Longitudinal Survey Weight Calibration Applied to the NSF Survey of Doctorate Recipients

Michael D. Larsen, Department of Statistics & Biostatistics Center, GWU
Siyu Qing, Department of Statistics, GWU
Beilei Zhou, Biostatistics Center, GWU
Mary A. Foulkes, Departments of Epidemiology/Biostatistics
and Health Policy & Biostatistics Center, GWU
The George Washington University

Michael D. Larsen, The George Washington University, Biostatistics Center
6110 Executive Blvd., Suite 750, Rockville, MD 20852
email: mlarsen@bsc.gwu.edu
Siyu Qing, The George Washington University, Department of Statistics
Rome Hall Room 553, 801 22nd St. NW, Washington, D.C. 20052
email: qingsy04@gwmail.gwu.edu
Beilei Zhou, The George Washington University, Biostatistics Center
6110 Executive Blvd., Suite 750, Rockville, MD 20852
email: bzhoubsc.gwu.edu
Mary A. Foulkes, The George Washington University, Biostatistics Center
6110 Executive Blvd., Suite 750, Rockville, MD 20852
email: mfoulkes@bsc.gwu.edu.

Abstract
The National Science Foundation's Survey of Doctorate Recipients is conducted every two or three years and collects detailed information on individuals receiving PhDs in science and engineering in the U.S. and some others with PhDs from abroad in these areas. Survey weights adjust for oversampling and nonresponse on a cross sectional basis. A significant portion of the sample (e.g., 60% on 3 or more surveys from 1993-2006) appears in multiple survey years and can be linked across time. No longitudinal weight exists that would enable estimation of statistical models or comparison of finite population characteristics using data from multiple survey waves together. This paper applies calibration estimation for construction of such a longitudinal weight for this survey. Previous results studied the process of weight construction through simulation. Here we report on applications to NSF survey data. Choices of multivariate calibration targets are compared in a series of analyses.

Keywords
Calibration weighting; Longitudinal study; Panel study; Raking; SESTAT; Survey sampling.

Acknowledgments
Funding has been provided by NIH National Institute of General Medical Sciences (NIGMS) cooperative agreement (1 U01 GM094142-01). Disclaimer: The work and opinions expressed here are the responsibility of the authors and neither the National Institutes of Health nor the National Science Foundation.

1 Introduction

The National Science Foundation's Survey of Doctorate Recipients (NSF SDR) is gathers detailed information on people receiving PhDs in science and engineering in the United States and some others with PhDs from abroad in these areas. It is conducted every two or three years. Each survey year, survey weights adjust for oversampling and nonresponse. This is done on a cross-sectional basis. The survey has many uses, including providing estimates for use in reports such as those by the NSF (2008, 2011). Every survey year the target population changes, because people enter (e.g., new Ph.D. recipients in the U.S.) or leave (e.g., deaths) the population. Variables cover labor force status, academic rank and tenure, salary, field and institution of degree and employment, age, sex, race/ethnicity, marital status, spouse employment, whether children are at home and their ages, U.S. citizenship, work responsibilities, management position, professional memberships, reasons for taking a post doctoral position, and questions about a career path job.

Every survey year, survey weights adjust for oversampling and nonresponse. This means that an analysis using the survey data with the survey weights in a given year is representative of a corresponding population. A large portion of the sample (e.g., 60% on 3 or more surveys from 1993-2006) appears in multiple survey years and can be linked across time. Despite that fact the survey weights are not designed for longitudinal analysis of data sampled over time. Longitudinal analysis, of course, is still possible, but such an analysis would typically be anchored in a sample year. It does mean that there are no longitudinal survey weights that would enable estimation of statistical models or comparison of finite population characteristics.

