CONSEQUENCES OF IMPROPER ANALYSIS OF PROBABILISTICALLY SAMPLED DATA AND SUGGESTIONS FOR AN ALTERNATE METHOD OF ANALYSIS

by

JOHN G. BELL

JAMES E. MCLEAN, COMMITTEE CHAIR

J. BRIAN GRAY
RICK A. HOUSER
STEPHEN J. THOMA
SARA TOMEK

A DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in the
Department of Educational Studies in
Psychology, Research Methodology and
Counseling in the Graduate School
of The University of Alabama

TUSCALOOSA, ALABAMA

2011

ABSTRACT

This study reviewed the purpose and practice of weighting, particularly as regards disproportionately sampled populations. Weighted data cannot be simply unweighted by multiplying by the reciprocal of the weights. Doing so produces erroneous results. When weights are used for the sole purpose of correcting bias due to oversampling the minority portion of a population containing a minority and a majority, proper analysis can be performed without the weights. The population proportion of a disproportionate sample can be restored by replicating the majority and minority portions of the data. These replications can then be concatenated to produce a dataset that may be used under the assumptions of simple random sampling.

ACKNOWLEDGMENTS

I would like to express my appreciation to the members of my committee, Drs. J. Brian Gray, Rick A. Houser, Stephen J. Thoma, and Sara Tomek. Special gratitude and appreciation goes to my chairman, Dr. James E. McLean, who patiently advised and supported me from this project's inception to its end.

Thanks, also, to my wife, Anne, and to my children, Maggie, Hannah, Duren, and Pierce. Their support was unwavering and invaluable.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I:

INTRODUCTION

A common problem in research practice is the scarcity of data for some members of a population. In some cases, the very members of a scarce population are the ones of interest to the researcher. Entire journals are devoted to research on members of minority populations (e.g., *Diaspora, Indigenous, and Minority Education; Cultural Diversity and Ethnic Minority Psycholog; The Journal of Educational Issues of Language Minority Students;* and *The Journal of Negro Education*.)

In the course of simple random sampling (SRS,) members of minority groups will, naturally, be less frequently chosen for inclusion in the sample, thus decreasing the sample size for any given member of the minority population. If the sample size for the minority group of interest is too small for research purposes, then research on that minority group suffers because of the dependence of many statistical methods on sufficient sample size.

In order to remedy this problem, samples are sometimes designed so that minority members are deliberately oversampled. That is, the sample is taken in such a way as to increase the probability that a minority member will be chosen for inclusion. This serves to increase the sample size for the minority group, but creates a new problem in that the minority now has proportionately greater representation in the designed sample than is true of the greater population. Crude analysis using the results of such a probabilistic sample without considering the sample design will result in biases toward the oversampled (minority) group as they are proportionately overrepresented in the sample.

Statement of the Problem

The National Center for Education Statistics (NCES), organized under the U.S. Department of Education (DOE), collects and maintains data useful to education researchers. Though extensive, these data are collected using probabilistic sampling rather than simple random samples (SRS.) NHES uses a sample design first described by Casady and Lepkowski (1993) and later modified by Tucker, Lepkowski, and Piekarski (2002.) A list-assisted method of random digit dialing (RDD) described by Brick Waksberg, Kulp, and Starer (1995) is used to accomplish the sample. Weighting schemes are then employed to correct biases in the sample, including proportional representation of minority members of the sampled population. Probabilistically-sampled data such as these violate important assumptions of many statistical methods, thus rendering them inappropriate for use with such data. Further, popular statistical packages such as SAS, SPSS, and Minitab are unable to fully compensate for the some complex designs, nor are these packages always able to correctly use the weighting schemes accompanying some designs. At present, only specialized software, such as AM, WestVar, Stata, and Sudaan can be used to correctly analyze most probabilistically sampled data. These packages lack the sophistication and ease of use associated with the popular statistical packages mentioned above.

Significance of the Study

Exacerbating this problem of misuse is the wide availability of NCES data. The data can be easily downloaded from the NCES website or CDs can be ordered, free of charge, from the National Center for Educational Statistics (NCES) website (http://nces.ed.gov.) This broad availability allows researchers to unknowingly produce erroneous analyses under the incorrect

assumption that the data are the product of a simple random sample. Below is from http://www.westat.com/Westat/pdf/wesvar/WV_4-3_Manual.pdf.

Many investigations have shown that ignoring the sample design and using simple random sampling methods leads to biased estimates (e.g., Landis, Lepkowski, Eklund, & Stehouwer, 1982; Kish, 1992; Korn & Graubard, 1995; Brogan, 1998.)

For example, Kim, Murdock, and Choi (2005) published in a peer-reviewed journal, but gave no indication that weights were used to compensate for the complex design of data.

Research Questions

This study addressed the following questions as they apply to the problem of using data gathered under a probabilistic sampling design:

1. Can data resulting from a probability sample be unweighted (or reweighted) to make them usable for analysis under the assumptions of SRS data;
2. Under what conditions do the weights matter? For instance, if using a sample that is composed entirely of the majority or minority subsample, are the results the same with or without the weights;
3. Can subsamples of the greater, probabilistically sampled data be replicated to restore, for each analysis, the proportions of the sample to those of the original population; and
4. Can a valid analysis be performed using the concatenated samples produced in research Question 3? Would degrees of freedom need to be adjusted to compensate for artificially high sample sizes?

Definition of Terms

Accuracy (of a sample estimate) - the degree to which a sample statistic corresponds to the population parameter it estimates.

Aggregate weights –adjust the entire sample, not individual observations to reflect specified quantities of some related variable.

Attributes – characteristics of interest associated with the members of a population.

Bootstrapping - a technique of variance estimation that employs the resampling of multiple subsamples taken, with replacement, from the greater sample.

Calibrated sample size - the sample size, calibrated against known values for the purpose of consistency (of sample size) and better representation of the population.

Cell weights – are the proportional contribution of a single observation from a sample to a population target frequency.

Design effect - is the ratio of the variance of an estimate using the sample produced by the sample design and the variance of that same estimate taken from a simple random sample of the same size.

Disproportionate stratification (Oversampling) – the division of a population into mutually exclusive groups (strata) based on known population proportions for some attribute for the purpose of sampling higher proportions of some strata.

Effective sample size – the sample size (SRS) required to produce the same variance of the estimation of some variable of interest as the variance of that same variance for the sample design used to initially collect the data.

Item – a question asked of survey respondents.

Item values – the answers given to a survey-taker by the respondent.

Iterative marginal weighting – a method of marginal weight creation that considers the attributes not only independently, but sequentially as well.

Jackknifing – a technique of variance estimation that employs the resampling of multiple subsamples taken, without replacement, from the greater sample.

Marginal weights – vary the influence of an observation based on multiple attributes.

Post-stratification – the assignment of observations to strata after the sample is collected.

Precision (of a sample) – the degree to which results from consecutive samples reproduce the results of previous samples.

Proportionate stratification – the division of a population into mutually exclusive groups (strata) based on known population proportions for some attribute for the purpose of creating similar proportions in the sample.

Replicate weights - are a record of the number of identical (i.e., replicated) observations that exist in a full data set.

Self-weighting sample – a sample designed so that it is representative of the intended population without the use of weights.

Stratification – the division of a population into mutually exclusive groups (strata.)

Values – the collected responses of the members of a population for a given variable.

Variables – observable characteristics of the members of a population that may vary among the members of that population.

Weights – are derived variables used to vary the influence of some observations on the data analysis.

Weighting effect – measures the effect of weights on the sample, separate from other design effects.

CHAPTER II:

REVIEW OF RELATED LITERATURE

Dorofeev and Grant (2006) noted that "little has been published on the subject of weighting" (p. 45). Despite their assertion that weighted data's "prevalence, particularly in larger surveys, makes it essential that practicing researchers be familiar with the purposes, principles, and method of weighting" (p. 45).

The techniques of data analysis taught in elementary statistics classes assume, even require, that the sample in question be a simple random sample (SRS.) That is, every member of the population must have an equal chance of being selected for inclusion in the sample. The expenses associated with data collection may, however, limit the size of the sample and thus the degree to which it represents its underlying population. Even in the case of inexpensively collected data, problems may arise. The absence of an exhaustive sampling frame, for instance, may make a usable sample impossible to obtain even with sufficient resources available to reach the desired sample size. In other cases, the sampling frame, itself, may contain bias. Thus, members of the population included in the incomplete sampling frame may bias the results of any analysis.

Important Concepts

The *accuracy* of a sample estimate is the degree to which a sample statistic corresponds to the population parameter it estimates (Dorofeev & Grant, 2006). Confidence in the accuracy of some statistics (such as demographic measures) can be of particular importance to the sample design if the variables in question give rise to known biases. With this knowledge, the accuracy of a sample with regards to its population can be improved by weighting the sample such that the

remaining variables in the sample are afforded more or less influence to according to the known bias of the weighted variables.

The *precision* of a sample, on the other hand, is the degree to which results from consecutive samples reproduce the results of previous samples (Dorofeev & Grant, 2006). Similar to the research design notion of reliability, precision measures the repeatability of the results of a sample given the same sampling design for successive samples. The law of large numbers tells us that precision increases with sample size. Sample design will also influence the precision of a sample.

The *effective sample size* of a complex sample is the simple random sample size required to produce the same variance of the estimation of some variable of interest as the variance of that same variable for the sample design used to initially collect the data (Dorofeev & Grant, 2006).

Stratification is the division of a population into mutually exclusive groups (strata) based on some characteristic of the population. A population may be stratified by race or income categories, for instance. In any case, stratification is often employed to gain a more representative cross section of a population based on some classifying variable. It is known, for instance, that voters generally behave differently in rural areas than in urban areas. Rather than taking a simple random sample across the entire state, a voting survey might, instead, use this knowledge by sampling a predetermined number of voters from each voting precinct. This would assure that voters of different types (at least with regards to urbanicity) would be proportionately represented in the sample.

The preceding example refers to *proportionate stratification*. If the proportionate representation of a population attribute is known, proportionate stratification makes use of this information to assure that the sample represents a cross section of the population with regards to

the value frequencies of the attribute. For instance, the proportion of a certain range of incomes may be known. A sampling design using proportionate stratification might assure that the correct proportion of people from different income levels are properly represented in the final sample.

Stratification may increase the accuracy of a sample by calibrating the selection to known values, but the standard error of an estimate taken from such a design can no longer be calculated by the familiar simple random sample formula. Moser and Kalton (1971) have shown that the standard error of the estimate of a proportion from a proportionately stratified sample is different from that of a simple random sample. They estimated the standard error of a stratified sample to be the following:

$$s.e.(\hat{p}) = \sqrt{\frac{\sum_i n_i \hat{p}_i (1 - \hat{p}_i)}{n^2}}$$

Where:

n = the total sample size

n_i= the sample size of the stratum i

$\hat{p}_i$ = the estimated proportion of stratum i having the attribute value.

Note that the expression above is maximized for equal proportions.

The standard error for a simple random sample, on the other hand, is simply:

$$s.e.(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Where:

$\hat{p}$ = the estimated proportion

n = the sample size.

Whereas proportionate stratification is performed for the purpose of ensuring that a population is properly represented with regard to the population proportions for some attribute, *disproportionate stratification (oversampling)* does just the opposite. Disproportionate stratification deviates purposefully from known proportions by selecting greater or lesser proportions of the population based on the some attribute of the population members. In the typical case, a sample may be designed to target the members of one particular group in order to assure sufficient sample sizes for the purpose of testing claims regarding that group. For example, a simple random sample of size *n* may contain too few members of a particular group of interest, such as a particular demographic attribute (race, geographic location, population density, etc.) to properly perform analysis of the group. In such a case, the population may be oversampled with regard to that attribute. To accomplish this oversampling, a telephone interviewer conducting a survey might, for instance, gather simple demographic information very early in the interview for the purpose of qualifying respondents. Once identified, members of the minority demographic may then be asked to complete the full interview while members of the majority demographic are skipped according to some sampling design scheme. Thus, the completed full sample will contain members who possess the attribute values being oversampled disproportionate to those found in the population. Members of these strata, knowingly and intentionally disproportionate to the population, are then typically weighted to eliminate biases introduced by their disproportionate (over or under) representation in the full sample.

Oversampling may also be accomplished by purposely increasing the sample size from subpopulations of the sampling frame that contain a disproportionately higher number of the minority population. NHES, for instance, takes a higher number of telephone numbers from

telephone exchanges (100-banks) in high-minority areas. By increasing the sample size from these areas, the sample size of the minority population is increased as well.

Ideally, strata membership is known in advance, thus allowing strata to be independently sampled, later to be combined to create the full sample. In other cases such as random digit telephone dialing (used by NHES,) the researcher may not know the strata membership of the person being interviewed until the interview is completed. Thus, sample members may be assigned to strata only after data have been collected. This is known as *post-stratification*.

Weighting

Samples are taken with the intent of perfectly representing populations with respect to the attributes measured. Researchers seek variables that allow the maximum amount of distinction among observations with regard to the attributes of interest. In this sense, variation in the data is necessary to the researcher. Nothing would be gained if data collected on certain attributes were constant values. Certain values of some variables may be associated with different means and variances for other variables. For instance, certain values of the variable for race (black, white, Hispanic, etc.) may be associated with different means and variances for the variables representing household income or grades in school. In the event that the sample frequencies associated with one attribute, expressed by the data as the value of a variable, inadvertently alters the mean and variance of the other variables of interest, the sample can be brought closer to a perfect representation of the population by weighting the data to mimic the population with regards to the mean/variance altering variable. Data from observations that are proportionally underrepresented in the sample may simply be multiplied by an over-weighting factor in order to increase their relative contribution. Likewise, overrepresented observations may be under-

weighted to diminish their contribution to the final analysis. This application of weighting to disproportionate stratification is the primary emphasis of this dissertation.

Weighting Methods

Throughout this document, the following convention will be adopted. Variable names will be written in all upper-case letters, RACE, for instance. The nominal values of a variable will be written with an upper-case first letter; White, for instance.

Cell Weights

Cell weighting is the most direct method of weighting. Cell weights are created by first identifying a variable, the values of which will be used to measure the relative contribution of each observation. Typically, this would be the mean/variance altering variable mentioned above. Values of the weighting variable may be collapsed into categories (ages 20-29, for instance,) but the resulting variable must be mutually exclusive with regards to the individual observations and it must complete the sample space. That is, every observation must belong to a group, but no observation may belong to more than one group. To create cell weights, a target value is established as the frequencies of the values of the weighting variable for the population (the population parameter value). This target is the expected sum of the weights for frequencies of that particular cell. Put another way, for any cell, the observed frequency multiplied by the weight should equal the target.

The following hypothetical data illustrates this simple weighting scheme (see Figure 1).

Age	Target	Observed	Weight
Infant	10	12	.83
Child	60	55	1.09
Adult	30	33	.90

Figure 1. Example Frequencies

The variable, AGE, has three values: infant, child, and adult. A population of N is known to have 10% infants, 60% children and 30% adults. Thus a population of N=100 would be expected to have 10 infants, 60 children and 30 adults. A random sample of size n=100 produced data from 12 infants, 55 children, and 33 adults. The cell weight for each of the three categorical values of the variable AGE is determined by simple division:

$$cell\ weight = \frac{Target\ Frequency}{Observed\ Frequency}$$

Thus, the data collected from infants and adults are *down-weighted* (weight < 1) to minimize their over representation in the sample and children are *up-weighted* (weight > 1) to reflect the fact that data were gathered from proportionally fewer children than are actually in the population.

Marginal Weights

Sometimes referred to as “rim weights,” marginal weights are multi-dimensional and independent. That is, they are created using multiple sample variables, considered independently. The marginal distribution of each independent variable may be known, but the relationships among these marginal distributions may not be known. Nor may the multivariate relationship between the marginal variables and the greater population be known. Without this direct knowledge, we are thus unable to directly calculate target weights as in the simpler case of cell

weighting. Extending the cell weighting example above, if personal income were to be included in the data, the previous weights calculation would be insufficient to accommodate the additional variable. Instead, new weights must be created using the frequency totals of each unique value combination of the variables in question. Ideally, the weights for each unique value combination will sum to the target value for that combination.

Marginal weights may be derived by solving a system of equations of the weighting variables. The case of three weighting variables is illustrated here, but the concept readily extends to any number of variables.

For three variables, P, Q, R; let $(P_i)_{i=1}^{n}$, $(Q_j)_{j=1}^{m}$, $(R_k)_{k=1}^{l}$ be the target sums of weights for each of the three variables. The total target sum of weights must satisfy:

$$\sum_{i=1}^{n} P_i = \sum_{j=1}^{m} Q_j = \sum_{k=1}^{l} R_k$$

The nonempty cell for each *i*, *j*, *k* triplet contains the frequencies of the three values for each of the three variables, denoted $n_{i,j,k}$. Let $w_{i,j,k}$ then be the weight for each cell that satisfies the following system:

$$\begin{cases} \sum_{j,k} w_{i,j,k}\, n_{i,j,k} = Pi & i = 1, \ldots, n \\ \sum_{i,k} w_{i,j,k}\, n_{i,j,k} = Qj & j = 1, \ldots m \\ \sum_{i,j} w_{i,j,k}\, n_{i,j,k} = Rk & k = 1, \ldots, l \end{cases}$$

Fully enumerated, this will produce then a system of n+m+l equations with n*m*l weights (minus the number of empty cells) (Dorofeev & Grant, 2006).

It should be noted that weights are usually required to be non-negative. This means that a solution to the above solution does not always exist. See Dorofeev and Grant (2006), page 57, for an example of such a system.

