A Simulation Study to Compare Weighting Methods for Nonresponses in the National Survey of Recent College Graduates

Amang Sukasih[1], Donsig Jang[1], Sonya Vartivarian[1],
Stephen Cohen[2], Fan Zhang[2]
[1]Mathematica Policy Research. Inc., 600 Maryland Ave., SW, Suite 550,
Washington, D.C. 20024
[2]National Science Foundation, Division of Science Resources Statistics,
4201 Wilson Blvd., Suite 965, Arlington, VA 22230

Abstract

Common methods to adjust sampling weights to account for survey nonresponse are the weighting cell technique, response propensity modeling, or a combination of both. Each raises several issues; for examples, which covariates to use to construct the weighting cells or to model response propensities, whether weights are used in modeling, and whether to weight the adjustment factor. To address these issues, we used a simulation based on a data from the National Survey of Recent College Graduates (NSRCG) maintained by the National Science Foundation to evaluate these weighting methods. In the end, we expect that the weighting adjustment will have successfully accounted for possible nonresponse bias.

Key Words: survey, nonresponse, weight adjustment, weighting cell, propensity model.

1. Introduction

1.1. Background

This paper focuses on methods to compensate for unit nonresponse through weighting adjustment in a survey data. A common method to compensate for unit nonresponse is to adjust the sampling weights. There are several methods including the weighting cell technique, response propensity modeling, or a combination of both. When performing weighting adjustments one may face with several issues as follows:

(a) Choosing the method: weighting cell, or propensity score modeling

(b) If weighting cell method is used, how to construct the cells: based on covariates, or based on propensity scores

(c) If response propensity modeling is used, how to estimate the model: design based (weighted) fitting, or random sample (unweighted) fitting

(d) Options in calculating adjustment factor: weighted response rate, unweighted response rate, or individual propensity score.

Weighting cell technique classifies respondents and nonrespondents into adjustment cells based on auxiliary information available for both groups so that samples within cells are homogeneous in their response propensities. Then, within each cell, respondents are weighted by the inverse of the response rate in the cell (for example, Lessler and Kalsbeek 1992). Response propensity modeling technique regress a binary response indicator to the survey on some predictors/covariates observed for both respondents and nonrespondents (Little 1992). Then, the predicted response propensities can be used further in weighting adjustment process as described in the next paragraph.

There are, in general, two different ways to form adjustment cells: one is based on a cross-tabulation of covariates (Oh and Scheuren 1983, Little 1986), and another is based on deciles of predicted propensity scores from a response propensity model (Eltinge and Yansaneh 1997). As an alternative to weighting cell adjustment methods, one may consider using the inverse of the estimated response probability (propensity score) for each individual respondent as the weighting adjustment factor (Czajka 1992).

When adjustment is done through weighting cell, the inverse of response rate can be used as the adjustment factor within each cell. This adjustment factor is then multiplied to the respondent’s sampling weight to produce the adjusted weight that accounts for the nonrespondents. The response rate can be calculated as a weighted ratio of the respondent counts to the sample counts, or as an unweighted ratio of the respondent counts to the sample counts (Little and Vartivarian 2003).

In this paper, we looked into these issues, and through a simulation study we investigated whether they provide better nonresponse adjustments with regard to correcting nonresponse bias and efficiency of the estimates for unit nonresponse in the National Survey of Recent College Graduates (NSRCG), a sample survey with a two-stage cluster sample design.

1.2. The National Survey of Recent College Graduates

The NSRCG is part of the Scientists and Engineers Statistical Data System (SESTAT) maintained by the National Science Foundation (NSF). SESTAT collects information about employment, educational, and demographic characteristics of scientists and engineers in the United States through three national surveys of this population: the National Survey of College Graduates (NSCG), the National Survey of Recent College Graduates (NSRCG), and the Survey of Doctorate Recipients (SDR).

The NSRCG covers a population of individuals who recently obtained bachelor's or master's degrees in a science, engineering or health field (SEH). The NSRCG provides information to the educational planners within the federal government and in academia, as well as the employers in all sectors (education, industry, and government) to understand and predict trends in employment opportunities and salaries in SEH fields for recent graduates, as this group of individuals has recently made the transition from school to the workplace or attending graduate school.

Different than the NSCG and SDR that use a one-stage stratified sample design and that both can be viewed as longitudinal surveys (sampled units are followed in subsequent survey rounds with a supplemental sample of new graduate cohorts), the NSRCG uses a two-stage sample design with a sample from schools at the first stage and a sample of graduates from selected schools at the second stage. At the first stage, lists of graduates who earned bachelor’s or master’s degrees during certain academic years are collected from sampled institutions. These lists are then used for construction of the sampling frame from which a sample of graduates is selected at the second stage.

In this simulation study we used a survey data from the 2006 NSRCG, which sampled 300 schools at the first stage and then sampled 27,000 graduates at the second stage. The list of graduates requested from sampled schools covered three academic years: 2002–2003 (AY03), 2003–2004 (AY04), and 2004–2005 (AY05).

