	.3562	2.8077
7	MWLO 0/0	881	78	.0885	11.2949
8	MWLO 0/1+	51	51	1.0000	1.0000
9	MWLO 1	176	78	.4432	2.2564
10	MWLO 2-4	141	78	.5532	1.8077
11	MWLO 5+	69	69	1.0000	1.0000
12	MWHI 0/0	3923	78	.0199	50.2949
13	MWHI 0/1+	102	78	.7647	1.3077
14	MWHI 1	500	78	.1560	6.4103
15	MWHI 2-4	276	78	.2826	3.5385
16	MWHI 5+	97	78	.8041	1.2436
17	MNWLO 0/0	2830	78	.0276	36.2821
18	MNWLO 0/1+	257	78	.3035	3.2949
19	MNWLO 1	777	78	.1004	9.9615
20	MNWLO 2-4	768	78	.1016	9.8462
21	MNWLO 5+	464	78	.1681	5.9487
22	MNWHI 0/0	1211	78	.0644	15.5256
23	MNWHI 0/1+	94	78	.8298	1.2051
24	MNWHI 1	244	78	.3197	3.1282
25	MNWHI 2-4	198	78	.3939	2.5385
26	MNWHI 5+	101	78	.7723	1.2949

[a]Design weight = 1 / p_i or Ni / n_i.

Weighting the Sample

Table 4 provides a layout of the cohort, the drawn sample, and the strata-specific base weights. After weighting the sample cases by the base weights, we can compare the full cohort with the unweighted and weighted sample cases with respect to a variety of statistical measures that will give us an indication of *whether* and *how well* the weighted sample can be used to estimate the full cohort parameters. These common statistical measures are reported in Table 5.

Table 5. Sampling Design Effects and MSE: A Comparison of Unweighted and Weighted Data.

Descriptive statistics	Full cohort	Unweighted drawn sample	Weighted drawn sample
Adult offender status			
n	27,160	1,992	27,161
M	0.1332	0.3062	0.1314
SE	0.0021	0.0103	0.0020
Variance	0.1155	0.2126	0.1141
SD	0.3398	0.4610	0.3378
MSE[a]	—	0.2425	0.1155
Root mean square error	—	0.4924	0.3399
Adult crime incidence			
n	3,617	610	3,568
M	2.5040	3.1672	2.5934
SE	0.0428	0.1408	0.0480
Variance	6.6383	12.0902	8.2196
SD	2.5765	3.4771	2.8670
MSE[a]	—	12.5301	8.2276
Root mean square error	—	3.5398	2.8684

Note. MSE = mean square error.
[a] $\mathrm{MSE}(\bar{X}) = \mathrm{Variance}(x) + (\mathrm{Bias})^2$, where $\mathrm{Bias} = (\mu - \bar{X})$.

Table 5 provides descriptive statistics for the two criterion measures we have been examining: proportion of adult criminals (prevalence) and mean number of adult crimes (incidence). With respect to prevalence, we see, as above, that the unweighted drawn sample is a poor representation of the full cohort. The mean, variance, standard deviation, and standard error are all much higher in the unweighted drawn sample. After weighting, the sample data now are virtually identical to the full cohort scores for all the usual descriptive statistics (mean, variance, standard deviation, and standard error).

Table 5 also reports another statistic, mean square error (MSE; root mean square error [RMSE] which equals √MSE). MSE is a very important measure and is calculated as follows:

$$\mathrm{MSE}(\bar{X}) = \mathrm{Variance}(x) + (\mathrm{Bias})^2, \text{ where } \mathrm{Bias} = (\mu - \bar{X}).$$

Essentially, the MSE of any sample statistic, like the mean, inflates the variance of the estimate by the *bias* surrounding the sample statistic. In the

present case, bias is the difference between the cohort mean (μ) and the sample mean ($\bar{X}$). Thus, as the sample mean departs from the population mean, MSE becomes increasingly greater than the variance. Likewise, as RMSE is analogous to the standard deviation, as the sample standard deviation departs from the population standard deviation, RMSE becomes increasingly greater than the standard deviation.

The MSE scores in Table 5 clearly indicate that the unweighted drawn sample is characterized by sampling bias and sample selection bias in particular. The MSE for the unweighted sample (0.2425) is greater than its own variance (0.2126) and it is much greater than the MSE of the weighted sample (0.1155). Moreover, the MSE of the weighted sample (0.1155) is nearly identical to its own variance (0.1141) and is identical to the cohort variance (0.1155). These results definitely indicate that the sample weighting process eliminated the design effects caused by disproportional stratified sampling.