Dorofeev & Grant, 2006, note that negative weights give rise to several problems. First, negative weights may be symptomatic of a problematic dataset, perhaps one with extreme outliers and/or influence points. Even for well behaved data, negative weights are problematic in terms of their use by practitioners. Consider one possible interpretation of a negative weight: the assertion that because a negative weight exists, there must be a corresponding negative proportion of observations in the population that possess some combination of the attributes that produce the negative weight. This is, of course, an absurdity.

An obvious advantage of marginal weighting is the inclusion of more than a single dimension. If the variance in the data is altered by the values of more than one variable, then a method that allows for the use of this additional information is useful. Another advantage of marginal weights is that, when compared to cell weights, marginal weights typically exhibit less variance and consequently have less impact statistically.

The disadvantage of marginal weighting is that, because the weighting variables are considered independently, rather than as a multivariate set, relationships between and among the weighting variables themselves are not considered (Dorofeev & Grant, 2006). A potential (almost certain) artifact of this omission is the fact that the individual marginal proportions may not be reconcilable with the resulting composite proportions at every value of the variable. For instance, the weight of variables for age and income may (likely will) indicate a lesser or greater proportion of low income people in the population than is actually indicated by the marginal variables.

Iterative marginal weighting, more commonly referred to as “raking” is a preferred method of marginal weight creation by the NHES. Proposed by Deming and Stephan (1940), raking is an iterative process made tenable for large datasets by modern computing. Like all

marginal weighting methods, raking considers the weighting variables independently, but is distinguished from other methods in that it also considers these variables sequentially. Further, raking can be done numerically, rather than mathematically, as with the system of equations approach. A single complete iteration consists of one stage for each weighting variable. If the sample is being weighted on three dimensions, for instance, each iteration will be comprised of three stages.

Marginal iterative weighting is accomplished as follows. Marginal target weights are first established for each dimension (weighting variable). If the sample is being weighted for the purpose of mimicking a population, these target weights are simply the (known or desired) population frequencies of each value for each weighting variable. For example, if age is to be a weighting variable, then the number of people of a certain age (say between 20 and 29) in the population would be the target weight for that age group. A study that is weighting on k variables, each with m_i levels, would thus have $\sum_{i=1}^{k} m_i$ target weights.

Once the target weights are known, the first iteration begins by creating a correction factor for each level of each variable. This is simply the target weight for the variable level in question divided by the total observed frequency for that variable level in the sample.

Starting with the sample variable exhibiting the greatest divergence from its population counterpart, each sample frequency is then multiplied by the correction factor for that level of the chosen (stage 1) variable. Subsequent stages then follow the same procedure for each of the remaining weighting variables.

The next iteration sees this process repeated using the frequencies generated by the previous iteration. Further iterations are performed until the totals for each level of each weighting variable become acceptably close to the corresponding target weight.

Once the iterations are complete, the final weights are calculated for each observation. The final weight for each observation is the product of the initial weight, usually 1, and the correction factors of the weighting variables corresponding to the levels of the observation.

A Numerical Example of Marginal Iterative Weighting

The example below, using data from Dorofeev and Grant (2006), illustrates the raking process using a simple dataset that resolves to very near its target weights at the end of only one two-stage iteration. More complex data will, naturally, require multiple iterations.

In this example, two weighting variables are used, GENDER and AGE. The first correction factor for the GENDER variable value Men is calculated as the target number of men divided by the total number of men or 20/13 = 1.5385. The correction factor for remaining value of GENDER, Women, is then calculated in the same way to produce an initial correction factor of 10/17 = .588235.

Each frequency value of men is then multiplied by the correction factor for men and recorded in a new frequency table for GENDER and AGE. The first cell of the new table would then be the original frequency of the first value of Men (4.0) multiplied by the correction factor for men, producing 4.0 * 1.5385 = 6.15. The remaining two new values for men are then similarly calculated using the same correction factor. The three new values for women are then calculated using the correction factor for women. This completes stage one of the first iteration.

Stage 2 of the first iteration proceeds as did Stage 1 by calculating the three correction factors for the three levels of AGE. Once the correction factors are calculated, they are then applied to the frequencies obtained in stage one.

Stage 1 of the second iteration (not shown) then continues by calculating new correction factor for GENDER, applying them to the values of gender and so on as in the first iteration.

The iterations continue until the corrected totals become tolerably near the target totals for both (all) variables (dimensions.)

At this point, the final weights are created by multiplying the initial factor (usually "1.0") times the correction factors applied to each cell at each stage. Once the final weight is obtained, the original data are then adjusted by multiplying the final weight by the corresponding frequency values in the dataset.

Observed Frequencies					
Age	**Men**	**Women**	**Total**	**Target**	
Age1	4.00	3.00	7.00	10	
Age2	5.00	8.00	13.00	10	
Age3	4.00	6.00	10.00	10	
Total	13.00	17.00	30.00		
Target	20.00	10.00			
CF	1.5385	.5882			
Stage 1 – Gender	**Men**	**Women**	**Total**	**Target**	**CF**
Age1	6.15	1.76	7.92	10	1.2629
Age2	7.69	4.71	12.40	10	.8066
Age3	6.15	3.53	9.68	10	1.0327
Total	20.00	10.00			10.18407
Target	20.00	10.00			
Stage 2 – Age	**Men**	**Women**	**Total**	**Target**	
Age1	7.77	2.23	10.00	10	0
Age2	6.20	3.80	10.00	10	0
Age3	6.36	3.64	10.00	10	0
Total	20.33	9.67			0
Target	20.00	10.00			
Final Weights	**Men**	**Women**			
Age1	1.94	0.74			
Age2	1.24	0.47			
Age3	1.59	0.61			

Note: the final weights above are produced after further iterations.

Figure 2. Example of Observed Frequencies

Aggregate Weights

Aggregate weights "adjust the results from a sample to ensure that the sums of the weighted quantities match target figures. These target figures may be established from official or industry statistics or from other reliable sources" (Dorofeev & Grant, 2006). Unlike the previously discussed cell weights and marginal weights which adjust the influence of the observations (respondents, in the case of a survey,) relative to one another, aggregate weights adjust the sample to reflect specified quantities of some variable related to the respondents. For instance, a study may not be interested in schools, *per se*, but rather, the total consumption or production of a school related item, say books, high calorie snacks, or SAT points. If the study were statewide in nature and aggregate weights were to be used, the sample would be adjusted to reflect the state totals for the number of textbooks purchased, SAT points scored, or the number of snacks sold.

Statistical Issues Related to Weighting

Recall that accuracy, as the term is used in reference to sampling theory, is the degree to which a sample statistic corresponds to the population parameter it estimates. Weighting seeks to increase the accuracy of analyses. Recall also that precision, in the current context, is the degree to which results from consecutive samples reproduce the results of previous samples. Accuracy and precision often represent a mutual trade-off. Increased accuracy by the use of weights is often obtained at the expense of precision because weights tend to increase the variance in the resulting sample.

Weights are, necessarily, not constant. The fact that the weights themselves have variance typically increases the variance of the adjusted sample produced by them. This greater variance produces smaller test statistics which lead to less frequent rejection of the null hypothesis for the

same level of alpha. Put another way, greater variance produces wider confidence intervals which are more likely to contain the hypothesized (null) value. Kish (1965) noted that weights used for the purpose of correcting bias are used reluctantly because they complicate the analysis and, usually increase the variance. These penalties, he further remarked, are, however, often less than the bias in the unweighted sample, hence their use.

Effective Sample Size, Design Effect and Weight Effect

While not directly important to the topic of this dissertation, the effective sample size and the design effect as explained by Dorofeev & Grant, 2006, are, nevertheless, important concepts in sample theory, thus their inclusion here. The *effective sample size* is a measure of this loss of precision due to the sample design. It is the size of the simple random sample exhibiting the same precision as the designed sample. The effective sample size, n_e is n, the size of the initial, simple random sample divided by the design effect. In other words, the effective sample size is the size of the simple random sample that would produce a parameter estimate with the same variance as the designed sample estimate.

$$n_e = \frac{nV_{ran}}{V}$$

Where:

n = the sample size.

V_{ran} = the variance of $\bar{x}$ for a simple random sample.

V = the variance of $\bar{x}$ calculated correctly (considering the sample design.)

The *design effect* (DE) takes the notion of effective sample size one step further. For some estimate, such as the mean, the design effect is the ratio of the variance of that estimate using the sample produced by the sample design and the variance of that same estimate taken

from a simple random sample of the same size as the designed sample. A ratio greater than 1.0 indicates a loss of precision attributable to the design and thus, at least in part, to the weighting.

Expressed mathematically, design effect is:

$$DE = \frac{V(\bar{x})}{V_{ran}(\bar{x})}$$

Where:

$$V_{ran}(\bar{x}) = \frac{V(x)}{n} \text{ and, thus } V(x) = nV_{ran}\bar{x}$$

Thus, the design variance is the SRS variance multiplied by the design effect.

Rearranging,

$$V(\bar{x}) = DE * V_{ran}(\bar{x})$$

Noting that $n_e = n/DE$,

$$V(\bar{x}) = \frac{V(x)}{V(x)} \cdot \frac{1}{1/V(\bar{x})} = \frac{V(x)}{V(x)/V(\bar{x})} = \frac{V(x)}{nV_{ran}(\bar{x})/V(\bar{x})} = \frac{V(x)}{n/DE} = \frac{V(x)}{n_e}$$

The standard error of a designed-sample estimate is then the square root of the design effect multiplied by the standard error of the estimate of that same parameter produced from a simple random sample.

$$SE_{design} = SE_{ran}\sqrt{DE}$$

No discussion of the effects of complex designs would be complete without an at least cursory discussion of design effect, hence its inclusion here. Note however, that design effect has two issues associated with it that are troubling to the primary discussion of sample weighting. First, weights make only a partial contribution to the design effect. In addition to weighting; stratification, clustering, and the complexity of the design all contribute to design effect. Further, the design effect calculations above are limited to a single, specific attribute for any sample

subset – not the entire sample. Fortunately, another method exists that measures the overall effect of effect of weighting on the entire sample.

Weighting Effect

A discussion of the weighting effect (WE) must begin with an understanding of the notion of the calibrated sample size, n_c. The calibrated sample size is simply the square of the sum of the weights divided by the sum of the squared weights. That is, the sample size is calibrated against known values for the purpose of consistency (of sample size) and better representation of the population.

$$n_c = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2}$$

Where:

n=the the sample size

w_i,…, w_n are the final weights.

This formula is useful to us in that it isolates the weights. Any use of this quantity will thus be free of contamination by non-weight design effects such as stratification and clustering. Further, the calibrated sample size of a sample reflects the sample as a whole, rather than individual items of even item values on a survey instrument. The above formula is discussed in very few papers (Conway, 1982; Sharot, 1986).) Readers of these should note that these papers use the term “effective sample size” for what we will call “calibrated sample size” (Dorofeev & Grant, 2006).

From this, the weighting effect is defined as:

$$WE = \frac{n}{n_c}$$

Where:

n = the sample size of the original, unweighted sample

n_c = the calibrated sample size.

Similar to the interpretation of the design effect, a weighting effect of 1.0 indicates that the weights have no effect. This result can only be obtained if the weights are constant, which is to say that the data are unweighted. Since the calibrated sample size cannot be greater than the original sample size, the weighting effect cannot be less than 1.0.

Variance Calculation

No test of statistical significance can be performed without an accurate measure of the variance of the variable(s) of interest. A simple matter for simple random samples, the calculation of variance is naturally more complex for stratified samples as each stratum will exhibit a different variance. This calculation is, however, a fairly straightforward sum of the "weighted variances" that uses the proportion of observations in each stratum to weight the variances of each individual stratum. The calculation of variance for samples subject to post-stratification is another matter. The complexity created by what may be multiple weighting stages with weights created by iterative and other means potentially necessitates the derivation of a different formula for each post stratification sample design. "In general," Dorfeev and Grant (2006) write, "much more research is required in this area."

The difficulties of variance calculation for complex designs have led to alternative, less direct, strategies for ascertaining the variance of such designs; among these are jackknifing and bootstrapping. The jackknife technique, originated by Quenouille (1949), employs a resampling scheme in which multiple subsamples of size n-1 are taken, without replacement, from the greater sample. Each successive subsample deletes one observation from the greater sample;

beginning with the first observation and proceeding sequentially with the remaining observations until a total of n-1 subsamples are created. Variance estimates are then generated from these subsamples and used to estimate the variance of the greater subsample. Like the direct estimation of variances for simple random samples and stratified samples above, jackknifing is a straightforward technique for simple random samples. The technique becomes difficult, however, for complex designs as it becomes design-specific. This design-specific complexity includes samples subjected to post-stratification. Further, Dorofeev and Grant (2006,) speaking of the jackknife formula for even a very simple stratified design, stated that "there is little information available about the performance of this formula for a range of surveys." Dorofeev and Grant (2006) pointed out that the asymptotic convergence assumptions in Shao and Tu (1995) are "too theoretical to check for any given survey."

The bootstrap technique, originated by Efron (1979), also employs resampling, but differs from the jackknife in that it samples with replacement. Bootstrapping falls victim to concerns for complex designs similar to those expressed for jackknifing.

Replicate Weights

A true weight is a multiplicative factor used to increase or decrease the contribution of some observations to the analysis. In this sense, *replicate weights,* as discussed by Dorofeev & Grant, 2006, are not true weights and are thus considered separately from the weights mentioned above. Replicate weights are, instead, simply a record of the number of identical (i.e., replicated) observations that exist in a full data set. When replicate weights are used, all but one of these identical records is removed from the stored data set. Analysis is then performed using this false "weight" to replicate the omitted records thus restoring, for analysis purposes, the original, full dataset. The reason for the use of replicate weights is simple – efficient storage. For very large

datasets with many identical records, the inclusion of only unique records plus a single additional variable to indicate the degree to which each of these records must be replicated is more efficient in terms of storage requirements than storing the original dataset in its entirety – duplicate records and all.

Taylor Series

Another method of variance estimation is *Taylor Series Estimation*. The Taylor Series technique uses two derived variables, in addition to the response data. The first, a stratum-level variable, indicates the variance estimation stratum from which a given observation (telephone number) is selected. The second variable, the primary sampling unit (PSU,) is an arbitrary numeric identification number which identifies the observation within the stratum. It is simply a count variable within each stratum. The Taylor Series Method uses these new variables to produce standard errors by repeatedly sampling with replacement. Although not identical, this method produces estimates in a manner similar to that of the jackknife method. The Taylor Series tends to produce slight underestimates of the parameters when compared to the NHES method of replication. Further, Taylor Series estimation packages currently available are not able to correctly estimate variances for two-phase samples. Since the NHES is a two-phase sample, this method will receive no further consideration as it is not a serious competitor to replication methods (NHES, 2007).

An Example of Real-world Weighted Data

The National Households Education Surveys program (NHES,) a product of the National Center for Education Statistics (NCES) under the U.S. Department of Education, creates, maintains, and distributes several large databases of primarily survey data. The surveys use a list assisted, Random Digit Dialing method of collection originally proposed by Casady and

Lepkowski (1993) and later modified by Tucker, Lepkowski, and Piekarski (2002). The surveys are conducted using a list-assisted random digit dialing method proposed by Brick et al (1995). The resulting design is a two-phase single stage, unclustered method that, unaltered, produces a self-weighting sample of telephone numbers.

In order to produce sample sizes sufficient for the reliable analysis of data from racial minorities, blacks and Hispanics were oversampled. This was accomplished by stratifying telephone exchanges based on the joint concentration of blacks and Hispanics. Telephone numbers for black and Hispanics were then oversampled by increasing the sample sizes for these high minority strata. In order to avoid further bias associated with the implementation of this portion of the design, weights (described later) were calculated to account for the probability of selection at each step in the process (NHES, 2007).

In the first stage of the collection process, a single-stage sample of telephone numbers from telephone exchanges with high minority representation was taken. These high minority strata were created based on 100-banks. A 100-bank is the set of all telephone numbers with the same first 8 digits, for example 205-348-60xx. 100-banks without at least one listed residential number were excluded from the sample. Potential survey respondents were then selected by a simple random sample (SRS) of the telephone numbers in each stratum (NHES, 2007).

Telephone numbers from these high minority strata were sampled at approximately double the rate of telephone numbers from low minority 100-banks. Once the sample was taken, the mailable status of each telephone number sampled was then determined by attempting to match each selected telephone number with a mailing address (NHES, 2007).The mailable status of the selected telephone numbers was included for the purpose of follow-up calls upon commencement of the actual survey.

In the second stage of the collection process, the telephone numbers from within each of the minority strata were subsampled based on the mailable status of the telephone number. In other words, only telephone numbers with known, associated addresses were included.

Additionally, two further sources of potential bias were noted by the investigators:

1. Difference in telephone coverage rates for different subgroups of the population; and
2. Nontelephone households (no telephone or cell-phone only.)

A screening survey was first given to successfully contacted respondents selected in the two phase process described above. Based on the information gathered, respondents to the screening survey were deemed eligible or not eligible to complete an extended survey interview.

The 2007 Parent and Family Involvement in Education Surveys

As part of the NHES, the Parent and Family Involvement in Education Surveys (PFI) were conducted in 1996, 2003, and 2007. The PFI interviewed the parents of children age 3 to grade 12 regarding many aspects of family involvement in children's education, both in and outside of school. Data were collected on 10,681 children (NHES, 2007).

Weighting Methodology of the 2007 Parent and Family Involvement in Education Surveys

NHES surveys are conducted entirely by landline telephone. Since all household do not have a landline telephone, an adjustment was made to compensate for potential bias created in the sample by failure to select potential respondents without landline telephones. The adjustment used results from the October 2005 and March 2006 Current Population Survey (CPS) as target population totals. The March 2006 CPS weights were then adjusted to the 2000 Decennial U.S. Census. Once done, the final weight, FPWT, known in the documentation as the full sample weight or the child weight was created using household-level weights and person-level weights.