Stratification of graduates implemented in the second-stage sampling was based on the following, yielding 756 sampling strata: three cohorts by degree year, two degree types (bachelor's and master's), 21 major fields of study, three race/ethnicity groups (non-Hispanic white; non-Hispanic Asian including Pacific Islanders and unknown races; and minority, including Hispanic, black, and American Indian), and two gender groups. For more details on the sampling design, see Jang et al. (2006).

2. Weighting Methods

In looking for a more systematic weighting adjustment procedure for the NSRCG, we proposed adjustment stages and weighting methods as follows. Sampled units can be classified by their disposition statuses: locating (located vs. not located), survey eligibility (eligibility known vs. eligibility unknown), and response (responded vs. not responded). Since the reason for nonresponse for each of these three dichotomous statuses is not the same and cases in each status can be characterized differently, this research recommended three stages of weighting adjustments to account for these three types of unit nonresponse separately:[1]

1. Adjustment for unlocated sample persons.
2. Adjustment for sample persons located but with unknown eligibility status.
3. Adjustment for sample persons located and eligible but did not complete the survey.

These adjustments are carried out sequentially (Iannacchione 2003). In these sequential adjustments, at any particular step, when the weighted procedure is implemented, weights from the previous step are used.

The following are the five weighting methods compared in the simulation:

- **Method 1 (SM1)**. Weighting cell is based on cross-tabulation of significant main effects with an unweighted ratio-adjustment factor
- **Method 2 (SM2).** Weighting cell is based on deciles of predicted propensity values from an unweighted model (not accounting for sample design) with an unweighted ratio-adjustment factor
- **Method 3 (SM3).** Weighting cell is based on deciles of predicted propensity values from an unweighted model with a weighted ratio-adjustment factor
- **Method 4 (SM4).** The inverse of individual predicted propensity values from an unweighted model is the adjustment factor

[1] This research focuses only on weighting adjustment for unit nonresponse and excludes any other post-nonresponse adjustments, such as post-stratification, trimming, or raking.

- **Method 5 (SM5).** The inverse of individual predicted propensity values from a weighted (design-based) model is the adjustment factor.

In both weighting cell and model-based response propensity methods, five sampling variables plus a variable that indicates whether or not the sampled student is a non-U.S. resident alien are used as the candidates for weighting cell construction and model building. These variables are: graduate cohort, degree level, degree field, race/ethnicity, gender, and residency status.

Not all variables may be included in each weighting stage; a modeling procedure is used to select variables used for weighting. In method SM1, first, a main-effect model with these six predictors is estimated. Then, significant main effects are identified and only these significant effects are used to construct cross-tabulation that creates weighting cells. It turned out that all six variables were significant. For small cells (cells with sample size less than 20) cell collapsing is carried out before nonresponse adjustment takes place. Cell collapsing process was designed to be as systematic as possible, where choosing the variables to collapse relied on the order of significance level of covariates, and was automated to avoid subjective judgment.

For weighting methods based on response propensity model (SM2-SM5), the 2006 NSRCG samples provide information on response propensities and survey outcomes in the recent college graduate population. The propensity score is estimated through a logistic regression model, where the logit function of response/nonresponse indicator variable is regressed with a set of covariates observed for both respondents and nonrespondents (the six main effects mentioned above and their interaction terms). First, CHAID (Chi-square Automatic Interaction Detection) using AnswerTree® software is used to select potential two- and three-way interactions (Magidson 1993). Then, an initial model that includes all main effects and CHAID identified interaction terms is fitted using a stepwise variable selection. The resultant model from this step is used as the final model in the simulation methods SM2-SM5.

Once this final model is determined, response propensity scores are estimated using either unweighted fitting which is a regular random sample modeling method (SM2, SM3, SM4) or weighted fitting (SM5) which is a design based modeling method for clustered data. These estimated response propensities are then used directly as the adjustment factors in methods SM4 and SM5, or used to form 10 adjustment cells in methods SM2 and SM3 (Eltinge and Yansaneh 1997). To form 10 adjustment cells based on estimated response propensities, these estimated response propensities are sorted in increasing order. Then, nine cut-off points based on the 10th,20th,30th,…,90th percentiles are estimated and used as the boundaries for the weighting cells.

When weighting cell method is used (SM1-SM3), the nonresponse adjustment factor within each cell can be calculated as the inverse of unweighted response rates in methods SM1 and SM2; or as the inverse of weighted response rates in method SM3.

3. Simulation Setting

We used the 2006 NSRCG sample in simulating data with nonresponses. However, since we do not have survey outcomes for the nonrespondents, we use respondents-only data and treat this as if it were a full sample. We will call these data sets the "full-sample data." When weighted by the 2006 NSRCG final analysis weights, this set of respondents represents the population of graduates for the NSRCG. These data serve as the benchmark when evaluating simulated data. Given these data, we generated replicates and simulated unit missingness within each replicate.