The Table 5 results with respect to the quantitative dependent variable (incidence of adult crime) yield an identical situation. The unweighted drawn sample is a poor representation of the full cohort. The mean, variance, standard deviation, and standard error are all much higher in the unweighted drawn sample. After weighting, the sample data now are much closer to the full cohort scores for all the usual descriptive statistics (mean, variance, standard deviation, and standard error). Likewise, the MSE results replicate the scenario that was obtained for the prevalence data. That is, the unweighted drawn sample is characterized by sampling bias and sample selection bias in particular. The MSE for the unweighted sample (12.5301) is greater than its own variance (12.0902) and it is much greater than the MSE of the weighted sample (8.2276). Moreover, the MSE of the weighted sample (8.2276) is nearly identical to its own variance (8.2196) and is close to the cohort variance (6.6383). As with prevalence, these incidence results definitely indicate that the sample weighting process eliminated the design effects caused by disproportional stratified sampling.

Ultimately, these descriptive statistical comparisons can only take us so far. The ultimate test of the prophylactic benefit of weighting the sample data remains to be tested using multivariate models in which the true population relationships and scores are compared with the sample estimates. These results are reported in Tables 6 and 7.

Table 6 shows a multiple logistic regression model that predicts adult crime status (i.e., prevalence; Yes vs. No) using four main effects (predictor variables that were found to be strongly associated with adult crime status; Tracy & Kempf-Leonard, 1996) and one interaction effect (Sex × Race/ethnicity) that was also found to have significant predictive power. For the full cohort, all of the predictor variables, sex, race/ethnicity, SES, and

Table 6. Logistic Regression of Adult Crime Status: Full Cohort and Drawn Sample.

	Full cohort			Drawn sample unweighted			Drawn sample weighted		
	Standardized coefficient	*p* value	Odds ratio	Standardized coefficient	*p* value	Odds ratio	Standardized coefficient	*p* value	Odds ratio
Sex	1.7663	.00000	5.8494	1.3259	.00000	3.7654	1.5567	.00000	4.7432
White	−0.2537	.00000	0.7759	−0.2063	.06239	0.8136	−0.5897	.00000	0.5545
SES	−0.1187	.00009	0.8881	−0.0673	.41019	0.9349	−0.1362	.00000	0.8727
One time	0.8356	.00000	2.3063	0.9197	.00000	2.5084	0.9224	.00000	2.5154
Recidivist	1.4768	.00000	4.3789	1.4960	.00000	4.4637	1.3855	.00000	3.9968
Chronic	2.1845	.00000	8.8862	2.3968	.00000	10.9884	2.0563	.00000	7.8169
SES × White	−0.2251	.00001	0.7984	−0.0929	.46030	0.9113	−0.1969	.00018	0.8212
Constant	−2.2560	.00000	0.1048	−1.8803	.00000	0.1525	−2.0256	.00000	0.1319

Note. SES = socioeconomic status.

Table 7. OLS of Adult Crime Incidence: Full Cohort and Drawn Sample.

	Full cohort			Drawn sample			Drawn sample weighted		
	Standardized coefficient	*T* score	*p* value	Standardized coefficient	*T* score	*p* value	Standardized coefficient	*T* score	*p* value
Constant		5.4987	.00000		0.4149	.67824		10.5115	.00000
Sex	0.1427	25.1262	.00000	0.0931	4.3919	.00001	0.0921	15.8126	.00000
White	−0.0308	−4.7444	.00000	−0.0334	−1.5972	.11039	−0.0980	−14.3964	.00000
SES	−0.0295	−4.5522	.00001	−0.0387	−1.8459	.06505	−0.0325	−4.7699	.00000
One time	0.0433	7.7120	.00000	0.0314	1.1873	.23527	0.0339	5.8978	.00000
Recidivist	0.1402	24.7308	.00000	0.1123	4.3300	.00002	0.1092	18.7895	.00000
Chronic	0.3378	59.4577	.00000	0.3830	14.9040	.00000	0.3100	53.1841	.00000

Note. OLS = ordinary least squares; SES = socioeconomic status.

delinquency status, had highly significant coefficients and strong predictive efficiency (odds ratios) in classifying cases as adult criminals. When we turn to the unweighted drawn sample, however, as we would expect from the discussion above, the model is not a good fit to the data. The results indicate that neither race/ethnicity nor SES nor the Sex × Race/ethnicity interaction effect is significant.

In the absence of having the true population scores, if one had used the unweighted drawn sample to estimate the population relationships, then a very distorted picture of adult crime status would have emerged. In a multivariate modeling situation, the sampling bias in the unweighted data grossly distorted which cohort characteristics were significantly associated with the criterion measure. Alternatively, when we examine the weighted drawn

sample, the results nicely replicate the full cohort results. Once again, all the predictor variables are significant and strongly associated with adult crime status, although the coefficients and odds ratios are slightly different owing to sampling error (not bias).

Table 7 shows an ordinary least squares (OLS) regression model that predicts the quantitative version of the criterion measure (incidence; number of adult crimes) using four main effects (sex, race/ethnicity, SES, and delinquency status) that were found to have significant predictive power. For the full cohort, all of these predictor variables had highly significant regression coefficients (standardized βs) in explaining the variation around the number of adult crimes.