Replicate weights were then created for the purpose of variance estimation. Note that what the NHES documentation calls replicate weights differs substantially from the use of the term by Dorofeev and Grant (2006) above. For the remainder of this paper, the term "replicate weights" will refer to the NHES use of the term.

The Full Sample Weight

The full sample weight (FPWT,) also known as the child weight for the PFI, was created using household-level weights and person-level weights.

The Household Weight

The household weight is a product of the following five factors:

1. A_j = the telephone-level base weight. This is the weight associated with the differential sampling based on the minority stratum of the telephone exchange and the mailable status of the telephone number described in phases one and two, above;

   B_j = the adjustment for subsampling of cases for nonresponse followup;

2. C_j = the adjustment for screener nonresponse;

3. D_j = the adjustment for the number of telephones in the household; and

4. E_j = a post stratification adjustment to compensate for the fact that only landline telephone households were eligible for the NHES (2007) surveys.

A description of each of these weights follows:

A_j, the telephone-level base weight, is the product of two factors, created by a two-stage process. In the first stage, a sample of telephone numbers was selected from the minority strata telephone exchanges. Telephone numbers from these high-minority exchange strata were sampled at approximately twice the rate of the low-minority strata. During this phase, an attempt

was made to match each telephone selected to an address listing. In the second phase, the telephone numbers were subsampled from within each minority stratum based on the mailable status of the telephone number, mailable status, refers to whether a mailing address was obtained for the phone number.

The following table, taken from the *NHES 2007 Methodology Report* (page 159) illustrates both the values used and the resulting weights for each stage.

Table 7-1. Weighting factors for the sampling of telephone numbers: 2007

Minority stratum	Phase 1 sample			Phase 2 sample			
	Number of telephone numbers in frame	Number of telephone numbers sampled	Weighting factor	Mailable status	Number of telephone numbers in Phase 1 sample	Number of telephone numbers sampled	Weighting factor
High minority	103,520,200	134,789	768.02	Mailable	83,876	82,366	1.02
				Not mailable	180,497	66,695	2.71
Low minority	179,711,500	117,037	1,535.51	Mailable	69,895	69,895	1.00
				Not mailable	141,899	59,534	2.38

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Household Education Surveys Program (NHES), 2007.

Figure 3. NHES' Table 7-1. Weighting Factors for the Sampling of Telephone Numbers: 2007

The High Minority stratum weight, for example, is given as:

$$768.02 = 103{,}520{,}200/134{,}789.$$

Likewise, for the mailable stratum, the weight is given as:

$$1.02 = 83{,}876/82{,}366.$$

The telephone-level base weight, A_j, then, for a survey respondent in the High Minority and Mailable strata would be 783.38 = 768.02*1.02.

B_j, the adjustment factor for the subsampling of cases for nonresponse followup, is needed because of the large number of followup calls required. In the event of a nonresponse on the first attempted contact, additional attempts (followups) are made. Sixty percent of telephone

numbers included in the sample required followup calls. The weighting factor given these nonresponse cases selected for additional (followup) attempts is:

$$B_j = \frac{\sum_{h \in NF} A_h}{\sum_{h \in NC} A_h}$$

Where:

h = the subscript for a single household

NF is the set of all screener nonresponse cases and

NC is the set of all screener nonresponse cases designated for additional attempts.

Nonresponse cases that were not chosen for inclusion in the subsampled were assigned a weighting factor of $B_j = 0$. Cases for which the interview was completed on the first attempt were assigned a weighting factor of $B_j = 1$.

For a selected telephone number *j*, the unadjusted household weight is then:

$$\text{UHW}_j = A_j B_j.$$

C_j, the nonresponse factor, adjusts for households that did not respond to the screening survey. Each telephone number was classified as a responder (R,) a nonresponder (NR,) or ineligible (I.) The base weights of the nonrespondent cases were distributed to the base weights of the respondent cases within a nonresponse adjustment cell. A Chi-square Automatic Interaction Detection (CHAID) analysis was used to identify the characteristics most associated with nonresponse. These, primarily geographic, characteristics were used to form the cells, c, of the nonresponse adjustment of the household weights. The adjustment is applied to each responding household, *j*, in adjustment cell, *c*.

$$C_{j(c)} = \frac{\sum_{h \in R_c \cup NR_c} UHW_h}{\sum_{h \in R_c} UHW_h}$$

D_j, the adjustment for the number of telephones in the household, is assigned a value equal to the reciprocal of the number of phones in the household, up to 3 phones. So,

$D_j = 1$ for households with exactly one telephone.

$D_j = \frac{1}{2}$ for households with exactly two telephones.

$D_j = \frac{1}{3}$ for households with three of more telephones.

If a household with two telephones is sampled twice, once for each telephone number, one of the responses is coded as a duplicate and removed from the sample. The value of D_j for the remaining number is set to equal 1.

The nonresponse adjusted household weight is then:

$$UHW_j' = A_j B_j C_{j(c)} D_j$$

$E_{j(d)}$, the final factor in the household weight, uses poststratification to obtain an adjustment for the household weight to account for sampling only landline phones. The poststratification process provides an adjustment to ensure that the weights sum to known national totals from the previously mention March 2006 CPS. Table 7-3 of from the NHES 2007 Methodology Report gives these totals by Census regions and the presence of children under 18 years of age as used in this step.

The final household weight for household *j* is then given by:

$$HHW_j = UHW_j' E_{j(d)}$$

Person Level Weights

In order to minimize the burden on families who agree to participate in the extended interviews, a sampling algorithm was implemented for all NHES 2007 surveys to limit the number of persons who participate in extended interviews per household (NHES, 2007).

The within-household sampling scheme followed the following guidelines:

1. No more than three persons were sampled in a given household;
2. Exactly one preschooler was sampled in every household that had at least one, and exactly one child enrolled in kindergarten through twelfth grade was sampled in every household that had at least one;
3. Because adult education participants were of particular interest, they were sampled at a higher rate than other adults; and
4. In households with eligible children, adults were sampled at lower rates than in households without eligible children. Additionally, adults in households with children sampled for both SR and PFI interviews were sampled at about half the rates of adults in households with only one child sampled (NHES, 2007, p. 159).

The person-level weight, PW_{jk}, a for person, *k*, from household, *j*, is the household weight, HHW_j, multiplied by 2 adjustment factors, A_{jk}, and B_{jk}, to account for this in-household sampling scheme.

A_{jk}, the adjustment for the probability of sampling a given person in the selected household, is always equal to 1.0 for the PFI because one, and only one child was selected from every eligible household. This factor is, however, non-trivial for the Adult Education for Work-Related Reasons survey (AEWR) because of the with-in household sampling scheme described above.

B_{jk}, adjusts for the probability of choosing person *k* from among all eligible persons in the given domain in household *j*. B_{jk} is given by:

$$B_{jk} = N_{jk}$$

Where N_{jk} is the numbers of persons in household *j* in the same sampling domain as person *k*. Thus, for a household consisting of N = 3 eligible members, the probability of being selected is $\frac{1}{3}$. The adjusted frequency is then $N\frac{1}{3} = 1$.

The unadjusted personal weight, UPW_k selected from household *j*, is then given by:

$$UPW_k = HHW_j A_{jk} B_{jk}.$$

This weight is then adjusted for nonresponse to the extended interview. This is accomplished by distributing the unadjusted person-level weights, UPW, of the nonrespondents to the unadjusted person-level weights of the respondents within a nonresponse adjustment cell. For the PFI, these nonresponse adjustment cells were created using home tenure, Census regions, age/grade and a homeschool indicator variable. These variables were chosen for two reasons: First, they are available for all sampled children (both respondents and nonrespondents,) and second, because they relate to response propensity.

These nonresponse adjustment cells are listed in Table 7-4 of the 2007 Methodology manual (see Figure 4).

Table 7-4. SR-NHES:2007 and PFI-NHES:2007 interview nonresponse adjustment cells

Explanatory variables	Number of respondents in cell	Completion rate (percent)
Homeowner/not enrolled, preschooler/homeschooler	84	93
Homeowner/not enrolled, preschooler/ not a homeschooler	1,937	78
Homeowner/grades K-5/homeschooler/Northeast/West	52	96
Homeowner/grades K-5/homeschooler/South/Midwest	73	78
Homeowner/grades K-5/not a homeschooler	3,361	74
Homeowner/grades 6-7	1,227	80
Homeowner/grades 8, 9, 10	2,083	77
Homeowner/grades 11, 12/Northeast/South	818	70
Homeowner/grades 11, 12/Midwest/West	748	78
Rent or other/Northeast	526	63
Rent or other/Midwest/South/homeschooler	52	80
Rent or other/Midwest/South/not a homeschooler/3-4 year olds	244	80
Rent or other/Midwest/South/not a homeschooler/5-20 year olds/not enrolled, preschooler, Kindergartener	159	63
Rent or other/Midwest/South/not a homeschooler/5-20 year olds/grades 1-8	665	70
Rent or other/Midwest/South/not a homeschooler/5-20 year olds/grades 9-12	290	63
Rent or other/West/not enrolled	118	81
Rent or other/West/preschool, grades K-1	209	68
Rent or other/West/grades 2-5	310	78
Rent or other/West/grades 6-12	358	72

SOURCE: U.S. Department of Education, National Center for Education Statistics, Parent and Family Involvement in Education (PFI) Survey of the National Household Education Surveys Program , 2007.

Figure 4. NHES' Table 7-4. SR-NHES:2007 and PFI-NHES:2007 Interview Nonresponse Adjustment Cells

The nonresponse adjustment factor for every respondent, k, in adjustment cell, c, $C_{k(c)}$ is given by:

$$C_{k(c)} = \frac{\sum_{h \in R_c \cup NR_c} UPW_h}{\sum_{h \in R_c} UPW_h}$$

For each selected person, *k*, the nonresponse-adjusted person-level weight, NPW_{jk} is then:

$$NPW_{jk} = UPW_{jk} C_{jk(c)}$$

Finally, the nonresponse-adjusted person-level weights, NPW, were raked to national totals. In addition to improved reliability of estimates, raking to national totals corrects the bias around households not included in the sample due to telephone status (no landline, unlisted

numbers, etc.) Raking to national population totals also allows sample estimates to be on the same scale as, and therefore comparable too, national figures. Specifically, the data were raked on the following three variables:

1. A cross between the race/ethnicity of the child and household income;
2. A cross between Census region and urbanicity; and
3. A cross between home tenure and grade in school.

These variables were chosen for two reasons: the variables' inherent importance to typical use of the data (e.g., grade) and telephone coverage (e.g., income and race/ethnicity.)

The national totals to which the sample was raked are given in Table 7-7 of the NHES:2007 Methodology Report (p. 171).

The final person-level weight for each sampled person, *k*, is then:

$$PW_{jk} = NPW_{jk}D_{jk(d)}$$

Where $D_{k(d)}$ is the adjustment factor produced by the raking process for cell *d* where the person k has the attributes corresponding to the levels of the dimensions of raking cell *d.*

Conventional Replicate Weights

Earlier, replicate weights were described as values used to replicate duplicate records deleted from the original full sample. As such, these values are said to make no contribution to the correction of bias in the sample; they merely restore the archived form of the sample to its full self. These values are not considered, by some, to be weights in the common use of the word. The term "weight" normally refers to a multiplicative factor used to increase or decrease the contribution of some observations to the analysis. But is that not what replicate weights of this definition do? Consider a full sample that contains two identical records. If one of these records is deleted and the remaining twin is assigned a replicate weight of "2," then the contribution of

that half of the pair is doubled by fully restoring the sample with no change in parameter values. By this definition, replicate weights are, in fact, weights, simply not weights employed to reduce design bias. It stands to reason that that replicate weights of this type would be integer values. A fraction of a record would never be deleted nor remain and thus never need to be fractionally restored.

Replicate Weights Used by the NHES

Replicate weights, as the term is used by the NHES, do not fit the description above, nor are they integer values. From this point forward, unless otherwise noted, the term "replicate weights" will be used as defined by the NHES method.

The NHES (and other NCES surveys) create replicate weights based on strata membership. In the present case, the process is as follows:

1. The full sample is divided into 80 groups (replicates) based on the sample design. For NHES:2007, the design factors used were minority status, listed stratum, and the sampling order of the telephone numbers.

2. Weights are calculated for each replicate using the same methods used to calculate the previously discussed full weight.

3. An estimate of the variance of some population parameter is calculated using a jackknife variance estimator of the form:

$$v(\hat{\theta}) = \frac{G-1}{G}\sum_{k=1}^{G}(\hat{\theta}_{(k)} - \hat{\theta})^2$$

Where: θ = the population parameter of interest

$\hat{\theta}$ = the estimate of θ based on the full sample

$\hat{\theta}_{(k)}$ = the estimate of θ based on the k_{th} replicate

G = the total number of replicates

For NHES:2007, the procedure followed that outlined by Kim, Navarro, and Fuller (2000). Specifically,

1. The entire sample of 467,167 selected telephone numbers were divided into the two minority strata used in phase 1 of the survey design;

2. The telephone numbers within the two strata were sorted to the same order as the phase 1 selection;

3. Eighty replicates were systematically formed by assigning the 1_{st}, 81_{st}, and 161_{st},… telephone numbers in the stratum to replicate 1; the 2_{nd}, 82_{nd}, and 162_{nd},… telephone numbers were assigned to replicate 2, etc., until every telephone number was assigned to one of the 80 replicates; and

4. Each telephone number was then assigned a replicate weight variable, REPL1,…, REPL80. The replicate phase 1 base weights were assigned to all 467,167 telephone numbers by multiplying the full sample base weight by either 0 or $\frac{80}{79}$, thus dropping one telephone number and adjusting the remaining one to proportionally match the original sample.

For example, to create the base weights for replicate 1, a replicate base weight of 0 is assigned to every member of REPL1 and a replicate base weight of $\frac{80}{79}$ is assigned to every member of REPL2 through REPL80. The phase 2 sample is then assigned a final base weight by applying a subsampling adjustment to the replicate phase 1 base weights within each of the phase 2 strata. Within each stratum, this adjustment “weights up” the replicate base weights of the phase 2 units to the replicate base weight totals of the phase 1 units.

5. The weighting steps used to create each of the sets of the full sample weights was applied to every one of the 80 replicate phase 2 base weights except that the raking convergence criteria was raised to within 10 for the replicate weights as opposed to a stricter requirement that the totals be within 1 for the full sample weights.

These resulting replicate weights for the PFI are named FPWT1, . . . FPWT80. They are employed in the analysis by calculating 81 estimates – one using each replicate weight, plus one using the full sample. The variation in the estimates computed using the replicate weights are then used to estimate the sampling errors from the full sample using specialty software such as AM and WesVar.

A Demonstration of the Impact of Weighting on Coefficient Estimates, Standard Error, and P-values Using NHES Data

Much theory regarding weighted samples has been discussed above. What follows is a practical demonstration of the effect of weighting on coefficient estimates, standard error, and p-values.

A simple grade prediction model will demonstrate that differences may be observed when probabilistically sampled data are analyzed in three different ways. As a foil to this demonstration, an ordinary least squares regression model will be created and used to predict students' grades using independent variables for race, income and television.

The model used to illustrate the impact of weights is:

Grades= Race, Income, TV

In the first analysis, no weights will be used. In the second analysis, the Final Child Weight will be used. In the third analysis, both the final child weight and the replicate weights will be used (the correct analysis.) A.M. Statistical Software performs the analyses using

ordinary least squares regression for the first two analyses and a Jackknife method for the last, replicated regression analysis. A.M. Statistical Software Beta Version 0.06.03. (c) was developed by The American Institutes for Research (A.I.R.) and Jon Cohen.

The Data

The analyses use data from the National Household Education Surveys Program of 2007, Parent and Family Involvement in Education Survey (PFI-NHES:2007.)

Variable Names, Descriptions, and Values

The model grades= Race, Income, TV will be specified using the following variables:

1. **SEGRADES**: Overall, what are the child's grades across all subjects?

Response	Value
Mostly A's	1
Mostly B's	2
Mostly C's	3
Mostly D's	4

2. **CBLACK** – Is the child Black of African American?

Response	Value
Yes	1
No	2

3. **TVWKDYNU** – How much time does the child spend watching television or videos on a typical weekday?

Response	Value
1-16	1-16

4. **HINCOME** – What is the total household income?

Response	Value
$5,000 or less	1
$5,001-$10,000	2
$10,001-$15,000	3
$15,001-$20,000	4
$20,001-$25,000	5
$25,001-$30,000	6
$30,001-$35,000	7

$35,001-$40,000	8
$40,001-$45,000	9
$45,001-$50,000	10
$50,001-$55,000	11
$60,001-$75,000	12
$75,001-$100,000	13
Over $100,000	14

Note: the following results are generated solely for the purpose of demonstrating the differences that may be observed when data of this type are analyzed in different ways. The reader should attempt no further interpretation of the models presented here as none the underlying assumptions of the models have been checked.

The A.M. output for the three analyses is:

Model: SEGRADES = CBLACK HINCOME TVWKDYNU

Regression: **No Weights**

Parameter	Estimate	SE	t	p > |t|
Constant	2.860	0.081	35.120	0.000
CBLACK	-0.078	0.042	-1.879	0.060
HINCOME	-0.022	0.004	-5.436	0.000
TVWKDYNU	0.428	0.018	23.608	0.000

Regression: **Final Weight Only**

Parameter	Estimate	SE	t	p > |t|
Constant	2.913	0.118	24.611	0.000
CBLACK	-0.051	0.059	-0.863	0.388
HINCOME	-0.033	0.006	-5.942	0.000
TVWKDYNU	0.394	0.024	16.154	0.000

Replicated Regression: **All Weights**

Parameter	Estimate	SE	t	p > |t|
Constant	2.913	0.130	22.434	0.000
CBLACK	-0.051	0.063	-0.807	0.422
HINCOME	-0.033	0.005	-6.324	0.000
TVWKDYNU	0.394	0.023	16.922	0.000

AM Statistical Software Beta Version 0.06.03. (c) The American Institutes for Research and Jon Cohen

For simplicity, define the three weighting levels as:

Level 1 – No weights are used.
Level 2 – Only Final Weight are used.
Level 3 – Both Final Weight and Replicate Weights are used.