3.1. Response Rate

We simulated the overall graduate response rate, broken down into three components corresponding to the three stages of nonresponse adjustment. In the 2006 NSRCG, the unweighted rate of each component is as follows:

- Location rate = 76.3%
- Known-eligibility rate among located = 90.1%
- Completion rate among eligible = 99.2%

However, since the last rate is large and close to 100%, we did not include it in our simulation. The combined rate for the first two rates is 76.3% × 90.1% = 68.7%. In our simulation we simulated three different (unweighted) graduate response rates as follows: 60%, 68.7%, and 80%. However, the same conclusions hold, and we therefore only focus on the response rate of 68.7% in this paper.

3.2. Nonresponse Mechanisms

The three different nonresponse mechanisms are considered in this simulation:

Missing Completely at Random (MCAR). Since MCAR does not depend on covariates (“coin toss”), the response probability/propensity is constant for everyone. Suppose R_i denotes indicator of response/nonresponse for graduate i; that is, $R_i = 1$ if responding; $R_i = 0$ otherwise. Then, R_i (for each weight adjustment step) will be generated as follow: $R_i \sim \text{Bernoulli}(P)$ for graduate i, where the value of P is given as follows (note that these response rates are similar to those existing in the original NSRCG data):

- Non-located adjustment: $P = 0.763$
- Unknown eligibility adjustment: $P = 0.90$

Missing at Random (MAR). In MAR, the missingness depends on observed values of covariates. That is, the probability to respond will differ from case to case depending on the values of their covariates. Three options were used as the response propensities:

(1) **MAR1**. Unweighted response rate in weighting cells constructed based on cross-classification of significant variables,

(2) **MAR2**. Unweighted response rate in 10 weighting cells constructed based on the estimated propensity scores,

(3) **MAR3**. Individual estimated propensity score calculated through a design-based logistic regression.

This response rate/probability under the three MAR schemes above is attached to each case and this value will be used as the probability parameter P_i to generate response indicator in a Bernoulli random number generator.

Not Missing at Random (NMAR). In the NMAR, the missingness depends on covariates as well as (unobserved) values of survey outcomes. Suppose we assume that nonrespondents in the NSRCG corresponded to the following groups (based on survey outcomes WRKG and/or SALARY):

– Graduates who did not have a job (WRKG = "No"),

– Graduates who had a job with high income (SALARY > 100,000).

First we assigned response probability from MAR3 to each individual, and then adjust this response probability based on the values of WRKG and SALARY. We would expect to observe a lower response probability for the cases in the above two groups and a higher probability for the rest. That is, for cases with WRKG = "No" or WRKG = "Yes" with SALARY > 100,000 we assign an average response probability of 0.45 in adjustment for location (we did not change the probability of known/unknown eligibility).

Thus, in summary this paper presents 5 data sets used in the simulation: MCAR data set, three MAR data sets corresponding to three different response propensity calculations, and NMAR data set.

3.3. Computer Programming and the Number of Replicates

When choosing the number of replicates, we considered not only the convergence of the statistics being evaluated, but also the length of time required to perform the whole process. We performed this simulation using *R*, a software for statistical computing and graphics (www.r-project.org). All calculation here, including survey estimation and design-based modeling, can be run under *R*. We ran the simulation 1,000 times. A run based on a larger number of replicates (2,000) on some of the data sets produced similar results. Therefore, we decided to use 1,000 replicates in all 5 data sets.

3.4. Evaluation

To measure bias correction through weighting, we compared the survey estimate calculated based on the full-sample survey data using the final survey weights (the "true value") to the estimate calculated based on the simulated "respondents-only" using the nonresponse-adjusted weights across 1,000 replicates. The survey estimates/statistics to be compared are:

- Total estimates: overall, by degree field and degree level.
- Median salary: by degree field and degree level.
- Mean salary: by degree field and degree level.
- Proportion of employed: by degree field and degree level.
- Proportion of unemployed looked for work: by degree field and degree level.

Suppose $\hat{\theta}_0$ denotes the estimate calculated based on the full sample (no unit missing), and $\hat{\theta}_{ir}$ denotes the estimate calculated based on each simulated data under method *i* (*i*=1,2,3,4,5) from replicate *r* (*r* = 1, …, 1000). We calculated the percentage of relative difference, defined as

$$RELDIFF_{ir} = \frac{\hat{\theta}_{ir} - \hat{\theta}_0}{\hat{\theta}_0} \times 100\%, \qquad (1)$$

so that the magnitude of difference from the "true value" (the bias) can be measured as a percentage. We investigated the plots of $\hat{\theta}_{ir}$ where the horizontal axis represents index of individual replicate and the vertical axis represents the relative differences of statistic being compared (for example, see Figure 1).

Also, the following mean of differences (*BIAS*) and square root mean square error (*RMSE*) can be used to measure the magnitude of bias and variability of the estimate from weighting adjustment for nonresponse:

$$BIAS_i = \frac{1}{1000}\sum_{r=1}^{1000}\left(\hat{\theta}_{ir} - \hat{\theta}_0\right), \tag{2}$$

$$RMSE_i = \sqrt{\frac{1}{1000}\sum_{r=1}^{1000}\left(\hat{\theta}_{ir} - \hat{\theta}_0\right)^2}. \tag{3}$$

In addition, we evaluated the effect of weighting adjustment to the variance of total estimate by comparing the design effects (*DEFF*) due to weight variation (Kish, 1992) across 1,000 replicates.