1.1 Longitunidal Analysis and the SDR

As described in Larsen *et al.* (2011), the type of analysis of change over time that can be accomplished with the Survey of Doctorate Recipients is focused on cohorts defined by survey years. If one wants to estimate rates of progression or factors associated with advancement in employment within a field of study, then one can do so using a particular cohort or survey year. A consequence of conducting cross-sectional analyses is that sample sizes are more limited than they would be if longitudinal analysis was planned into the design. Another limitation occurs when estimating statistical models of change over time. Ideally one would use all respondents from all survey years. What should one do with the cross-sectional survey weights that each respondent has for each survey in which they participate? If there were one longitudinal survey weight for each unique respondent, then combining respondents from different survey years would be more readily doable.

1.2 Surveys Designed for Longitudinal Analysis

As described in Larsen *et al.* (2011), some surveys are designed with planned longitudinal, panel, or time series analyses in mind. These surveys include the American Community Survey (ACS; http://www.census.gov/acs/www/, U. S. Census Bureau 2009; chapter 4) and the Current Population Survey (CPS; http://www.bls.gov/cps/; http://www.census.gov/cps/, U. S. Census Bureau 2006),

There are many other surveys – longitudinal surveys and panel surveys – that are designed to measure change over time. Examples include the Survey of Income and Program Participation (SIPP), the National Longitudinal Surveys (http://www.bls.gov/nls/), the Panel Study of Income Dynamics (http://psidonline.isr.umich.edu/), the 2009 Panel Survey of Consumer Finances (http://www.federalreserve.gov/pubs/oss/oss2/scfindex.html), and the Medical Expenditure Panel Survey (http://www.meps.ahrq.gov/mepsweb/). An example in the area of environmental surveys is the National Resources Inventory (Breidt and Fuller 1999). See also Duncan and Kalton (1987), Fuller (1999), and McDonald (2003) and references therein.

1.3 Outline

This paper explores the construction of longitudinal weights for cross-sectional sample surveys using calibration estimation (Deville and Särndal 1992 and references given below). Section 2 discusses survey calibration weighting and estimation. Section 3 outlines a proposal for the formation of longitudinal survey weights from cross-sectional weights. Results of a simulation using this proposal were described in Larsen *et al* (2011). Section 4 describes application of methods to data from the NSF Survey of Doctorate recipients. Section 5 discusses findings, limitations, and future work.

2 Calibration Weighting

This section is repeated from Larsen *et al.* (2011). It provides necessary background for understanding calibration estimation and weighting.

Calibration estimation and calibration weighting methods were described by Deville and Särndal (1992). The connection to raking adjustment was demonstrated in Deville, Särndal, and Sautory (1993). Reviews of the literature and methods for calibration in sample surveys can be found in Kim and Park (2010) and Särndal (2007). Calibration methods in survey sampling allow one to adjust survey weights so that they are close to initial weights, such as the sampling design weights, but satisfy certain constraints. The closeness of the weights is described by a distance function. For example, if x_k is a value for a variable X on subject k in the sample and the total for variable X in the population is known to be t_x, then a constraint could be that the weighted total of the x-values in the sample equal t_x: $\sum_{k \in s} x_k w_k = t_x$.

Let $\{d_k\}$ be original survey (design) weights. Let $t_x = \sum_U x_k$ is a known total in the population with indices U; x_k can be a vector. The calibrated weights $\{w_k\}$ are "close" to $\{d_k\}$ but satisfy a set of calibration equations: $\sum_s w_k x_k = \sum_U x_k$. There are various ways to compute the weights, including in the R survey package (Lumley 2011). Calibration weighting can match (published) control totals and reduce mean squared error. A reduction in mean squared error might occur when the x variable is sufficiently correlated with an outcome y variable.

Calibration can be implemented in a way to control the minimum and maximum value of weights and to match one or more control totals. It is therefore a very flexible methodology. Indeed, Zhang (2000) describes how calibration can produce adjusted weights equivalent to those produced with post stratification.

In the context of nonresponse weighting, one can specify the desired post stratification adjustments in terms of control totals for calibration weighting. For example, the goal could be to have the sum of weights for respondents in a weighting class or post stratification cell match the sum of weights of sampled units in that cell. One might also want to place an upper bound on the largest weight in the cell. Then the survey calibration algorithm provides a procedure for adjusting the current weights. The Research Triangle Institute (RTI 2008) implements a general methodology that enables this form of calibration. Inherent in the use of calibration, cell-based adjustment, and raking is the need to select variables and subgroups to define the control targets. These methods will be more successful in removing non-response bias if cells and control variables are related to probabilities of non-response and to variables used for analyses. Mirel *et al.* (2010) used the RTI SUDAAN program to compare weighting class and more general calibration adjustments for weights in the NHANES (2003-2004).