The value in the table below represents the weighting levels that produce a change when moving from the previous level. For instance, a value of "2" indicates that a change is noted when the method of analysis moves from level 1 (no weights) to level 2 (final weight only.) A value of "2, 3" indicates a change for that statistics when moving among all 3 levels of weighting.

Parameter	Estimate	SE	t	p > |t|
Constant	2	2, 3	2, 3	2, 3
CBLACK	2	2, 3	2, 3	2, 3
HINCOME	2	2, 3	2, 3	2, 3
TVWKDYNU	2	2, 3	2, 3	2, 3

Summarizing, as we move from

1. the use of no weights to;

2. the use of only the final weight to; and

3. the use of both the final weight and the replicate weights (the correct analysis):

The regression coefficients change with the use of the final weight, but remain the same when the replicate weights are used. This demonstrates that parameter estimates can be correctly calculated using only the final weight. The final weight alone does not, however, allow for the correct calculation of the standard error. This requires some form of variance estimation such as replication.

CHAPTER III:

METHODS

Purpose

The purpose of this study was to create a method for using data from oversampled populations that minimizes the use of weights, which have been shown to increase variance in the data.

Data

For purposes of clarity, the proposed method will first be illustrated using simple, contrived numbers. The method will then be demonstrated and tested using randomly generated data.

Research Design and Data Analysis Procedure

To demonstrate the sampling method and subsequent analyses proposed in research questions, consider the following population:

i	RACE	freq(x)
1	1	2
2	1	2
3	1	2
4	1	2
5	1	2
6	1	2
7	2	4
8	2	4
9	2	4

Where:

i = the observation number

Race = the race of the respondent. Race is coded Race = 1 for the proportional majority value and race = 2 for the proportional minority value.

freq(x) = the number of respondents for some value of some survey item.

The population enumerated above has the following parameters: n = 9, $\mu = 2.667$, and $\sigma^2 = .889$

Additionally, the known population proportion for Majority, prop(Maj) = $\frac{2}{3}$ and the population proportion for Minority, prop(Min) = $\frac{1}{3}$, which we will concisely write as 2:1 (Majority:Minority.) From this population, take a disproportionate sample comprised of the last six observations, that is i = 4, ... i=9. Note that the Majority:Minority proportions of this sample are 1:1.

The disproportionate sample will thus be:

I	RACE	freq(x)
4	1	2
5	1	2
6	1	2
7	2	4
8	2	4
9	2	4

Method - Concatenated Subsamples

The population proportion, 2:1, can be restored for this sample by replicating the majority portion of the sample twice and then concatenating the three subsamples (2 majority and 1 minority) to create a proportionally correct super sample.

Specifically, the newly created super sample will be:

i	RACE	freq(x)
4	1	2
5	1	2
6	1	2
4	1	2

5	1	2
6	1	2
7	2	4
8	2	4
9	2	4

The mean and population variance of this super sample are, again, 2.667 and .889, respectively.

This is, of course, a contrived example meant only to illustrate the proposed method. A fuller demonstration follows:

Using SAS 9.2, a more thorough example was created by generating random numbers for 10,000 observations to approximate the sample size of the PFI-NHES:2007. For this example, assume that the majority proportion of the population is $prop(maj) = \frac{3}{4}$ and the minority proportion of the population is $p(maj) = \frac{1}{4}$. As above, we can write this as a population Majority:Minority proportion of 3:1.

The steps for the Method 2 demonstration follow:

1. A dataset was created for the majority population by generating 7500 normally distributed observations with an arbitrarily chosen mean of $\mu = 12$ and variance, $\sigma = 1$. Similarly, 2500 members of the minority population were generated, but with a mean and variance different from those of the majority population. The minority population was, again, normally distributed, but with an arbitrarily chosen mean of $\mu = 10$ and variance, $\sigma = 2$. The RAND function was used in both cases.

2. These two datasets (majority and minority) were then concatenated using PROC APPEND to create a population of 10,000 observations with the Majority:Minority proportions assumed above.

3. A new variable valued with random numbers was then generated and the dataset was sorted by this random number to create a randomly sorted dataset as might be encountered in a real-world situation. (This step is necessary only for the purpose of beginning with a data set that mimics a real-world example)

4. The population dataset is then sorted first by RACE and then by random number.

5. The first 2500 majority members and the first 2500 minority members (all of the minority members) are then taken from the population to create a random sample of each group.

6. These two datasets are then concatenated using PROC APPEND to create a new dataset, randomly sampled from the population, but with different proportions than the population. The minority members of the population are now proportionately $\frac{1}{2}$ of the sample ($prop(min) = \frac{1}{2}$.) Thus, the minority is oversampled with regards to the majority.

7. The population proportions are then restored using the oversampled data by further concatenating the datasets in proportion to the population proportions to create a proportionally correct super sample. That is, the minority members are included once and the majority members are replicated to create 3 replicate samples of the majority. These 4 datasets are then concatenated and the resulting dataset has the same majority/minority proportions as the original population.

The method can now be tested by comparing parameters calculated from the original population to statistics calculated from the super sample. PROC MEANS was employed in the comparison and the results are presented below:

Original Population

The MEANS Procedure

Variable	N	Mean	Variance	Std Error	Pr > |t|
y	10000	11.5140362	2.5371972	0.0159286	<.0001
x1	10000	11.5278108	2.4614370	0.0156890	<.0001
x2	10000	11.5042531	2.5409510	0.0159404	<.0001
x3	10000	11.4951090	2.4515898	0.0156576	<.0001

Super Sample

The MEANS Procedure

Variable	N	Mean	Variance	Std Error	Pr > |t|
y	10000	11.5050608	2.5459953	0.0159562	<.0001
x1	10000	11.5149305	2.4785656	0.0157435	<.0001
x2	10000	11.5027447	2.5245080	0.0158887	<.0001
x3	10000	11.5009026	2.4822937	0.0157553	<.0001

Note, in particular, that standard error estimates are the same as the parameters to 4 significant digits for x1, and x2; and the same to 3 significant digits for y and x3.

In order for the proposed method to work,

1. The sample must contain response data for both the oversampled and undersampled members of the population; and
2. The population proportions for the oversampled and undersampled members of the sample must be known.

Nonresponse bias, for instance, cannot be addressed using this method because, by its very nature, response data do not exist for non-responders. Thus, in the case of the PFI data, the

goal of eliminating all weights will be unattainable. Instead, the secondary goal will be to minimize the number of weights used, thereby decreasing the variance of the variables and hence decreasing the standard error of the analyses.

For the purpose of addressing the research questions of this dissertation, I will compare my method of calculating the first three moments of a disproportionately sampled population to the calculation of the same three moments using the weights method. My primary concern is not with differences between my estimates of the first three moments and the population values for these three moments (the parameters). Instead, my primary concern is the difference between estimates of the first three moments produced by my method and estimates of the first three moments produced by the weights method. Estimates produced by both of these methods will then be compared to the population parameters. Thus, the distributions will be compared on the basis of the first three moments using the population parameter moments as the standard of comparison.

Step 1: Initially generate a population of 100,000 observations with 2 variables, X and Majority. Of these, 80,000 observations will be designated as the majority population (Majority = 1) while the remaining 20,000 observations will be designated as the minority population (Majority = 0). Thus, the population will have a majority:minority proportion of 4:1.

The variable X for the majority population will be generated to have an arbitrary mean and variance of (0, 1). The variable X for the minority population will be generated to have an arbitrary mean and variance of (2, 3). All of the samples below will then be drawn from this population.

Step 2: Create five randomly selected majority/minority samples of n = 10,000 as follows:

Sample 1 – 90% majority : 10% minority

Sample 2 – 80% majority : 20% minority

Sample 3 – 70% majority : 30% minority

Sample 4 – 60% majority : 40% minority

Sample 5 – 50% majority : 50% minority

For convenience, the research questions are restated below. The method used to address each research question follows:

For RQ1, can data resulting from a probability sample be unweighted (or reweighted) to make them usable for analysis under the assumptions of SRS data?

Method: Produce a counter-example by using each of the 5 majority:minority samples above. To do this:

1. Calculate the weights needed to restore the population proportion.
2. Multiply the data by the reciprocal of the weights.
3. Calculate the first three moments of the resulting reweighted data.
4. Calculate the first three moments of the weighted data.
5. Compare the results of steps 3 and 4.

For RQ2, under what conditions do the weights matter? For instance, if using a sample that is composed entirely of the majority or minority subsample, are the results the same with or without the weights?

Method: For each of the five samples:

1. Calculate the weights needed to restore the population proportion.
2. Using these weights, calculate the first three moments of the distribution separately for the majority and minority members of the population.

3. Without using these weights, calculate the first three moments of the distribution separately for the majority and minority members of the population.
4. Compare the results of steps 2 and 3.

For RQ3, can subsamples of the greater, probabilistically sampled data be replicated to restore, for each analysis, the proportions of the sample to those of the original population?

Method: From each of the 5 samples, replicate the majority and minority observations to restore the population proportion of majority:minority, 4:1.

The replication scheme for each will be determined by dividing the population proportion (4:1) by the sample proportion (9:1, 4:1, 7:3, 6:4, and 5:5). From this quotient proportion, the simplest integer equivalent is then found. The replication scheme can then be read directly from the resulting integer proportion.

For the 90:10 sample, for example, $\frac{4:1}{9:1} = 4:9$, thus the majority portion of the sample will be replicated 4 times and the minority proportion of the sample will be replicated 9 times.

For the 5 samples above, the majority:minority replication schemes are similarly found as follows:

For the 90:10 sample, $\frac{4:1}{9:1} = 4:9$
For the 80:20 sample, $\frac{4:1}{4:1} = 1:1$
For the 70:30 sample, $\frac{4:1}{7:3} = 12:7$
For the 60:40 sample, $\frac{4:1}{3:2} = 8:3$
For the 50:50 sample, $\frac{4:1}{1:1} = 4:1$

Frequencies of the values (0 and 1) for the variable Majority will be generated to confirm that the population proportion has been restored for each sample.

For RQ4, can a valid analysis be performed using the concatenated samples produced in research question 3? Would degrees of freedom need to be adjusted to compensate for artificially high sample sizes?

Method: For each of the five samples, calculate the weights needed to restore the population proportion:

1. Using these weights, calculate the first three moments of the distribution.
2. Concatenate the subsamples produced in research question 3.
3. Calculate the first three moments of the resulting super sample.
4. Compare the moments produced by steps 2 and 3.

For Research Question 2, Research Question 3, and Research Question 4, 100 iterations of the methods described above were performed and then summarized. In addition to increases reliability, these iterations allowed the calculation of standard errors for each of the three moments. The SAS code used for a single iteration of this process appears as Appendix C of this dissertation.

CHAPTER IV:

RESULTS

The purpose of this study was to explore the uses and effects of weights on disproportionately sampled populations and to suggest an alternative method of analysis. It was hypothesized that the alternative method might prove easier to implement, statistically superior, or both. To this end, four research questions were posed, the methods in Chapter III were employed, and results were obtained as follows.

Research Question 1

Research Question 1 asked if data resulting from a disproportionate probability sample can be unweighted (reweighted) to make them usable for analysis under the assumptions of simple random sampling. Although somewhat naïve, the fact that the question is even asked among otherwise learned professionals indicates the existence of a basic misunderstanding of weights and "weighted data." Data that are disproportionately sampled are rarely, if ever, presented as "weighted data." Instead, the weights are calculated and provided alongside the native data for use in an appropriate analysis. The weights are included for the purpose of correcting known bias in the sample due to disproportionate sampling.

If the data were, in fact, weighted, that is, if the data values already included a multiplicative factor reflecting the bias correction of the weighting scheme then the data could easily be unweighted by simply multiplying the weighted data by the reciprocal of the weights. This would restore the native data, but to use this now unweighted data as though it were the result of a simple random sample would be to ignore, and thus accept, the known bias the weights were intended to correct.

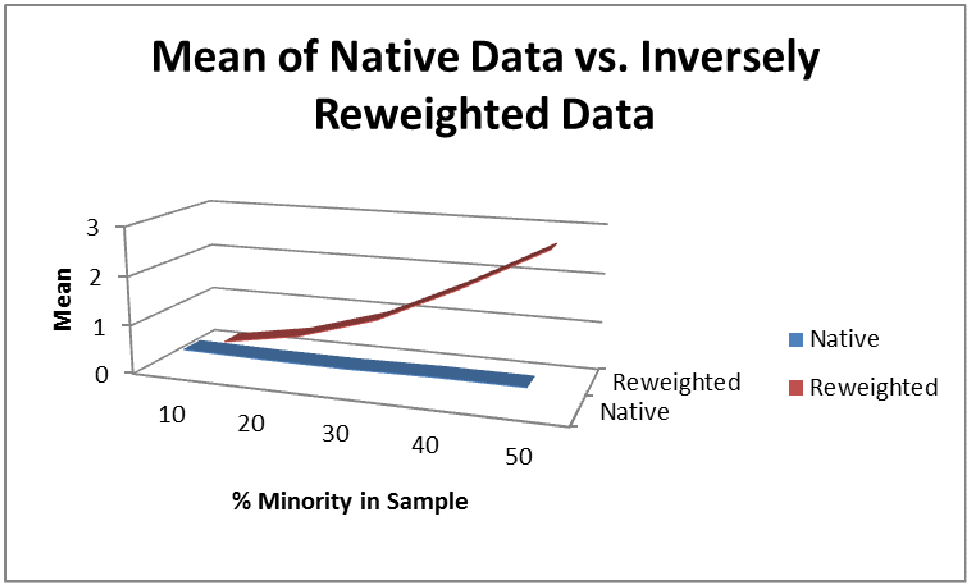
If, as is the more common case, the data were not presented as "weighted," that is the data were presented as native data with the bias correcting weights included for use in the proper analysis, the act of "unweighting" data that were never weighted place would produce absurd results as illustrated in Table 4.1.

Tables 4.1a, 4.1b, and 4.1c contain the first three moments for the distributions produced by the five subsamples. Each subsample is first appropriately analyzed by employing the weights to calculate the first three moments. The native data are then "unweighted" by multiplying each data value by the reciprocal of the weight for that observation. The first three moments of these "unweighted" data were then calculated as if the data were from a simple random sample.

Table 4.1a

Means of the Native Data and Means of the Inversely Reweighted Data

Mean	% Minority Sample				
	10	20	30	40	50
Native	0.44967	0.42758	0.42243	0.45408a	0.45497
Reweighted	0.13094	0.42758	0.94104	1.74454	2.65951

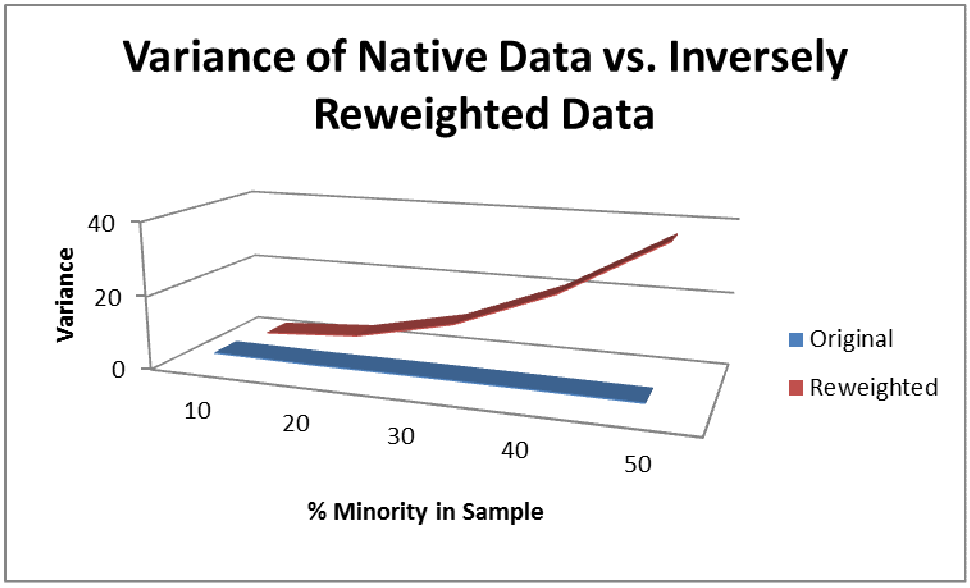


The differences between the means generated by the correctly employed weights and the means generated by the "reweighted" data are obvious. For the 90:10 sample the mean is 0.44967, while the reweighted data produce a mean value of 0.13094. The means for the proportional sample are the same for both groups because the weights are equal to "1." Thus, no effective weights are created as no bias exists in the sample. The means for each of the four disproportional samples (90:10, 70:30, 60:40, and 50:50) are clearly different for the native data and the reweighted data. Further, the means increase as the percentage of minority observations increases.