4. Simulation Results

Discussion on the simulation findings will focus on the following:

- Response propensities
- Bias and efficiency of the Total, Median, Mean and Proportion
- Design effects due to the weights

4.1. Response Propensities

The response propensity is calculated as the response rate within each weighting cell in the weighting cell adjustment method, or the estimate of propensity score given the covariates for each case in the model-based adjustment method. Once calculated, this response propensity is then used as the weighting adjustment factor, which is calculated as the reciprocal of response propensity. Here, we assessed the variability of response propensities produced under the five simulation methods by using a 95 percent confidence interval of response propensities $\overline{p} \pm 1.96\, s_p$, where $\overline{p}$ is the mean response propensities and s_p is the standard deviation of response propensities for each replicate.

In this application, when missing data is MCAR, there is little variability in the response propensities, as expected.[2] Weighting adjustment may not be a practical concern in

[2] Method SM1 produces response propensities with less variability than methods SM2, SM3, SM4, and SM5. This is to be expected since with the covariate adjustment cells in SM1 we allowed for collapsing and often ended up with few cells under MCAR; however, in the other methods we "forced" these weighting methods to have 10 cells or as many as individual covariate patterns which lead to more variability. Though, in the data with MCAR such differences across methods are small.

MCAR. When missing data is MAR (MAR1, MAR2, and MAR3), in any adjustment method the variability of response propensities becomes large. Weighting method SM1 produces response propensities with more variability than other methods do, although these differences are not too striking. Under the NMAR data, in any weighting methods the response propensities are moderate—larger than those under MCAR data but smaller than those under MAR data. Method SM1 produces response propensities with more variability than other methods do.

When comparing between method SM2 and method SM3 (weighted and unweighted response rates within the same adjustment cells), in any missing data mechanisms, in general the unweighted response rates (method SM2) have slightly smaller variability than do the weighted response rates (method SM3), though such differences are almost negligible. Also, when comparing between the unweighted (random sample) model and the design-based model (method SM4 vs. method SM5), there exist small differences between the two, especially on the right tail of propensity distribution where the design-based propensity scores tend to be larger than those based on the unweighted model.

4.2. Nonresponse Bias Correction

a. Total

The sum of weights across all samples represents an estimate of total population overall (frame total). When there are nonresponses and the sampling weights for respondents are adjusted to account for these nonrespondents, the sum of the adjusted weights is expected to be equal to the frame total. We first discuss total estimates overall, and then we discuss total estimates within certain domains.

Totals, Overall. Under MCAR, method SM3 produced adjusted weights that sum exactly to the frame total for each simulation replication; while in method SM5 the sum of adjusted weights actually varies a little to the frame total across replicates. The estimate of grand total based on these weights is unbiased[3] and very efficient (the estimate has small variability across replication). On the other hand, methods SM1, SM2, and SM4 produced adjusted weights that still result in an unbiased estimate of grand total but are inefficient, as the variability of the estimates across replicates is large.

Under all MAR data (MAR1, MAR2, MAR3), methods SM3 and SM5 also produce adjusted weights that sum to the frame total. Under data with MAR1, there is a tendency for SM2 to produce weights that underestimate the frame total, though this underestimation is minor (see Figure 1). Recall that in the simulation MAR1 response mechanism is generated with response probability calculated as the unweighted response rate within the weighting cells based on the covariates. However, SM2 adjusts the weights using unweighted response rate calculated within the 10 cells (based on response propensity deciles). Thus, it is important to know whether the (unweighted) response rate under SM2 is an unbiased estimate of MAR1 response probabilities within each of 10 cells. If that is the case, then the adjustment under SM2 may produce unbiased estimate of totals. However, when that is not the case, then the adjustment under SM2 may produce biased estimate of totals.

[3] Throughout the discussion of simulation result, the term “unbiased” is used to indicate the expected value or the average across sample replications (or repeated samples).