When we examine the unweighted sample, however, the model is not a good fit to the data. The results indicate that neither race/ethnicity nor SES nor one-time offenders effect is significant. Like with the logistic regression models, if one had used the unweighted drawn sample to estimate the population relationships, then a very distorted picture of the extent of adult crime would have emerged. In a multivariate modeling situation, the sampling bias in the unweighted data grossly distorts which cohort characteristics are significantly associated with the criterion measure.

As was the case above, when we examine the weighted drawn sample, the results nicely replicate the full cohort results. Once again, all the predictor variables are significantly associated with adult crime status, although the coefficients are slightly different owing to sampling error (not bias).

Conclusion

The purpose of this article was to demonstrate that inferential and statistical modeling problems can arise when researchers fail to appreciate the implications of sampling design, and especially, when disproportionate stratified sampling is used in a study. Disproportionate sampling involves a purposeful over-sampling of certain population sub-groups. Increasingly, researchers have turned to complex sampling designs. Such designs ensure that theory can be optimally tested by providing a sufficient inclusion of substantively meaningful segments of the population. This is critically necessary when certain demographic correlates are known to be strongly associated with the criterion measure. We have demonstrated that serious inference problems arises when an analyst fails to appreciate that complex sampling designs bring about serious statistical issues that must be addressed when using samples to estimate population parameters.

First, we showed that a drawn sample with over-sampling of population sub-groups creates a design effect situation. We examined two different

criterion measures: (a) prevalence of adult crime and (b) incidence of adult crime. In both cases, the drawn sample was a very poor representation of the true scores in the full birth cohort. Owing to the design effect caused by oversampling, the drawn sample inflated the proportion of adult criminals and the mean number of adult offenses. The mean, variance, standard deviation, and standard were all much higher in the unweighted drawn sample. Moreover, the drawn sample was characterized by significant sampling bias. Had unweighted sample data been used, research would have been unable to generate valid findings generalizable to the full cohort.

Second, we weighted the sample data and demonstrated that the design weights eliminated the design effects and rendered the drawn weighted sample statistically equivalent to the full cohort for all descriptive statistics pertaining to the prevalence and incidence of adult crime.

Third, we also estimated multivariate models predicting dichotomous adult crime status and mean number of adult crimes. It was found that sampling bias in the unweighted data grossly distorted which cohort characteristics were significantly associated with the criterion measure. Alternatively, after weighting the data, when we examined the weighted drawn sample, the results nicely replicated the findings for the full cohort—The predictors once again reached statistical significance with similar effect sizes.

This article has thus demonstrated that complex sampling designs, especially disproportionate stratified sampling, are associated with significant design effects. In such instances, sample data must be weighted to remedy the design effects and/or possible selection effects due to disproportionate sampling.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Chantala, K. (2006, October). *Guidelines for analyzing add health data*. Carolina Population Center, University of North Carolina at Chapel Hill. Retrieved from http://www.cpc.unc.edu/projects/addhealth/data/guides/wt-guidelines.pdf

Chantala, K., & Tabor, J. (2010, August). *National longitudinal study of adolescent health: Strategies to perform a design-based analysis using the add health*

data. Carolina Population Center, University of North Carolina at Chapel Hill. Retrieved from http://www.cpc.unc.edu/projects/addhealth/data/guides/weight1.pdf

Cochran, W. G. (1977). *Sampling techniques*. New York, NY: John Wiley.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago, IL: Rand McNally.

Fowler, F. J. (1984). *Survey research methods*. Newbury Park, CA: Sage.

Frankel, M. R., McWilliams, H. A., & Spencer, B. D. (1983). Carolina Population Center, University of North Carolina at Chapel Hill. *NLSY79: Technical sampling report*.

Harris, K. M. (2012). *The add health study: Design and accomplishments*. Carolina Population Center, University of North Carolina at Chapel Hill. Retrieved from http://www.cpc.unc.edu/projects/addhealth/data/guides/DesignPaperWIIV.pdf

Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage.

Kalton, G. (1983). *Introduction to survey sampling*. Newbury Park, CA: Sage.

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

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*, *8*, 183-200.

Moore, W., Pedlow, S., Krishnamurty, P., & Wolter, K. (2000). *National Longitudinal Survey of Youth 1997 (NLSY97): Technical Sampling Report*. Chicago, IL: National Opinion Research Center.

Tracy, P. E. (1981). *Ecology and delinquency: The development of a composite measure of social class*. Philadelphia: Center for Studies in Criminology and Criminal Law, Wharton School, University of Pennsylvania.

Tracy, P. E., Wolfgang, M. E., & Figlio, R. M. (1990). *Delinquency careers in two birth cohorts*. New York, NY: Plenum Press.

Author Biographies

Paul E. Tracy is professor of criminology and director of graduate studies, School of Criminology and justice studies at the University of Massachusetts Lowell. His research interests concern criminal careers and the effectiveness of criminal sanctions.

Danielle Marie Carkin is a doctoral student in criminology at the School of Justice Studies at the University of Massachusetts Lowell. Her research interests concern criminal careers, crime over the life course, and desistance and persistence in offending