Table 4.1b

Variances of the Native Data and Variances of the Inversely Reweighted Data

Variance	% Minority Sample				
	10	20	30	40	50
Original	3.6759	3.4871	3.5903	3.5455	3.4199
Reweighted	1.5732	3.4871	9.5391	20.6148	36.6643

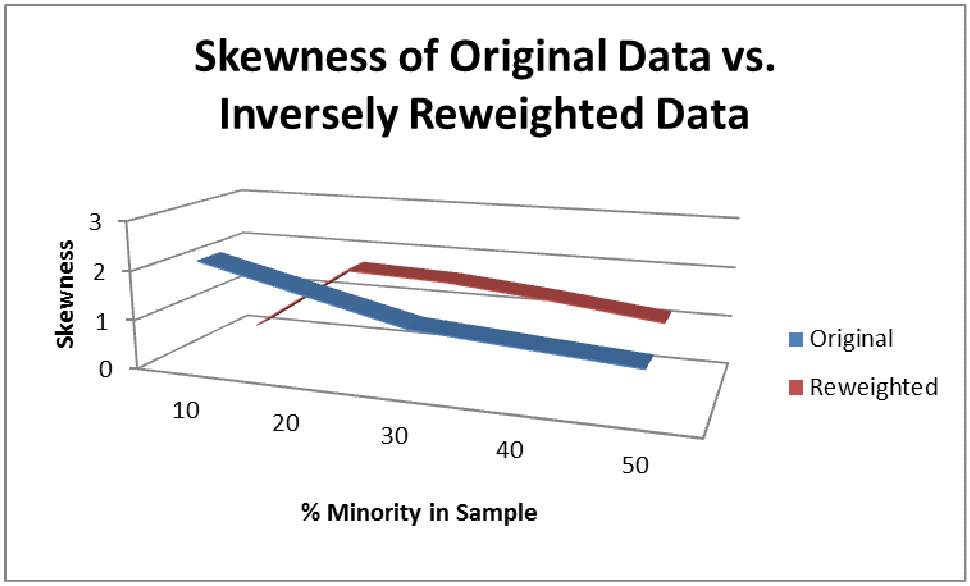


As with the means, the variances of the four disproportionate samples are clearly different for the native data and the reweighted data with an even more pronounced upward trend as minority proportionality increases.

Table 4.1c

Skewness of the Native Data and Skewness of the Inversely Reweighted Data

Skewness	% Minority Sample				
	10	20	30	40	50
Original	2.15962	1.66843	1.18617	1.0334	0.88753
Reweighted	0.26125	1.66843	1.59004	1.39083	1.14411



As with the first two moments, the coefficient of skewness of the four disproportionate samples are different for the native data and the reweighted data, but unlike the first two moments, the skewness exhibits no clear trend as minority proportionality increases.

The counter examples above show that data of this type may not be simply "reweighted" and that doing so produces erroneous results. The answer to Research Question 1 is, thus, "no."

Research Question 2

Research Question 2 asks, specifically, if the data can be correctly evaluated without the use of the weights for each level of the grouping variable. That is, for the grouping variable Majority, can the data be analyzed under the assumptions of a simple random sample if the researcher limits the analysis to either the majority or the minority group.

Table 4.2a, 4.2b, and 4.2c show the mean mean, mean variance, and mean coefficient of skewness of the 100 generated populations and samples. The standard error of each of these moments is included in each of the tables.

The moments are calculated for the 100 populations of randomly generated numbers and then the mean of each moment is calculated along with its standard error. These same moments are then similarly calculated by two methods for each of the five minority percentages of the 100 samples.

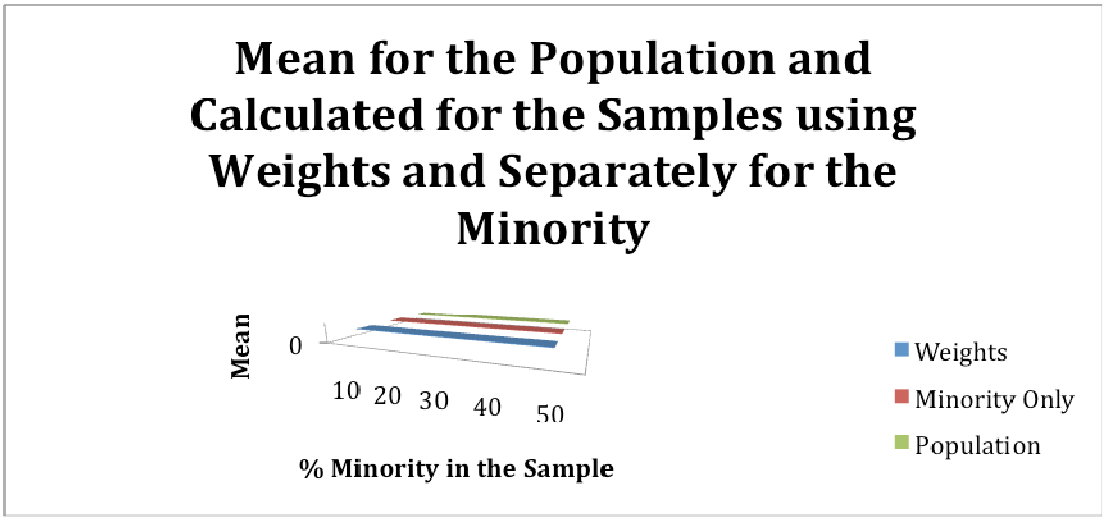
First, the first three moments are calculated for each of the five samples using the weights and a BY statement in SAS 9.2. The BY statement instructs SAS to perform separate analysis for each value of the "by" variable. Thus for the "by" variable, Majority, the moments are calculated separately for the majority and minority observations.

Second, the majority and minority observations were divided into separate datasets and the moments were calculated separately, as above, for each dataset under the assumptions of simple random sampling.

Table 4.2a

Mean Means for the 100 Populations and Calculated Separately using the Weights and Separately for the Minority

Mean Means	Percent Minority									
	10 Mean	10 SE	20 Mean	20 SE	30 Mean	30 SE	40 Mean	40 SE	50 Mean	50 SE
Weights	1.99128	0.02946	1.9992	0.02129	1.99434	0.01824	2.00404	0.01533	2.01093	0.01401
Minority Only	1.99128	0.02946	1.9992	0.02129	1.99434	0.01824	2.00404	0.01533	2.01093	0.01401
Population	2.00065	0.00227	2.00065	0.00227	2.00065	0.00227	2.00065	0.00227	2.00065	0.00227

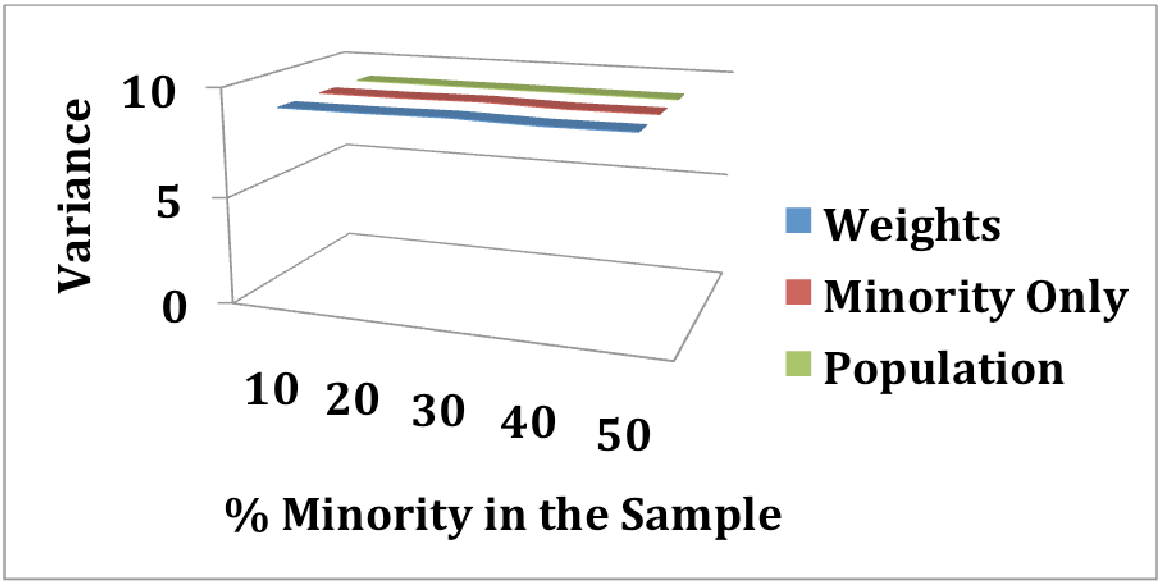


The values of the mean are identical for either method of analysis for all five samples. This indicates that, in the case of a disproportionate sample, the mean can be correctly calculated without the weights provided no other weights are required and the analysis is conducted on the data for a single level of the grouping variable on which the weights were calculated.

Table 4.2b

Mean Variances for the 100 Populations and Calculated Separately using the Weights and Separately for the Minority

Mean Variance	Percent Minority									
	10 Mean	10 SE	20 Mean	20 SE	30 Mean	30 SE	40 Mean	40 SE	50 Mean	50 SE
Weights	9.02767	0.14711	9.03728	0.09974	9.09239	0.08765	9.0104	0.07192	9.04607	0.06223
Minority Only	9.02767	0.14711	9.03728	0.09974	9.09239	0.08765	9.0104	0.07192	9.04607	0.06223
Population	9.0134	0.00974	9.0134	0.00974	9.0134	0.00974	9.0134	0.00974	9.0134	0.00974

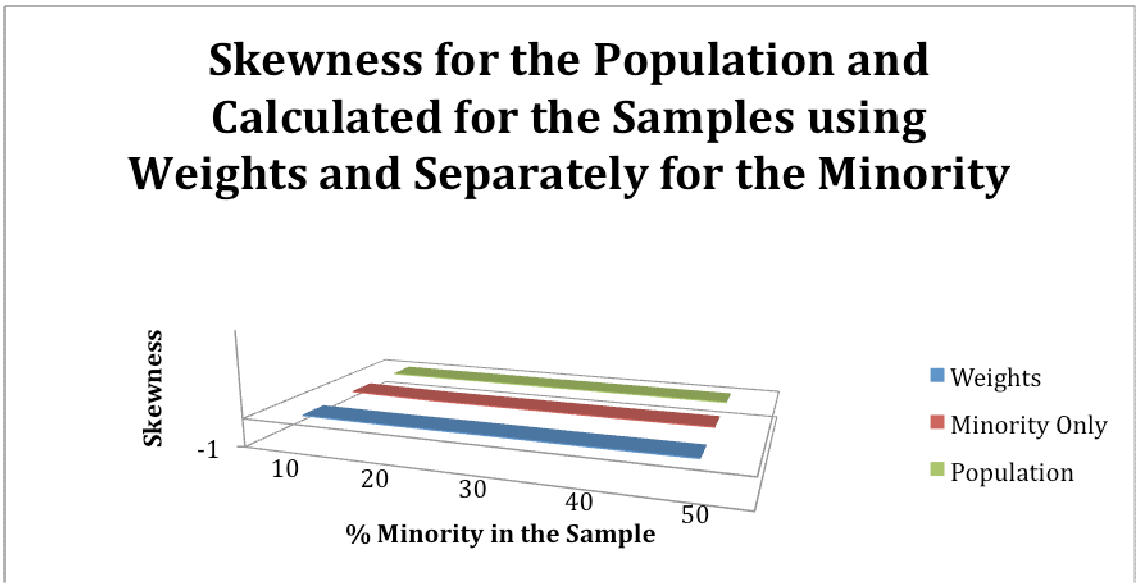


As with the means, the values of the variances are identical for either method of analysis for all five samples. This indicates that, in the case of a disproportionate sample, the variance can be correctly calculated without the weights provided the analysis is conducted on for a single level of the grouping variable on which the weights were calculated and no other weights are required.

Table 4.2c

Skewness for the Population and Calculated Separately using the Weights and Separately for the Minority

Mean Skewness	Percent Minority									
	10 Mean	10 SE	20 Mean	20 SE	30 Mean	30 SE	40 Mean	40 SE	50 Mean	50 SE
Weights	0.00378	0.02225	-0.00942	0.01731	-0.00059	0.01466	-0.00227	0.009895	0.01147	0.01151
Minority Only	0.00378	0.02225	-0.00942	0.01731	-0.00059	0.01466	-0.00227	0.009895	0.01147	0.01151
Population	0.00022	0.00191	0.00022	0.00191	0.00022	0.00191	0.00022	0.00191	0.00022	0.00191



As with the first two moments, the values of the skewness are identical for either method of analysis for all five samples. This indicates that, in the case of a disproportionate sample, the variance can be correctly calculated without the weights provided the analysis is conducted on a single level of the grouping variable on which the weights were calculated and no other weights are required.

Note: similar results were produced and identical conclusions were reached for the corollary analysis of the majority samples.

The answer to Research Question 2 is then, “the weights are required only if analysis is performed on the entire sample. The weights are not needed if analysis is done on a single level of the weighting variable and no other weights are required.”

Research Question 3

Research Question 3 asks if the majority:minority proportions of a disproportionately sampled population can be restored by replicating the majority and minority observations of the disproportionate sample and concatenating the replicated datasets to create a single super sample.

From each of the five samples, taken from 100 generated populations, the majority and minority observations were replicated to restore the population proportion of majority:minority, 4:1. Specifically, the replications for the five samples were as follows:

For the 90:10 sample, 4: 9

For the 80:20 sample, 1: 1

For the 70:30 sample, 12: 7

For the 60:40 sample, 8: 3

For the 50:50 sample, 4: 1

Frequencies of the values (0 and 1) for the variable majority were generated to confirm that the population proportion has been restored for each sample. In each of the five cases, the 4:1 proportion of the population was indeed restored according to the replication scheme for each, above. The answer to Research Question 3 is thus, “yes.”

Research Question 4

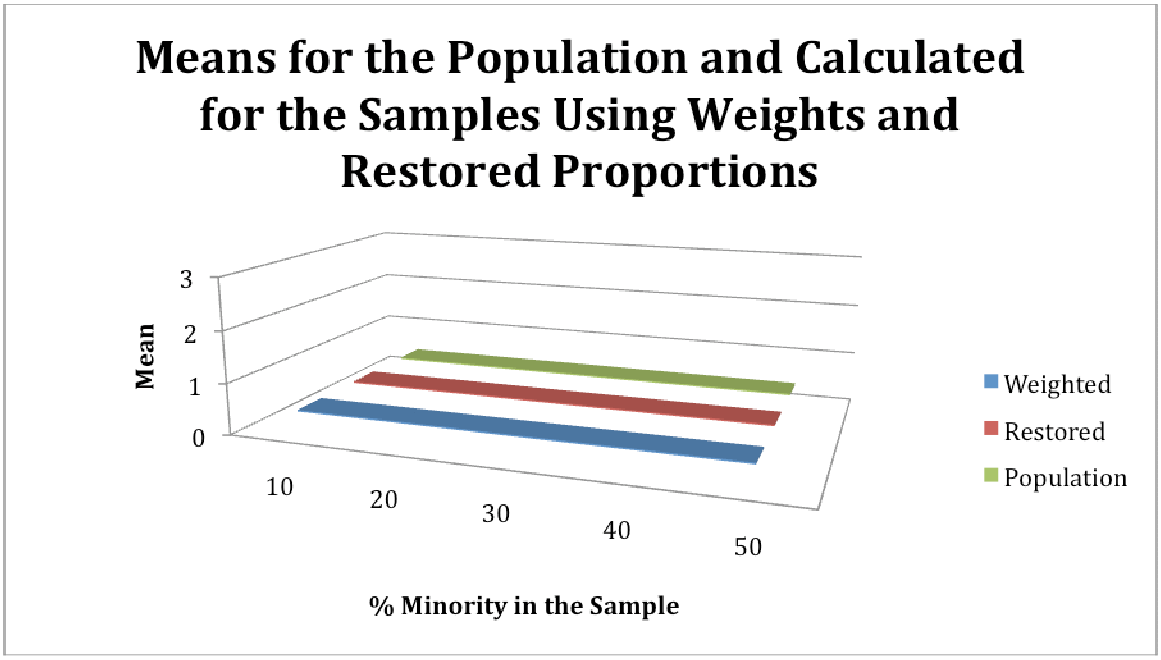
Research Question 4 asks if the concatenated datasets (super samples) from Research Question 3 can then be correctly analyzed under the assumptions of simple random sampling. The question is addressed by calculating the first three moments of each of the five samples taken from the 100 generated populations by two different methods. First, the moments are calculated from the original five samples using the weights method. Second, the moments are calculated for the proportion-restored super samples under the assumptions of simple random sampling. The means of each resulting moment are calculated along with the standard error for each. While the primary concern is between the moments produced by these two methods, the moments of the original population are also included to provide a standard of comparison.

Table 4.4a shows the mean means of each of the five samples taken from the 100 populations calculated first using the traditional weights method and then using proportionately restored super samples from Research Question 3.