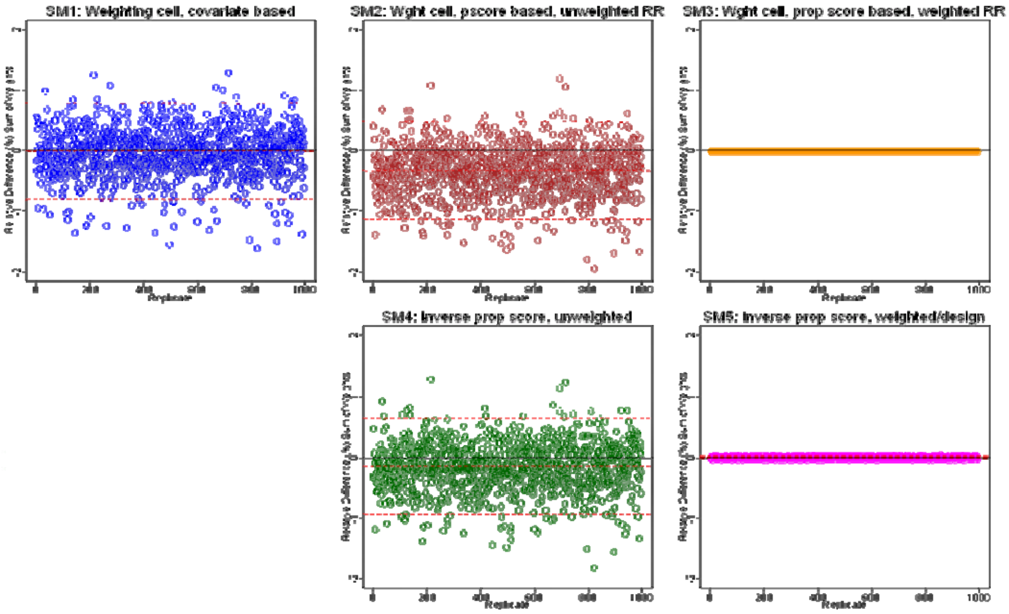

Figure 1: Estimate of grand total by weighting methods SM1-SM5, for MAR1 data.

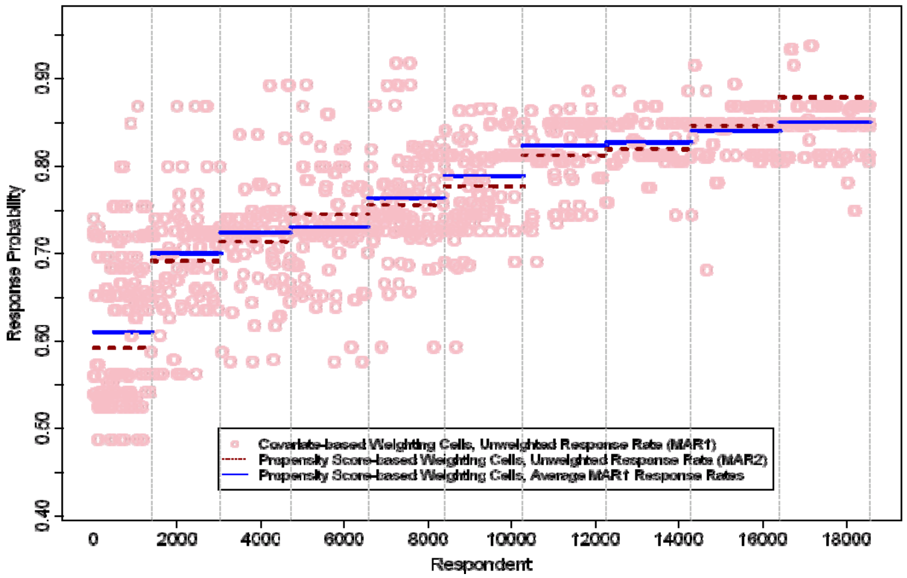

Figure 2: Response probabilities for MAR1, MAR2, and average MAR1 response probabilities within 10 cells, 2006 NSRCG Data

To investigate this, in Figure 2 we plotted the unweighted response rates based on the covariate based weighting cells used to generate MAR1 presented by the circles, and the unweighted response rates based on the propensity score based weighting cells (method SM2) within 10 cells presented by the dotted lines. To check whether unweighted response rate in SM2 is unbiased estimate of MAR1 response probabilities, we calculated the mean of MAR1 response probabilities within each of 10 cells, and plotted them as the solid lines. One can see that in the first and the last cells, the unweighted response rate (dotted line) is biased for the solid line, which means the adjustment factor under method SM2 is biased. As a consequence, when the response propensity correlates with sampling weight, unweighted response rate used as adjustment factor may produce biased estimate. As an example in the NSRCG, Asian and Minority are oversampled thus have higher sampling weights than Whites; however, the response rate in Asian and Minority is lower than that in White. Thus, when nonrespondents are cases with larger sampling weights; the estimate is underestimated.

Under data with MAR2 and MAR3, methods SM1, SM2, and SM4 produced weights that sum to the frame total (unbiased) but less efficient variance, where the variability of this statistic across 1,000 replicates is less than that under the data with MCAR.

Under the NMAR data, simulation result shows a different pattern. Yet, methods SM3 and SM5 produced weights that still sum to frame total. However, methods SM1, SM2, and SM4 resulted in a sum of weights that clearly underestimates the population total. However, weighting adjustments are not meant to address NMAR data, and so we do not address these results further.

Note that in methods SM1, SM2, and SM4, where the estimate of grand total is either downward biased or unbiased with large variation, even though the absolute value of deviation of the sum of weights from the frame total is not trivial (ranging from 23,000 to 36,000 graduates), the percentage of relative differences is small, less than ± 2 percent.