Table 4.4a

Mean Mean for the 100 Populations and Calculated for the Samples using Weights and Restored Proportions

Mean Mean	Percent Minority									
	10 Mean	10 SE	20 Mean	20 SE	30 Mean	30 SE	40 Mean	40 SE	50 Mean	50 SE
Population	0.40006	0.000526	0.40006	0.000526	0.40006	0.000526	0.40006	0.000526	0.40006	0.000526
Weighted	0.4035	0.006786	0.40506	0.005223	0.40317	0.004928	0.40479	0.004838	0.40376	0.004928
Restored	0.40358	0.006779	0.40506	0.005223	0.4029	0.004909	0.40479	0.004838	0.40376	0.004928

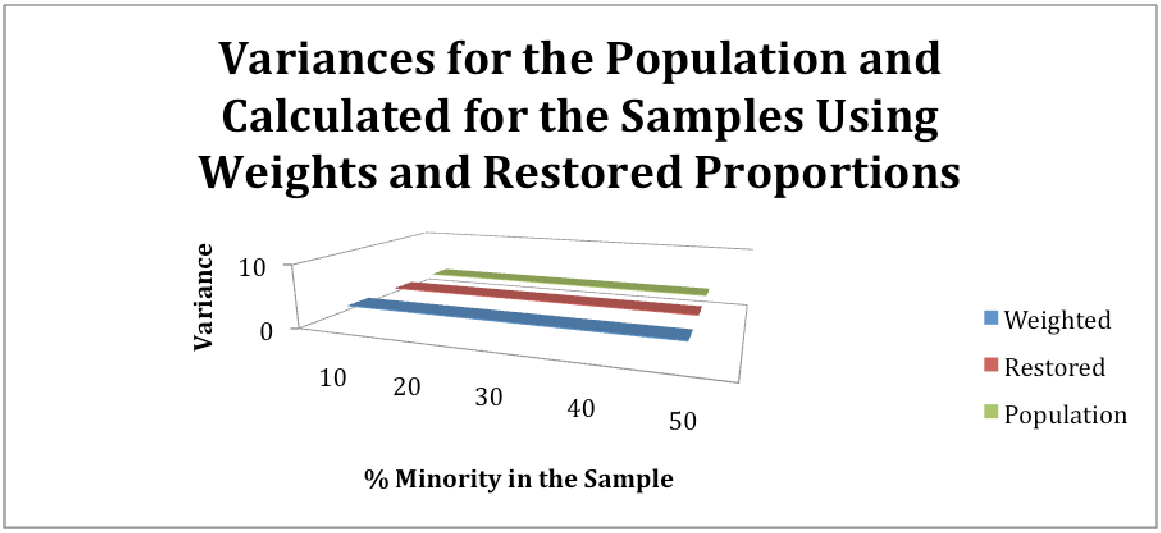


The values for the first moment, the mean, indicate near agreement between the weights method and the restored proportions method for all five cases. This says that the correct mean of a disproportionately sampled population can be calculated without the use of weights by creating a proportionately correct super sample. The import of this is that datasets containing disproportionately sampled data could be constructed, used and published by researchers that do not require the use of weights. This is advantageous in that it will allow casual users of the data to calculate the mean without considering the disproportionate design of the sample. Thus, the design of the sample need not be considered nor must specialized software or programming be used to obtain the correct mean for samples constructed in this manner.

Table 4.4b

Mean Variance for the 100 Populations and Calculated for the Samples using Weights and Restored Proportions

Mean Variance	Percent Minority									
	10 Mean	10 SE	20 Mean	20 SE	30 Mean	30 SE	40 Mean	40 SE	50 Mean	50 SE
Population	3.2436	0.002287	3.2436	0.002287	3.2436	0.002287	3.2436	0.002287	3.2436	0.002287
Weighted	3.23557	0.03676	3.24601	0.024145	3.25772	0.021793	3.24648	0.019048	3.26069	0.016472
Restored	3.2332	0.036751	3.24601	0.024145	3.25391	0.021624	3.24377	0.019032	3.25873	0.016462

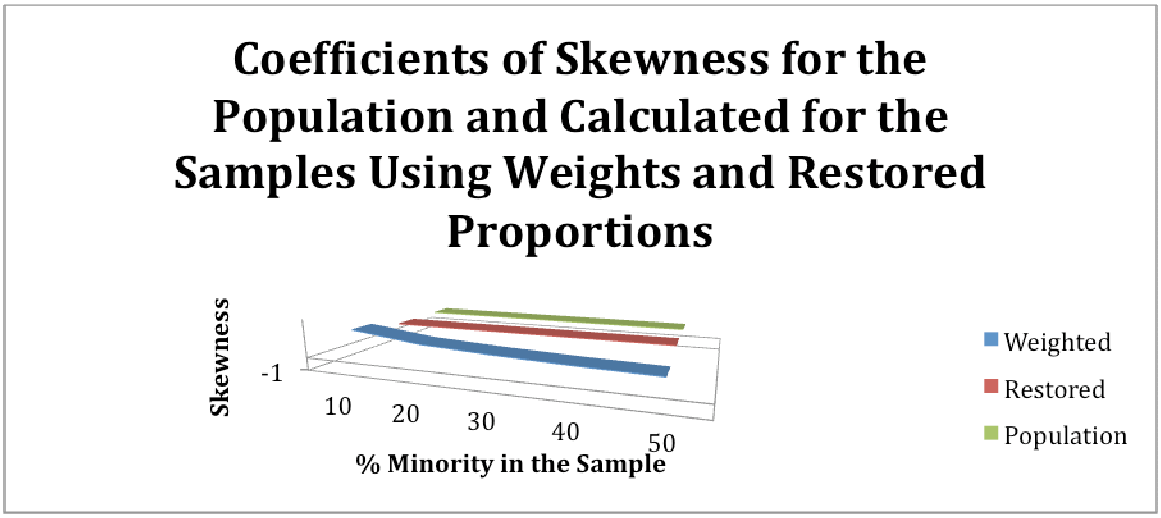


The values for the second moment, the variance, are also in near agreement for the two methods for the single proportionate sample (80:20) and are in similar agreement for the two methods for the four disproportionate samples. Note that because of the large sample sizes involved in this simulation, the degrees of freedom do not need to be adjusted to achieve satisfactory results. For smaller sample sizes such that the proportional difference between *N* and *n-1* are greater, this will not be the case as replication does, in fact, inflate the variance. As such, this must be considered a limitation of the method.

Table 4.4c

Skewness for the Population and Calculated for the Samples using Weights and Restored Proportions

Mean Skewness	Percent Minority									
	10 Mean	10 SE	20 Mean	20 SE	30 Mean	30 SE	40 Mean	40 SE	50 Mean	50 SE
Population	1.44911	0.002134	1.44911	0.002134	1.44911	0.002134	1.44911	0.002134	1.44911	0.002134
Weighted	2.09299	0.035173	1.42536	0.018467	1.13924	0.016826	0.94325	0.009952	0.81275	0.008065
Restored	1.41952	0.024821	1.42536	0.018467	1.4349	0.016258	1.44245	0.013901	1.45794	0.012338



The values for the third moment, the coefficient of skewness, are in near agreement for the single proportionate sample (80:20). In the case of all four disproportionate samples, the four restored proportion samples produce a coefficient of skewness that is much closer to the population parameter than the values produced by the weights method. While this is not of importance to analysis techniques that are completely parameterized by the first two moments, it may allow more reliable analyses using methods such as factor analysis, discriminant analysis, and other multivariate techniques. Thus, the restored proportions method produces a statistically superior result when compared to the weights method.

Summary

This chapter contains four findings. First, it is shown that “weighted data” cannot be simply unweighted, or “reweighted” by some inverse factor of the weights. There are 2 reasons for this. The first reason, demonstrated by the results, is that “weighted data” are commonly not weighted by any multiplicative factor contained in the data value. Instead, the weights are calculated by the researcher and supplied alongside the data for use in the appropriate analysis. Thus to factor out a value that was never factored in is an absurd act leading to absurd results. The second reason that weights cannot simply be factored out is that, in the case of data values that do contain a multiplicative weighting factor, to remove this factor would be to accept the bias in the data that the weights were created to correct.

Second, it is found that even in the case of a disproportionate sample, the first three moments of the distribution can be correctly calculated under the assumptions of simple random sampling. This means that proportion correcting weights can safely be ignored if the researcher is only concerned with analysis of one level of the grouping variable upon which the weights were created.

Third, it is found that the proportions of a disproportionately sampled population can be restored by replicating the majority and minority portions of disproportionate samples and that the first three moments of the distribution can be correctly estimated using the super sample that results from the concatenation of these replications. Further, it is found that the coefficient of skewness is nearer the population parameter for this moment when calculated using the restored proportions method than with the weights.

CHAPTER V:

SUMMARY

This study considered probabilistically sampled data. Specifically, the problems inherent in the analysis of disproportionately sampled data were considered and a method was proposed to both eliminate the complexity of data analysis associated with the weights method of analysis and the statistical disadvantages of weights, particularly with regard to variance inflation.

Results from the first research question show that "weighted data" may not be simply unweighted by using a reciprocal factor of the weights. Firstly, disproportionately sampled data are typically not distributed as weighted data values. Instead, the weights are calculated and supplied for use in the analysis. Thus, to "unweight" data that were never weighted would be an arbitrary and absurd act. Table 4.1 provides illustrates this with five counter examples.

Results from the second research question show that, yes, disproportionately sampled data can, in fact, be safely and correctly used without the weights provided the analysis is isolated to the members of a single value of the variable on which the weighting calculations were performed. For instance, if a population was disproportionately sampled with regard to two values of race (majority and minority), the researcher would be able to analyze data from either the majority group or the minority group under the assumptions of simple random sampling.

Results from the third research question show that, yes, population proportions can be restored using disproportionately sampled data. This is accomplished by finding the lowest integer quotient where the population proportion is the dividend and the sample proportion is the divisor. Table 4.2 enumerates the population proportion (80:20) and the proportions of the five

samples (90:10, 80:20, 70:30, 60:40, and 50:50). The same table shows the restored proportion (80:20) for each of these.

Results from the fourth research question show that not only can the proportionately restored super samples from Research Question 3 be used under the assumptions of simple random sampling to calculate the first three moment of a population distribution, but that the results for both the variance and the coefficient of skewness are superior (nearer the population parameters) than the corresponding numbers calculated by the weights method.

In summary, probabilistically sampled data may not be unweighted and appropriately analyzed without producing absurd results or ignoring the bias of the sample design that produced the data. The minority of majority portions of a disproportionate sample may be correctly analyzed separately for either group without the use of the weights. The proportions of a population can be restored using disproportionate samples and these samples can be concatenated and analyzed as if they were the result of simple random sampling.

Discussion

Compounding the issue of misuse of disproportionately sampled data is easy access to such data afforded researchers. A Jstor search (www.jstor.org) using the terms "NCES NHES data" returned 31 articles from peer reviewed journals. An identical search using the search term "NCES NHES data weights" returned only seven articles. A search for the term "weight" in each of these seven articles showed that only four of these actually used the sampling weights provided by NHES. One paper, for instance, generated a search result because of references to "birth weights." Still, two other papers use weights calculated by the authors for other purposes, but make no mention of the use of weights provided by NHES. Further, one of the papers

returned was a theory paper written by an NHES consultant (J. Michael Brick) on the very subject of sample bias and weights.

Fatalities and the Organization of Child Care in the United States, 1985-2003 uses the NHES data, but makes no mention of the use of weights (Wrigley & Dreby, 2005). As part of the study, the authors calculate fatality rates for different types of child care arrangements. The authors devote a section of the article to selection bias, yet do not mention the sampling bias present in the sample. This omission assumes that the number of child care fatalities is not influenced by any of the weighting factors included with the NHES data. Thus, the authors assume that fatalities occur at the same rates for different races. This may, in fact, be true, but the article in its present form cannot adequately address the question.

Early Child Development and Social Mobility also uses NHES data, but does not mention the use of weights in the analysis (Barnett and Belfield, 2006). On page 76, the authors "found that just 12 percent of young children had parents who reported participating in a parenting education program or support group." This finding assumes that participation in parenting education programs or support groups is unrelated to race or any of the other factors the weights were created to address. As race and income are known to be closely related, it may be true that persons with higher incomes are more likely to participate in such programs, introducing an unaddressed bias in the results.

As these two articles illustrate, the potential for erroneous results exists for researchers who use disproportionately sampled data without considering the sample design. If the weights are not used in the analysis, the results must be considered suspect.

This implies that of the 31 peer reviewed articles accessible by Jstor that mention the terms “NCES,” “NHES,” and “data,” only three empirical papers may have accounted for the sample design of the data by using the provided weights.

As a result, the analyses used in these papers must be considered suspect. Without the weights, it is unlikely that even the mean of the distributions are calculated correctly. Furthermore, incorrectly calculated variances will produce incorrect standard errors which will, in turn, produce erroneous inference decisions regarding those (likely incorrect) means. Given the fact that popular ordinary least squares (OLS) techniques such analysis of variance (ANOVA) and linear regression rely heavily on the mean and variance of a distribution, any analysis using these techniques must be considered suspect even if the remaining assumptions of the models are met.

If the sampling design is ignored, means, for instance, will be biased toward the oversampled population. That is, if the mean of the minority (oversampled) population is less that than the mean of the majority population, then the mean of the population will be underestimated if the data are analyzed under the assumptions of simple random sampling. Conversely, the means will be overestimated if the mean of the oversampled minority is greater than that of the majority portion of the sample. The variance and coefficient of skewness will be similarly underestimated or overestimated if the disproportionate sampling is not taken into account.

Limitations of the Study and Resulting Methods

A number of limitations must be noted. The study presented here is based on simulated data and is thus somewhat artificial in its approach. Real world data will likely not conform to known distributions. These deviations will likely produce deviation from the results above. The

implications of outliers and influence points, for instance, have not been addressed. Replication of a small minority sample containing outliers may exaggerate the effect of these outliers on the final results of any analysis using the restored proportions method. If the outliers lie in the direction of the majority population, differences between the two groups could be obscured. Conversely, outliers that lie further from the majority may serve to artificially exaggerate these differences.

Real world data, like the NHES data illustrated earlier in this dissertation, may use multiple weights to correct for biases other than oversampling. The alternate methods discussed here would find limited application in such cases. To analyze the minority portion of a sample as in Research Question 2 would be to ignore and thus accept biases other than that resulting from minority oversampling. In some cases, the weight associated with oversampling could be factored out using these methods, but the remaining weights would need to remain in use if a proper analysis is to be conducted.

The study presented here assumes a population majority:minority proportion of 80:20. In reality, the population proportion in the United States for racial minorities in nearer 70:30. Thus this may have been a wiser benchmark for researchers concerned with racial minorities. The technique of oversampling is not limited to research on racial minorities. Medical researchers interested in cancer rates, for instance, use oversampling to gain sufficient sample sizes of cancer patients. Research on any rare event could thus find benefit from oversampling.

The restored proportions method requires extant data and as such, it cannot be used to correct non-response bias.

Further Research

Many variables remain to be explored with regards to this study. Population proportions other than the 80:20 majority:minority example used here should be studied in a similar manner to generalize the findings of the 80:20 case. In particular, research is needed on the behavior of the distributions under these conditions in the presence of rare events, such as cancer and other medical conditions.

Additional means and variances should be considered. The populations represented in this study were generated to have a mean of 0 and a variance of 1 for the majority population and a mean of 2 and a variance of 9 for the minority population. Results may differ as these parameters change – particularly as the sample and population diverge.

In order to be useful, the methods should be applied to real world data such as the NHES dataset. While all of the weights cannot be eliminated in a dataset of this type, worthwhile statistical improvements may be attainable.

The data used here were generated to be normally distributed. Other distributions should be explored; in particular, heavily skewed distributions such as the Chi-squared distribution may produce contrary results.

REFERENCES

AM Statistical Software. (2011). *Manual*. Retrieved July 1, 2011, from http://am.air.org/help/JSTree/MainFrame.asp

Barnett, W. S., & Belfield, C. R. (2006). Early childhood development and social mobility. *The Future of Children*, 16(2), 73-98

Brick, J.M., Waksberg, J., Kulp, D., & Starer, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly, 59*(2), 218–235.

Brogan, D. J. (1998). *Pitfalls of using standard statistical software packages for sample survey data.* Retrieved on July 1, 2011, from http://www.hcp.med.harvard.edu/statistics/survey-soft/docs/donna_brogan.html

Casady, R.J., & Lepkowski, J.M. (1993). Stratified telephone survey designs. *Survey Methodology, 19*(1), 103–113.

Conway, S. (1982). *The weighting game.* Paper presented at the Market Research Society Conference, Metropole Hotel, Brighton.

Deming, W.E., & Stephan, F.F. (1940). On a least square adjustment of a sampled frequency table: When the expected marginal totals are known. *Annals of Mathematical Statistics, 11*, 427–444.

Deming, W. E. (1943). *Statistical adjustment of data.* New York: Wiley

Dorofeev, S., & Grant, P. (2006). *Statistics for real life sample surveys: Non-simple-random samples and weighted data*. New York: Cambridge University Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1-26

Hagedorn, M., Roth, S.B., O'Donnell, K. Smith, S., & Mulligan, G. (2008). *National Household Education Surveys Program of 2007: Data File User's Manual, Volume I.* (NCES 2009-024). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Hagedorn, M., O'Donnell, K., Smith, S., & Mulligan, G. (2008). *National Household Education Surveys Program of 2007: Data File User's Manual, Volume III, Parent and Family Involvement in Education Survey.*(NCES 2009-024). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Hagedorn, M., Roth, S. B., Carver, P., Van de Kerckhove, W., & Smith, S. (2009). *National Household Education Surveys Program of 2007: Methodology Report.* (NCES 2009-047). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Kim, J. K., Navarro, A., & Fuller, W. (2000). *Variance estimation for 2000 Census Coverage Estimates.* Proceedings of the Survey Research Methods Section, American Statistical Association, Alexandria, VA.

Kish, L. (1995). *Questions/answers from The Survey Statistician, 1978-1994.* Paris: The International Association of Survey Statisticians, Section of the International Statistical Institute.

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

Kim, J, Murdock, T. & Choi, D. (2005). Investigation of parents' beliefs about readiness for kindergarten: An examination of National Household Education Survey (NHES: 93). *Educational Research Quarterly, 29*(2), 3-17.

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

Landis, R.J., Lepkowski, J.M., Eklund, S.A., & Stehouwer, S.A. (1982). *A statistical methodology for analyzing data from a complex survey: The first national health and nutrition examination survey (DHHS Pub. No. 82-1366).* Vital and Health Statistics,Series 2, No. 92. Washington, DC: National Center for Health Statistics.