For the grand total estimation, methods SM3 and SM5 are superior than methods SM1, SM2, and SM4 with regards to the bias and variance of the estimate. We note that with method SM3, using the weighted response rates within cells forces the sum of weights to equal the frame total, where the total is calculated as the sum of final weights across all respondents in the original data. In this case it can be shown analytically that the weighting adjustment method SM3 always produces adjusted weights that sum back to the true grand total in the following paragraphs. However, this may not apply to subgroup analyses.

Let Y_i and w_i, respectively, denote a surveyed variable and sampling weight for individual sampled person i. For a grand total estimation, $Y_i = 1$ for all i. In addition, let n_c denotes the sample size within weighting cell $c\ (c = 1, \cdots, C)$, and w_{cs} denotes the sampling weight for individual s within cell c. The total estimate $\hat{T}$ defined as

$$\hat{T} = \sum_{i=1}^{n} w_i Y_i = \sum_{i=1}^{n} w_i = \sum_{c=1}^{C} \sum_{s=1}^{n_c} w_{cs} \tag{4}$$

is an unbiased estimate of the corresponding overall total population T. Let n_c^* denotes the number of respondents in weighting cell c. When the adjustment factor A_c is calculated as the inverse of weighted response rate (method SM3), then it can be shown that the estimate calculated based on respondents only $\hat{T}^*$ (with the nonresponse adjusted weights) will produce the same total estimate $\hat{T}$ as follows:

$$\hat{T}^* = \sum_{c=1}^{C}\sum_{s=1}^{n_c^*} A_{cs} w_{cs} = \sum_{c=1}^{C} A_{cs} \sum_{s=1}^{n_c^*} w_{cs} = \sum_{c=1}^{C} \frac{\sum_{s=1}^{n_c} w_{cs}}{\sum_{s=1}^{n_c^*} w_{cs}} \sum_{s=1}^{n_c^*} w_{cs} = \sum_{c=1}^{C}\sum_{s=1}^{n_c} w_{cs} = \hat{T}\,. \tag{5}$$

When an unweighted response rate is used (method SM2), the same result is not guarantee but it can be easily obtained by post-stratification to the desired frame total. However, in this study, we have not considered the final stages of weighting, which typically includes such post-stratification.

To analytically prove that weighting adjustment method SM5 also produces unbiased with very small variance grand total estimate is not trivial. This proof will involve evaluating mathematical expectation of the inverse response propensity. Let R_i denote a binary response indicator, X_i is covariates, and ϕ_i denotes the true response propensity for individual i. Under a response propensity model and a missing at random mechanism we assume that the expected value of R_i given the covariates X_i and the sample design is equal to the true response propensity, or

$$E_m\left(R_i \mid \mathrm{X}_i, w_i\right) = \phi_i \tag{6}$$

where E_m denotes the expected value with respect to the model. The true response propensity ϕ_i is estimated by $\hat{\phi}_i$, and in the weighting method SM5 the individual inverse of $\hat{\phi}_i$ is used as the weighting adjustment factors. The estimate of total using the weights adjusted by inverse of individual response propensity is calculated as

$$\hat{T}^* = \sum_{i=1}^{n^*} \frac{1}{\hat{\phi}_i} w_i = \sum_{i=1}^{n} R_i \frac{1}{\hat{\phi}_i} w_i\,. \tag{7}$$

It can be shown that this total estimate is an unbiased estimate for population total. First, by implementing Taylor Series approach we can show that

$$E_m\left(\frac{R_i}{\hat{\phi}_i}\middle| \mathrm{X}_i, w_i\right) \approx 1\,. \tag{8}$$

Then we can implement a double expectation (E_d and E_m, respectively, denote the expected value with respect to the model and the design) as follows:

$$E_d\left[E_m\left(\hat{T}^* \mid \mathrm{X}_i, w_i\right)\right] = E_d\left[\sum_{i=1}^{n} w_i E_m\left(\frac{R_i}{\hat{\phi}_i}\middle| \mathrm{X}_i, w_i\right)\right] \approx E_d\left[\sum_{i=1}^{n} w_i\right] \approx T\,. \tag{9}$$

NSRCG Totals, Within Domains. We calculated survey estimates (total graduates, median and mean salary, proportion of employed, and proportion of unemployed looked for work) for each domain defined as a cross-classification between two levels of degree (bachelor’s and master’s) and eight groups of degree fields resulting in 16 domains of analyses. For the MCAR data, when comparing across five methods in any domain of analysis, any methods of the five weighting methods produced unbiased total estimate with a large variance across replicates. The variability is about the same across five weighting methods. Therefore, there is no difference in the estimates across five weighting methods under data with MCAR as the missing mechanism.

For the MAR and NMAR data, given a particular missing data, in general the five weighting methods seem to produce the same pattern. There is no specific pattern of bias/unbiasedness that can be attributed to the specific weighting method. The bias/unbiasedness depends on the domain of analysis and specific missing data mechanism we are looking at. That is, whenever there is a bias in estimation for a specific domain of analysis and missing data mechanism, all five methods tend to produce a bias estimate with the same direction, either underestimation or overestimation.

b. Median Salary

In general, for a given domain of analysis and missing data mechanism, the estimate of median salary and the variability across 1,000 replicates are about the same across the five weighting methods. The estimates are unbiased under the MCAR and MAR data. Under the NMAR data, however, the estimate of median salary is underestimated for all domains. (Recall that our simulation is set up to randomly exclude large portion of cases in the high salary group.) This explains that NMAR data cannot be taken care of through any of our weighting methods SM1-SM5.

c. Mean Salary

The conclusion for estimate of mean salary is the same as for median salary. In general, for a given domain of analysis and missing data mechanism, the estimate of mean salary is the same across the five weighting methods, with the same variability as well. In addition, for all domains of analysis the estimate of mean salary is unbiased under the MCAR and MAR missing data mechanisms. Under the NMAR data, the estimate of mean salary is underestimated for all but for a few domains, where the estimate of mean salary is either unbiased or only slightly underestimated.

d. Proportion Employed

Across domains, the proportion of employed in the SESTAT population is from moderate to large. For example, in the 2006 NSRCG, these numbers range from 67 percent to 93 percent across domains defined by degree level and degree field. We compared the estimate resulting from the simulation to the number based on the full sample. Note that in the NMAR data, the unemployed graduates were randomly excluded in each simulated data.

Given a specific domain of analysis and under a specific missing data mechanism, the five weighting methods produced an estimate that is about the same. Under MCAR and MAR data the estimates based on all weighting methods are unbiased. In some domains these estimates of proportion have large variance, but in other domains the variability is small. Since the variance of proportion is a function of the proportion itself and sample size, the variability that we saw in these plots could possibly due to either the magnitude of proportion or the sample size, or both. Under the NMAR data, the proportion of employed graduates is overestimated; which means that the weighting adjustment was not able to correct nonresponse bias under the NMAR data.

e. Proportion Unemployed Looking for Work

The denominator for this proportion is unemployed graduates, which is only 2,278 cases in the 2006 NSRCG data. The estimates of proportion range from 12 to 48 percent across our 16 domains of analyses. These estimates are based on as few as 31 cases in the smallest domain (master's graduates in the Health Sciences), and as many as 457 cases in the largest domain (bachelor's graduates in the Social and Related Sciences). Note that in our NMAR data simulation we randomly excluded a large portion of unemployed graduates and then further broke down this sample by our 16 domains of analyses. Thus, estimation of the "proportion of unemployed looked for work" under NMAR data sets can be considered as small domain estimation.

Given a specific domain of analysis and a specific missing data mechanism, the five weighting methods produced estimates that are about the same. The plots show that under all missing data mechanisms and for all domains of analyses the estimates based on all weighting methods are unbiased, but with large variance. Under the NMAR data, the variance is quite large. We conjecture that this is because of the small domain sample size rather than the weighting adjustment method.

4.3. Design Effect Due to Weight Variation

When there is nonresponse in the sample, weighting adjustment may add more variability within the respondent weights. We assessed possible variance inflation due to the variability added to the adjusted weights that results from using a particular weighting method. Variance inflation due to weight variation can be measured through the design effect (DEFF) due to weight variation. The design effect computed here is only for the grand total estimate, which is calculated based on all respondents. We compared the design effect calculated from the survey final weights in the full-sample data to the nonresponse-adjusted weights in the simulated data.

Under the MCAR data, there is no increase in the design effect for any weighting methods used, though variability across replications is not trivial. Under the MAR data, in general the design effect was increased, except for methods SM4 and SM5 when the data is MAR1. Under the NMAR data, it is clear that all weighting methods inflate the grand total variance as the design effect increased.

5. Conclusion and Discussion

When the missing data mechanism is MCAR, any method of weighting adjustment should work. The point estimates and their variances show similar results across weighting methods for all domain analyses and types of statistics. This missing data mechanism is not a practical concern, as a simple ratio weighting adjustment technique may provide satisfactory compensation for unit nonresponse. On the other hand, when the missing data mechanism is NMAR, the method of weighting adjustment for nonresponse may not be successful in correcting nonresponse bias, as expected. Therefore, we present our conclusion for missing data under MAR assumption as follows.

Based on our simulation investigation using the 2006 NSRCG, our main conclusions under MAR are as follows:

- Overall, all five methods considered for simulation are comparable.

- Weighting cells based on covariates can lead to issues with respect to small cell sizes and requires collapsing. Though in practice such collapsing strategy can be ad hoc, in our study simulation small cells were effectively handled with a systematic, automated method of collapsing. As the number of covariates including paradata available for use in the covariate cell creation increases, the covariate cell adjustment becomes a less desirable method because of sparseness and the increased need to collapse cells, ultimately limiting the ability to incorporate additional covariates.

- Alternatively, inverse propensity estimate adjustments maximize the utilization of auxiliary information and nonresponse bias reduction. However, there is often the concern that this may cause the most variable weights, and thus in turn, larger variances of estimates. In this study, more variable weights were observed, but the impacts on survey estimates were minimal in our NSRCG application due to a few cases having the largest weights. In practice, some of this weight variation may be dealt with via weight trimming.