Moser, C. A., & Kalton, G. (1971). Survey methods in social investigation (2nd ed.). London: Heinemann Educational Books.

Quenouille, M. (1949). Approximation tests of correlation in time series. *Journal of the Royal Statistical Society B, 11*, 18-84.

Shao, J., & Tu, D. (1995). *The jackknife and bootstrap, Springer series in statistics.* New York: Springer-Verlag.

Sharot, T. (1986,) Weighting survey results. *Journal of Market Research Society, 28*(3), 269-284.

Tucker, C., Lepkowski, J.M., & Piekarski, L. (2002). The current efficiency of list-assisted telephone sampling designs. *Public Opinion Quarterly, 66*(3), 321-338.

Wild, C. J., Seber, & George, A. F. (2000). *Chance encounters: A first course in data analysis and inference.* New York: John Wiley & Sons.

Wrigley, J., & Dreby, J. (2005). Fatalities and the organization of child care in the United States, 1985-2003. *American Sociological Review*, 70(5), 729-757.

Appendix A

Instructions for Data Access and Analysis of the

NHES:2007 PFI using PC SAS 9.2 and AM software

1. Go to http://nces.ed.gov/nhes/dataproducts.asp#2007dp.
2. Under "Accessing NHES 1999-2007 Public Use Data through Direct Downloads," Click on "2007 data products."
3. Click on "PFI data." Download and save. Note the location.
4. Click on "SAS setup file" and agree to the terms of use. The document will automatically open.
5. Select the entire document using "Ctrl – A."
6. Right click anywhere in the document choose "copy."
7. Open SAS 9.2
8. Right click in the editor window and choose "paste"
9. Change the path in the libname statement to the preferred destination folder for the SAS dataset that will be created by this program.
10. Change the path in the filename statement to the location of the folder containing the data file downloaded in step 3. Note – the path in this statement must contain the file name, such as "pfi07asc.dat."
11. Click "Run" and check the SAS Log for errors.
12. If the program ran successfully, a SAS dataset will appear in the folder listed in the libname statement. Close SAS.
13. Return to the NCES Data Products page. Near the top of the page, click on the "EDAT application" and agree to the terms. First time users will have to register and create a password.
14. Login to the EDAT application. Choose NHES:2007, and click "Next."
15. Choose the 2007 "Parent and Family Involvement in Education Catalog" and click "Next."
16. Under "TAG FILE OPTIONS," choose "New Tag File," name the new tag file and click "Create."
17. In the Variable Search panel on the left side of the screen, choose the variables to be included in the analysis. For instance, click on "ID Variables" to reveal the variable available in that group. A padlock symbol indicates a mandatory variable. Mandatory variables are automatically included in any analysis.
18. Once all of the variables of interest from a particular group are selected, click on the next group, repeating the process until the entire list is exhausted.
19. Under "Download Options," choose "Download Data and Syntax Files." When the Step 1 information window appears, click "Next Step."

20. Choose “SAS” from the Step 2 window and click “Next Step.”
21. Create a folder for the SAS dataset, preferably c:\EDAT\NHES.
22. Click the orange download button to begin the download.
23. The window containing the new SAS dataset will automatically open. Move the files to the c:\EDAT\NHES folder.
24. Point a browser to http://am.air.org/ and click “Download.”
25. If you have not registered with AIR, do so now. Otherwise, sign in.
26. Click on “Download AM Software Version x.xx.xx.” Open the folder and run the AMinstall.exe file. Follow the direction for installing the software.
27. Note the location of the shortcut to the AM software.
28. Next, click on the “AM’s Data Transfer Component” under download link is step 26 and run the .exe file as before.
29. Double click on the AM software shortcut.
30. Click on the “Manual” button on the AIR website and follow the instructions.
31. Under the File menu, choose “Import” and then “General Import.”
32. Choose the folder containing the SAS dataset created earlier.
33. In the “Files of type” drop down menu, choose “SAS for Windows V7/8/9 (*.sas7bdat)”
34. If a Dialog Box appears regarding user define formats, click “Done.”
35. The Variables window should be populated with variables from the PFI.
36. Begin the analysis by Choosing Statistics, Replication Procedures For Basic Statistics, Regression.
37. Drag the dependent variable from the variable list to the dependent variable window in the regression dialog box.
38. Drag the independent variables from the variable list to the independent variable window in the regression dialog box.
39. Choose “JK1” for the PFI
40. Drag the 80 replicate weights, FPWT1-FPWT80 into the Replicate Weights Box.
41. Click “OK”

Appendix B

The SAS Code for Example in Chapter III

```
*NOTE:To change the pop size,
change n=xxx in first 2 data steps,
plus racecount le xxx on line 70;

*OPTIONS PAGESIZE=150 LINESIZE=133;
*First, we create a dataset for the 7500 majority members
of the population, arbitrarily distributed N(12,1);

data majority;
n=7500;
call streaminit(123);
do i=1 to n;
y = RAND('NORMAL',12,1);
x1 = RAND('NORMAL',12,1);
x2 = RAND('NORMAL',12,1);
x3 = RAND('NORMAL',12,1);
race=1;
output;
end;
drop n i;
run;

*Next, we create a dataset for the 2500 minority members of the
population, arbitrarily distributed differently at N(10,2);

data minority;
n=2500;
call streaminit(123);
do i=1 to n;
y = RAND('NORMAL',10,2);
x1 = RAND('NORMAL',10,2);
x2 = RAND('NORMAL',10,2);
x3 = RAND('NORMAL',10,2);
race=2;
output;
end;
drop n i;
run;

*Now append (stack) the two datasets to create the population
from which we will oversample one of the groups;
```

```
proc append base=majority data=minority;
run;

* At this point, we have a population with 3/4 proportion
of majority and a 1/4 proportion of minority;

* We now sort the dataset randomly and take the first 2500
majority members and the first 2500 minority members. At this
point, we have a sample from the original population that
is 1/2 and 1/2, so the minorities are oversampled with regard
to the population proportions;

data population; set majority;
call streaminit(123);
random = RAND('NORMAL',0,1);
run;

proc sort data = population;
by race random;
run;

data disproportionate;
set population;
by race;
do;
if first.race then racecount=0;
racecount+1;
end;
if racecount le 2500;
run;

*proc print;
*run;

*Now,restore the majority and minority proportions by duplicating
the undersampled majority and appending the datasets to create a
representative sample consistent with the population proportions;


data minority1;
set disproportionate;
if race=2;

data duplicate1;
set disproportionate;
```

```
if race=1;

data duplicate2;
set disproportionate;
if race=1;

data duplicate3;
set disproportionate;
if race=1;

Proc append base=minority1 data=duplicate1;  run;
Proc append  base=minority1 data=duplicate2; run;
Proc append  base=minority1 data=duplicate2; run;

Data recpop;
Set minority1;
dataset=2;
proc sort;
by race;
run;
*proc print;
*run;

proc means data=population n mean var stderr probt;
var y x1 x2 x3;
run;
proc means data=recpop n mean var stderr probt;
var y x1 x2 x3;
run;
```