- As an alternative to the covariate-based weighting class adjustments and the individual inverse propensity estimate adjustments, the hybrid technique—propensity cell adjustments are attractive in a sense that this method makes the response propensity distribution smooth (and thus making the weight variation less) while utilizing all auxiliary information. The propensity cell adjustments are able to avoid the sparseness covariate cell adjustments face and the variability that inverse propensity weights face.

- Note, however, with disproportionate sampling rates, the propensity cell method with unweighted adjustment factor, for estimates of the grand total, might over/underestimate because each of the weighting cells based on propensity values might cut across many sampling cells with different sampling rates, and these rates may be related to nonresponse. In this case, inverse of weighted response rate as adjustment factor is recommended. This simulation did not consider post-stratification, which would also accomplish unbiased estimates of domain totals.

- Our simulation concluded that when the weighting cells are constructed based on deciles of propensity scores, the weighted response rate results in unbiased survey estimate with minimum variance for our main estimate of interest, the grand total, though there was no clear pattern for domain totals, means, medians and proportions.

Therefore, based on this study and when considering the overall total estimate, under the MAR missingness—which is a common assumption in practice—the weighting cells method based on grouping the propensity scores with the adjustment factor calculated as weighted response rate (method SM3), or the weighting method that uses the adjustment factor calculated as the inverse of individual estimated propensity score with the design-based fitting technique used to estimate model parameters (method SM5) provide a reasonable method for the NSRCG nonresponse adjustment. Because method SM3 uses weighted ratios for the adjustment factors, it produces adjusted weights that sum back to the frame total. We also note that for the estimation of grand total, method SM5 provides total estimates close to the frame total. This is important to SESTAT since grand totals are the main estimate of interest. Further, methods that involve modeling can better handle future paradata, and this is a strong reason to consider using propensity methods in SESTAT that avoid the issues of cell collapsement. Finally, knowing that all weighting methods we examined in the simulation study perform similarly for other estimates (domain totals, means, medians and proportions) is important as every weighting method we assess should adjust for nonresponse bias under missing at random (MAR).

The method recommended here may not provide similar results under different data sets. Therefore, we suggest that the judgment to choose a particular weighting-adjustment method should be based on the specific survey design used, as well as a thorough empirical investigation of the missing data. Given the data set, we also recommend that the statistician who constructs the weights should investigate whether a slight modification of the procedures/methods (for example, weighted vs. unweighted) produces significantly different results, and whether such differences (if pronounced) result in different survey estimates. Readers can refer to Korn and Graubard (1999) for diagnostic techniques when using survey weights.

Acknowledgements and Disclaimer

Work on this article was supported and funded by the National Science Foundation, contract SRS-04-20325. The views expressed here are those of the authors and not necessarily those of the National Science Foundation or Mathematica Policy Research.

References

Czajka, J. L., S.M. Hirabayashi, R.J.A. Little, and D.B. Rubin. “Projecting from Advance Data Using Propensity Modeling: An Application to Income and Tax Statistics.” Journal of Business & Economic Statistics, American Statistical Association, vol. 10, no. 2, April 1992, pp. 117-31.

Eltinge, J. L., and I.S. Yansaneh. “Diagnostics for Formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey." Survey Methodology, vol. 23, 1997, pp. 33-40.

Iannacchione, V.G. “Sequential Weight Adjustments for Location and Cooperation Propensity for the 1995 National Survey of Family Growth.” Journal of Official Statistics, vol. 19, no. 1, 2003, pp. 31-43.

Jang D., M. Satake, M.E. Bozylinsky, H. Xu , and X. Lin. “Sample Design for the 2006 National Survey of Recent College Graduates.” Report submitted to National Science Foundation. Princeton, NJ: Mathematica Policy Research, Inc., March 2006.

Kish L. *Weighting for unequal Pi*. Journal of Official Statistics. Vol 8, 1992, pp. 183 – 200.

Korn, E. L., and B. I. Graubard. *Analysis of Health Surveys*. New York: Wiley, 1999.

Lessler, J.T., and W.D. Kalsbeek. *Nonsampling Error in Surveys*. New York: Wiley, 1992.

Little, R.J.A.”Survey Nonresponse Adjustment for Estimates of Means.” International Statistical Review, vol. 54, 1986, pp. 138-157.

Little, R.J.A. “Models for Nonresponse in Sample Surveys.” Journal of the American Statistical Association, vol. 77, 1992, pp. 237-250.

Little, R.J., and S. Vartivarian. “On weighting the rates in nonresponse weights.” Statistics in Medicine, vol. 22, 2003, pp. 1589-1599.

Oh, H.L., and F.J. Scheuren. “Weighting Adjustment for Unit Nonresponse.” In Incomplete Data in Sample Surveys, vol. 2, Theory and Bibliographies, edited by W.G. Madow, I. Olkin and D.B. Rubin. New York: Academic Press, 1983.