Appendix C

The SAS Code for One Iteration of Results in Chapter IV

```
*NOTE:To change the pop size,
change n=xxx in first 2 data steps;

*First, we create a dataset for the 80000 majority members
of the population, arbitrarily distributed N(0,1);

data majority;
n=80000;
call streaminit(111);
do i=1 to n;
x = RAND('NORMAL',0,1);
majority=1;
output;
end;
drop n i;
run;

*Next, we create a dataset for the 20000 minority members of the
population, distributed differently at an arbitrary N(2,3);

data minority;
n=20000;
call streaminit(111);
do i=1 to n;
x = RAND('NORMAL',2,3);
majority=0;
output;
end;
drop n i;
run;


*Now append (concatenate) the two datasets to create the population
from which we will oversample one of the groups. Also create a count variable
within each level of majority to be used in subsetting below;

proc append base=majority data=minority;
run;

proc sort;
by majority;
run;
```

```
data population;
set majority;
by majority;
do;
if first.majority then majcount=0;
majcount+1;
end;
if majority=0 then target=20000; else target=80000;
if majority=0 then observed=20000; else observed=80000;
weight=target/observed;
run;

* At this point, we have a population with a 4/5 proportion
of majority and a 1/5 proportion of minority. That is, 4:1 Maj:Min. ;

*Now take a 90:10 sample of the population. The observations are already
randomly sorted, so we can just take them in the order they appear in the
population. ;

data majmin9010;
set population;
     if majority = 1 and (0 < majcount le 9000) then keeper=1;
else if majority = 0 and (0 < majcount le 1000) then keeper=1;
if keeper=1;
if majority=0 then target=2000; else target=8000;
if majority=0 then observed=1000; else observed=9000;
weight=target/observed;
dataset="majmin9010";
run;

*Now take a 80:20 sample of the population. The observations are already
randomly sorted,
so we can just take them in the order they appear in the population. ;

data majmin8020;
set population;
     if majority = 1 and (0 < majcount le 8000) then keeper=1;
else if majority = 0 and (0 < majcount le 2000) then keeper=1;
if keeper=1;
if majority=0 then target=2000; else target=8000;
if majority=0 then observed=2000; else observed=8000;
weight=target/observed;
dataset="majmin8020";
run;

*Now take a 70:30 sample of the population. The observations are already
randomly sorted, so we can just take them in the order they appear in the
population. ;

data majmin7030;
set population;
     if majority = 1 and (0 < majcount le 7000) then keeper=1;
else if majority = 0 and (0 < majcount le 3000) then keeper=1;
if keeper=1;
if majority=0 then target=2000; else target=8000;
if majority=0 then observed=3000; else observed=7000;
weight=target/observed;
```

```
dataset="majmin7030";
run;

*Now take a 60:40 sample of the population. The observations are already
randomly sorted, so we can just take them in the order they appear in the
population. ;

data majmin6040;
set population;
     if majority = 1 and (0 < majcount le 6000) then keeper=1;
else if majority = 0 and (0 < majcount le 4000) then keeper=1;
if keeper=1;
if majority=0 then target=2000; else target=8000;
if majority=0 then observed=4000; else observed=6000;
weight=target/observed;
dataset="majmin6040";
run;

*Now take a 50:50 sample of the population. The observations are already
randomly sorted, so we can just take them in the order they appear in the
population. ;

data majmin5050;
set population;
     if majority = 1 and (0 < majcount le 5000) then keeper=1;
else if majority = 0 and (0 < majcount le 5000) then keeper=1;
if keeper=1;
if majority=0 then target=2000; else target=8000;
if majority=0 then observed=5000; else observed=5000;
weight=target/observed;
dataset="majmin5050";
run;

* Begin RQ1 here. Use all 5 samples to demonstrate.;

data RQ1_9010;set
majmin9010;unweight=1/weight;dataset="RQ1_9010";xunweighted=x*unweight;run;
data RQ1_8020;set
majmin8020;unweight=1/weight;dataset="RQ1_8020";xunweighted=x*unweight;run;
data RQ1_7030;set
majmin7030;unweight=1/weight;dataset="RQ1_7030";xunweighted=x*unweight;run;
data RQ1_6040;set
majmin6040;unweight=1/weight;dataset="RQ1_6040";xunweighted=x*unweight;run;
data RQ1_5050;set
majmin5050;unweight=1/weight;dataset="RQ1_5050";xunweighted=x*unweight;run;

* Calculate the first three moments of the RQ1
first with the weights (RQ1w), then using the unweighting factor (RQ1u);

proc univariate data=RQ1_9010 vardef=df noprint;
output out=momentsRQ1_9010w mean=mean var=var skewness=skew;
var x;weight weight;run;

proc univariate data=RQ1_9010 vardef=df noprint;
output out=momentsRQ1_9010u mean=mean var=var skewness=skew;
var xunweighted;run;
```

```
proc univariate data=RQ1_8020 vardef=df noprint;
output out=momentsRQ1_8020w mean=mean var=var skewness=skew;
var x;weight weight;run;

proc univariate data=RQ1_8020 vardef=df noprint;
output out=momentsRQ1_8020u mean=mean var=var skewness=skew;
var xunweighted;run;

proc univariate data=RQ1_7030 vardef=df noprint;
output out=momentsRQ1_7030w mean=mean var=var skewness=skew;
var x;weight weight;run;

proc univariate data=RQ1_7030 vardef=df noprint;
output out=momentsRQ1_7030u mean=mean var=var skewness=skew;
var xunweighted;run;

proc univariate data=RQ1_6040 vardef=df noprint;
output out=momentsRQ1_6040w mean=mean var=var skewness=skew;
var x;weight weight;run;

proc univariate data=RQ1_6040 vardef=df noprint;
output out=momentsRQ1_6040u mean=mean var=var skewness=skew;
var xunweighted;run;

proc univariate data=RQ1_5050 vardef=df noprint;
output out=momentsRQ1_5050w mean=mean var=var skewness=skew;
var x;weight weight;run;

proc univariate data=RQ1_5050 vardef=df noprint;
output out=momentsRQ1_5050u mean=mean var=var skewness=skew;
var xunweighted;run;

data momentsRQ1_9010w; set momentsRQ1_9010w; dataset="momentsRQ1_9010w";run;
data momentsRQ1_9010u; set momentsRQ1_9010u; dataset="momentsRQ1_9010u";run;
data momentsRQ1_8020w; set momentsRQ1_8020w; dataset="momentsRQ1_8020w";run;
data momentsRQ1_8020u; set momentsRQ1_8020u; dataset="momentsRQ1_8020u";run;
data momentsRQ1_7030w; set momentsRQ1_7030w; dataset="momentsRQ1_7030w";run;
data momentsRQ1_7030u; set momentsRQ1_7030u; dataset="momentsRQ1_7030u";run;
data momentsRQ1_6040w; set momentsRQ1_6040w; dataset="momentsRQ1_6040w";run;
data momentsRQ1_6040u; set momentsRQ1_6040u; dataset="momentsRQ1_6040u";run;
data momentsRQ1_5050w; set momentsRQ1_5050w; dataset="momentsRQ1_5050w";run;
data momentsRQ1_5050u; set momentsRQ1_5050u; dataset="momentsRQ1_5050u";run;


Proc append base=momentsRQ1_9010w data=momentsRQ1_9010u;  run;
Proc append base=momentsRQ1_9010w data=momentsRQ1_8020w;  run;
Proc append base=momentsRQ1_9010w data=momentsRQ1_8020u;  run;
Proc append base=momentsRQ1_9010w data=momentsRQ1_7030w;  run;
Proc append base=momentsRQ1_9010w data=momentsRQ1_7030u;  run;
Proc append base=momentsRQ1_9010w data=momentsRQ1_6040w;  run;
Proc append base=momentsRQ1_9010w data=momentsRQ1_6040u;  run;
Proc append base=momentsRQ1_9010w data=momentsRQ1_5050w;  run;
Proc append base=momentsRQ1_9010w data=momentsRQ1_5050u;  run;

data RQ1; set momentsRQ1_9010w; run;
proc print data=RQ1 noobs;
var dataset mean var skew;
```

```
run;

* Begin RQ2 here.;

*First get the population moments for both values of MAJORITY;

data maj; set population; if majority=1;run;
data min; set population; if majority=0;run;

proc univariate data=maj vardef=df noprint;
output out=majpopmoments mean=mean var=var skewness=skew stdmean=se n=n;
var x;
run;
data majpopmoments; set majpopmoments; dataset="majpopmoments
";majority=1; run;

proc univariate data=min vardef=df noprint;
output out=minpopmoments mean=mean var=var skewness=skew; * stdmean=se n=n;
var x;
run;
data minpopmoments; set minpopmoments; dataset="minpopmoments
";majority=0; run;

*Now calculate the moments of the 9010 sample seperately for both values of
MAJORITY using a by statement and the weights that were calculated for the
entire sample;

proc univariate data=majmin9010 vardef=df noprint;
output out=momentsRQ2_9010w_by mean=mean skewness=skew stdmean=se n=n;
by majority;
var x; weight weight; run;

data momentsRQ2_9010w_by; set momentsRQ2_9010w_by;
dataset="momentsRQ2_9010w_by";
var=(se*sqrt(n))**2; drop se n; run;

*Now calculate the moments of the 9010 sample separately for both values of
MAJORITY;

data maj9010; set majmin9010;if majority=1;run;
data min9010; set majmin9010;if majority=0;run;

proc univariate data=maj9010 vardef=df noprint;
output out=momentsRQ2_maj9010 mean=mean var=var skewness=skew;
var x;
run;
proc univariate data=min9010 vardef=df noprint;
output out=momentsRQ2_min9010 mean=mean var=var skewness=skew;
var x;
run;

data momentsRQ2_maj9010; set momentsRQ2_maj9010; dataset="momentsRQ2_maj9010
";majority=1; run;
data momentsRQ2_min9010; set momentsRQ2_min9010; dataset="momentsRQ2_min9010
";majority=0; run;
```

```
*Now calculate the moments of the 8020 sample separately for both values of
MAJORITY using the weights that were calculated for the entire sample;

proc univariate data=majmin8020 vardef=df noprint;
output out=momentsRQ2_8020w_by mean=mean skewness=skew stdmean=se n=n;
by majority;
var x; weight weight; run;

data momentsRQ2_8020w_by; set momentsRQ2_8020w_by;
dataset="momentsRQ2_8020w_by";
var=(se*sqrt(n))**2; drop se n; run;

*Now calculate the moments of the 8020 sample separately for both values of
MAJORITY;

data maj8020; set majmin8020;if majority=1;run;
data min8020; set majmin8020;if majority=0;run;

proc univariate data=maj8020 vardef=df noprint;
output out=momentsRQ2_maj8020 mean=mean var=var skewness=skew;
var x;
run;
proc univariate data=min8020 vardef=df noprint;
output out=momentsRQ2_min8020 mean=mean var=var skewness=skew;
var x;
run;

data momentsRQ2_maj8020; set momentsRQ2_maj8020; dataset="momentsRQ2_maj8020
";majority=1; run;
data momentsRQ2_min8020; set momentsRQ2_min8020; dataset="momentsRQ2_min8020
";majority=0; run;

*Now calculate the moments of the 7030 sample separately for both values of
MAJORITY using the weights that were calculated for the entire sample;

proc univariate data=majmin7030 vardef=df noprint;
output out=momentsRQ2_7030w_by mean=mean skewness=skew stdmean=se n=n;
by majority;
var x; weight weight; run;

data momentsRQ2_7030w_by; set momentsRQ2_7030w_by;
dataset="momentsRQ2_7030w_by";
var=(se*sqrt(n))**2; drop se n; run;

*Now calculate the moments of the 7030 sample separately for both values of
MAJORITY;

data maj7030; set majmin7030;if majority=1;run;
data min7030; set majmin7030;if majority=0;run;

proc univariate data=maj7030 vardef=df noprint;
output out=momentsRQ2_maj7030 mean=mean var=var skewness=skew;
var x;
run;
proc univariate data=min7030 vardef=df noprint;
output out=momentsRQ2_min7030 mean=mean var=var skewness=skew;
var x;
```

```
run;

data momentsRQ2_maj7030; set momentsRQ2_maj7030; dataset="momentsRQ2_maj7030
";majority=1; run;
data momentsRQ2_min7030; set momentsRQ2_min7030; dataset="momentsRQ2_min7030
";majority=0; run;

*Now calculate the moments of the 6040 sample separately for both values of
MAJORITY using the weights that were calculated for the entire sample;

proc univariate data=majmin6040 vardef=df noprint;
output out=momentsRQ2_6040w_by mean=mean skewness=skew stdmean=se n=n;
by majority;
var x; weight weight; run;

data momentsRQ2_6040w_by; set momentsRQ2_6040w_by;
dataset="momentsRQ2_6040w_by";
var=(se*sqrt(n))**2; drop se n; run;

*Now calculate the moments of the 6040 sample separately for both values of
MAJORITY;

data maj6040; set majmin6040;if majority=1;run;
data min6040; set majmin6040;if majority=0;run;

proc univariate data=maj6040 vardef=df noprint;
output out=momentsRQ2_maj6040 mean=mean var=var skewness=skew;
var x;
run;
proc univariate data=min6040 vardef=df noprint;
output out=momentsRQ2_min6040 mean=mean var=var skewness=skew;
var x;
run;

data momentsRQ2_maj6040; set momentsRQ2_maj6040; dataset="momentsRQ2_maj6040
";majority=1; run;
data momentsRQ2_min6040; set momentsRQ2_min6040; dataset="momentsRQ2_min6040
";majority=0; run;

*Now calculate the moments of the 5050 sample separately for both values of
MAJORITY using the weights that were calculated for the entire sample;

proc univariate data=majmin5050 vardef=df noprint;
output out=momentsRQ2_5050w_by mean=mean skewness=skew stdmean=se n=n;
by majority;
var x; weight weight; run;

data momentsRQ2_5050w_by; set momentsRQ2_5050w_by;
dataset="momentsRQ2_5050w_by";
var=(se*sqrt(n))**2; drop se n; run;

*Now calculate the moments of the 5050 sample separately for both values of
MAJORITY;

data maj5050; set majmin5050;if majority=1;run;
data min5050; set majmin5050;if majority=0;run;
```

```
proc univariate data=maj5050 vardef=df noprint;
output out=momentsRQ2_maj5050 mean=mean var=var skewness=skew;
var x;
run;
proc univariate data=min5050 vardef=df noprint;
output out=momentsRQ2_min5050 mean=mean var=var skewness=skew;
var x;
run;

data momentsRQ2_maj5050; set momentsRQ2_maj5050; dataset="momentsRQ2_maj5050
";majority=1; run;
data momentsRQ2_min5050; set momentsRQ2_min5050; dataset="momentsRQ2_min5050
";majority=0; run;

* Now append to a single dataset for easier comparisons;

Proc append base=majpopmoments data=minpopmoments; run;

Proc append base=majpopmoments data=momentsRQ2_9010w_by; run;
Proc append base=majpopmoments data=momentsRQ2_maj9010; run;
Proc append base=majpopmoments data=momentsRQ2_min9010; run;

Proc append base=majpopmoments data=momentsRQ2_8020w_by; run;
Proc append base=majpopmoments data=momentsRQ2_maj8020; run;
Proc append base=majpopmoments data=momentsRQ2_min8020; run;

Proc append base=majpopmoments data=momentsRQ2_7030w_by; run;
Proc append base=majpopmoments data=momentsRQ2_maj7030; run;
Proc append base=majpopmoments data=momentsRQ2_min7030; run;

Proc append base=majpopmoments data=momentsRQ2_6040w_by; run;
Proc append base=majpopmoments data=momentsRQ2_maj6040; run;
Proc append base=majpopmoments data=momentsRQ2_min6040; run;

Proc append base=majpopmoments data=momentsRQ2_5050w_by; run;
Proc append base=majpopmoments data=momentsRQ2_maj5050; run;
Proc append base=majpopmoments data=momentsRQ2_min5050; run;

proc sort; by majority;

data RQ2; set majpopmoments;

proc print data=RQ2 noobs;
var dataset majority mean var skew;
run;


*RQ3 Begins here;

*Now, restore the majority and minority proportions by duplicating
the undersampled majority and appending the datasets to create a
representative sample consistent with the population proportions;

*First restore the 90:10 sample;
data maj9010_1; set majmin9010; if majority=1; run;
data maj9010_2; set majmin9010; if majority=1; run;
data maj9010_3; set majmin9010; if majority=1; run;
```

```
data maj9010_4; set majmin9010; if majority=1; run;
data min9010_1; set majmin9010; if majority=0; run;
data min9010_2; set majmin9010; if majority=0; run;
data min9010_3; set majmin9010; if majority=0; run;
data min9010_4; set majmin9010; if majority=0; run;
data min9010_5; set majmin9010; if majority=0; run;
data min9010_6; set majmin9010; if majority=0; run;
data min9010_7; set majmin9010; if majority=0; run;
data min9010_8; set majmin9010; if majority=0; run;
data min9010_9; set majmin9010; if majority=0; run;


Proc append base=maj9010_1 data=maj9010_2;  run;
Proc append base=maj9010_1 data=maj9010_3;  run;
Proc append base=maj9010_1 data=maj9010_4;  run;
Proc append base=maj9010_1 data=min9010_1;  run;
Proc append base=maj9010_1 data=min9010_2;  run;
Proc append base=maj9010_1 data=min9010_3;  run;
Proc append base=maj9010_1 data=min9010_4;  run;
Proc append base=maj9010_1 data=min9010_5;  run;
Proc append base=maj9010_1 data=min9010_6;  run;
Proc append base=maj9010_1 data=min9010_7;  run;
Proc append base=maj9010_1 data=min9010_8;  run;
Proc append base=maj9010_1 data=min9010_9;  run;

data restored9010; set maj9010_1; run;


*Now restore the 80:20 sample;
data maj8020_1; set majmin8020; if majority=1; run;
data min8020_1; set majmin8020; if majority=0; run;

Proc append base=maj8020_1 data=min8020_1;  run;

data restored8020; set maj8020_1; run;

*Now restore the 70:30 sample;
data maj7030_1;  set majmin7030; if majority=1; run;
data maj7030_2;  set majmin7030; if majority=1; run;
data maj7030_3;  set majmin7030; if majority=1; run;
data maj7030_4;  set majmin7030; if majority=1; run;
data maj7030_5;  set majmin7030; if majority=1; run;
data maj7030_6;  set majmin7030; if majority=1; run;
data maj7030_7;  set majmin7030; if majority=1; run;
data maj7030_8;  set majmin7030; if majority=1; run;
data maj7030_9;  set majmin7030; if majority=1; run;
data maj7030_10; set majmin7030; if majority=1; run;
data maj7030_11; set majmin7030; if majority=1; run;
data maj7030_12; set majmin7030; if majority=1; run;
data min7030_1; set majmin7030; if majority=0; run;
data min7030_2; set majmin7030; if majority=0; run;
data min7030_3; set majmin7030; if majority=0; run;
data min7030_4; set majmin7030; if majority=0; run;
data min7030_5; set majmin7030; if majority=0; run;
data min7030_6; set majmin7030; if majority=0; run;
data min7030_7; set majmin7030; if majority=0; run;
```

```
Proc append base=maj7030_1 data=maj7030_2;  run;
Proc append base=maj7030_1 data=maj7030_3;  run;
Proc append base=maj7030_1 data=maj7030_4;  run;
Proc append base=maj7030_1 data=maj7030_5;  run;
Proc append base=maj7030_1 data=maj7030_6;  run;
Proc append base=maj7030_1 data=maj7030_7;  run;
Proc append base=maj7030_1 data=maj7030_8;  run;
Proc append base=maj7030_1 data=maj7030_9;  run;
Proc append base=maj7030_1 data=maj7030_10;  run;
Proc append base=maj7030_1 data=maj7030_11;  run;
Proc append base=maj7030_1 data=maj7030_12;  run;
Proc append base=maj7030_1 data=min7030_1;  run;
Proc append base=maj7030_1 data=min7030_2;  run;
Proc append base=maj7030_1 data=min7030_3;  run;
Proc append base=maj7030_1 data=min7030_4;  run;
Proc append base=maj7030_1 data=min7030_5;  run;
Proc append base=maj7030_1 data=min7030_6;  run;
Proc append base=maj7030_1 data=min7030_7;  run;

data restored7030; set maj7030_1; run;

*Now restore the 60:40 sample;

data maj6040_1;  set majmin6040; if majority=1; run;
data maj6040_2;  set majmin6040; if majority=1; run;
data maj6040_3;  set majmin6040; if majority=1; run;
data maj6040_4;  set majmin6040; if majority=1; run;
data maj6040_5;  set majmin6040; if majority=1; run;
data maj6040_6;  set majmin6040; if majority=1; run;
data maj6040_7;  set majmin6040; if majority=1; run;
data maj6040_8;  set majmin6040; if majority=1; run;
data min6040_1;  set majmin6040; if majority=0; run;
data min6040_2;  set majmin6040; if majority=0; run;
data min6040_3;  set majmin6040; if majority=0; run;

Proc append base=maj6040_1 data=maj6040_2;  run;
Proc append base=maj6040_1 data=maj6040_3;  run;
Proc append base=maj6040_1 data=maj6040_4;  run;
Proc append base=maj6040_1 data=maj6040_5;  run;
Proc append base=maj6040_1 data=maj6040_6;  run;
Proc append base=maj6040_1 data=maj6040_7;  run;
Proc append base=maj6040_1 data=maj6040_8;  run;
Proc append base=maj6040_1 data=min6040_1;  run;
Proc append base=maj6040_1 data=min6040_2;  run;
Proc append base=maj6040_1 data=min6040_3;  run;

data restored6040; set maj6040_1; run;

*Now restore the 50:50 sample;

data maj5050_1;  set majmin5050; if majority=1; run;
data maj5050_2;  set majmin5050; if majority=1; run;
data maj5050_3;  set majmin5050; if majority=1; run;
data maj5050_4;  set majmin5050; if majority=1; run;
data min5050_1;  set majmin5050; if majority=0; run;

Proc append base=maj5050_1 data=maj5050_2;  run;
```

```
Proc append base=maj5050_1 data=maj5050_3;  run;
Proc append base=maj5050_1 data=maj5050_4;  run;
Proc append base=maj5050_1 data=min5050_1;  run;

data restored5050; set maj5050_1; run;


* Produce the freqs to demonstrate the restored proportions;

proc freq data=majmin9010 noprint;
tables majority / out=majmin9010freq; run;
data majmin9010freq; set majmin9010freq; dataset="majmin9010freq  "; run;

proc freq data=restored9010 noprint;
tables majority / out=restored9010freq; run;
data restored9010freq; set restored9010freq; dataset="restored9010freq"; run;

proc freq data=majmin8020 noprint;
tables majority / out=majmin8020freq; run;
data majmin8020freq; set majmin8020freq; dataset="majmin8020freq  "; run;

proc freq data=restored8020 noprint;
tables majority / out=restored8020freq; run;
data restored8020freq; set restored8020freq; dataset="restored8020freq"; run;

proc freq data=majmin7030 noprint;
tables majority / out=majmin7030freq; run;
data majmin7030freq; set majmin7030freq; dataset="majmin7030freq  "; run;

proc freq data=restored7030 noprint;
tables majority / out=restored7030freq; run;
data restored7030freq; set restored7030freq; dataset="restored7030freq"; run;

proc freq data=majmin6040 noprint;
tables majority / out=majmin6040freq; run;
data majmin6040freq; set majmin6040freq; dataset="majmin6040freq  "; run;

proc freq data=restored6040 noprint;
tables majority / out=restored6040freq; run;
data restored6040freq; set restored6040freq; dataset="restored6040freq"; run;

proc freq data=majmin5050 noprint;
tables majority / out=majmin5050freq; run;
data majmin5050freq; set majmin5050freq; dataset="majmin5050freq  "; run;

proc freq data=restored5050 noprint;
tables majority / out=restored5050freq; run;
data restored5050freq; set restored5050freq; dataset="restored5050freq"; run;

data RQ3; set majmin9010freq; run;

proc append base=RQ3 data=restored9010freq; run;

proc append base=RQ3 data=majmin8020freq; run;
proc append base=RQ3 data=restored8020freq; run;

proc append base=RQ3 data=majmin7030freq; run;
```

```
proc append base=RQ3 data=restored7030freq; run;

proc append base=RQ3 data=majmin6040freq; run;
proc append base=RQ3 data=restored6040freq; run;

proc append base=RQ3 data=majmin5050freq; run;
proc append base=RQ3 data=restored5050freq; run;

proc print data=RQ3 noobs;
var dataset majority COUNT PERCENT;
run;


*RQ4 Begins here;

* Calculate the first three moments of the population
first without the weights, then using the weights to verify
the weights and calculations;

proc univariate data=population vardef=df noprint;
output out=popmoments mean=mean var=var skewness=skew;
var x;
run;
data popmoments; set popmoments; dataset="popmoments  ";run;

proc univariate data=population vardef=df noprint;
output out=popmomentsw mean=mean var=var skewness=skew;
var x;
weight weight;
run;
data popmomentsw; set popmomentsw; dataset="popmomentsw ";run;

* Calculate the first three moments of the majmin9010
first with the weights, then using my method;

proc univariate data=majmin9010 vardef=df noprint;
output out=moments9010w mean=mean var=var skewness=skew;
var x;
weight weight;
run;

data moments9010w; set moments9010w; dataset="moments9010w";run;

proc univariate data=restored9010 vardef=df noprint;
output out=moments9010r mean=mean var=var skewness=skew;
var x;
run;
data moments9010r; set moments9010r; dataset="moments9010r";run;

* Calculate the first three moments of the majmin8020
first with the weights, then using my method;

proc univariate data=majmin8020 vardef=df noprint;
output out=moments8020w mean=mean var=var skewness=skew;
var x;
weight weight;
run;
```

```
data moments8020w; set moments8020w; dataset="moments8020w";run;

proc univariate data=restored8020 vardef=df noprint;
output out=moments8020r mean=mean var=var skewness=skew;
var x;
run;
data moments8020r; set moments8020r; dataset="moments8020r";run;

* Calculate the first three moments of the majmin8020
first with the weights, then using my method;

proc univariate data=majmin7030 vardef=df noprint;
output out=moments7030w mean=mean var=var skewness=skew;
var x;
weight weight;
run;
data moments7030w; set moments7030w; dataset="moments7030w";run;

proc univariate data=restored7030 vardef=df noprint;
output out=moments7030r mean=mean var=var skewness=skew;
var x;
run;
data moments7030r; set moments7030r; dataset="moments7030r";run;

* Calculate the first three moments of the majmin6040
first with the weights, then using my method;

proc univariate data=majmin6040 vardef=df noprint;
output out=moments6040w mean=mean var=var skewness=skew;
var x;
weight weight;
run;

data moments6040w; set moments6040w; dataset="moments6040w";run;

proc univariate data=restored6040 vardef=df noprint;
output out=moments6040r mean=mean var=var skewness=skew;
var x;
run;
data moments6040r; set moments6040r; dataset="moments6040r";run;

* Calculate the first three moments of the majmin5050
first with the weights, then using my method;

proc univariate data=majmin5050 vardef=df noprint;
output out=moments5050w mean=mean var=var skewness=skew;
var x;
weight weight;
run;
data moments5050w; set moments5050w; dataset="moments5050w";run;

proc univariate data=restored5050 vardef=df noprint;
output out=moments5050r mean=mean var=var skewness=skew;
var x;
run;
data moments5050r; set moments5050r; dataset="moments5050r";run;
```

```
Proc append base=popmoments data=popmomentsw;  run;
Proc append base=popmoments data=moments9010w;  run;
Proc append base=popmoments data=moments9010r;  run;
Proc append base=popmoments data=moments8020w;  run;
Proc append base=popmoments data=moments8020r;  run;
Proc append base=popmoments data=moments7030w;  run;
Proc append base=popmoments data=moments7030r;  run;
Proc append base=popmoments data=moments6040w;  run;
Proc append base=popmoments data=moments6040r;  run;
Proc append base=popmoments data=moments5050w;  run;
Proc append base=popmoments data=moments5050r;  run;

data RQ4;
set popmoments;
run;
proc print data=RQ4 noobs;
var dataset mean var skew;
run;
```