In some survey settings, researchers have used calibration to adjust weights to match *estimated* control totals. Estimated control totals have their own degrees of uncertainty associated with them. Variance estimation with calibrated estimators when the calibration is based on estimated totals receives further comment in the discussion section below.

3 Longitudinal Calibration

Material in this section is repeated and reorganized from Larsen *et al.* (2011). It provides necessary background for understanding the proposal for longitudinal calibration estimation and weighting. Larsen *et al.* (2011) contains details on the simulation performed for that paper.

The principle motivation for creating longitudinal weights is a desire to be able to take multiple survey years together. Combining data from survey years increase sample size versus a single cohort. Although the NSF SDR survey is large by most standards, the number of individuals in certain discipline by rank by demographic group combinations in a single survey year can be small. One complication with combining data from different survey years is that each individual in each year has survey weight for that year.

Calibration weights for estimation with longitudinal data in the National Long Term Care Survey (NLTCS; http://www.nltcs.aas.duke.edu/) has been considered by Ash (2005). Cross-sectional weights for this survey are computed so that weights sum to population totals. This is an example of classical post stratification. When the interest is the difference between totals at two time points, there are two sets of population totals (earlier totals, later totals) that are available. Ash (2005) uses calibration estimation to adjust weights for both sets of known total controls. The author investigated one- and two-step calibration approaches, which differ in whether the various calibration totals are used simultaneously or one after another in weight adjustment. The NLTCS uses repeated replications in variance estimation.

The interest in the current paper differs from the interest of Ash (2005) in a few important ways. First, the goal here is to use several survey years together, not only two. Second, the known population totals are not available; rather, estimated totals can be produced in each survey year. Third, a broader set of estimands is being considered; these are describe further below. Otherwise, the current paper shares much of the same interest as the paper by Ash (2005).

Three requirements are considered when producing longitudinal weights. First, the weight needs to be calculable from existing data, which means either the public use data sets or the restricted use versions that NSF releases under strict licensing. The exact population totals and the exact definition of post stratification cells are not known to the researchers outside of the organization that produced the data. Second, the weight needs to be useful for reproducing key cross-sectional analyses. This is both a requirement for consistency and an attempt to produce advantages in estimation via correlations. If a calibrated set of weights could not reliably reproduce analyses of interest (not with exact correspondence necessarily but with reasonable proximity in some metric), then users would be unlikely to utilize the new weight set. Third, the weight should be low in variability, because high variability weights are associated with low precision in estimation. The third requirement potentially affects all weight adjustment procedures and applications. In the area of nonresponse adjustment, fine adjustments to weights often have the potential to remove more nonresponse bias than coarse adjustments, but the resulting weights are often more variable, which can negatively affect the standard errors for some estimators.

The process of calibrating cross-sectional weights to produce a set of longitudinal weights for analysis of data from combined survey years can be divided into five steps.

1. Selection of initial weights for each subject that appears in at least one survey year.
2. Selection and computation/estimation of control targets from one or more survey years.
3. Selection of a calibration method from the available options. Some calibration methods require making choices such as minimum and maximum allowable weight.
4. Computation of calibrated weights.
5. Evaluation of the calibrated weights in terms of analyses of interest. The evaluation includes computation of point estimates as well as standard errors.

Table 1: Prototype scenario for longitudinal weighting.

Year	Year 1	Year 2	Year 3
Population	U1	U2	U3
Domain	d1	d2	d3
Variables	X1, Y1	X2, Y2	X3, Y3
Sample	s_1	s_2	s_3

Table 2: Overlap of populations in prototype scenario for longitudinal weighting. Simulation population sizes. Row numbers pertain to left portion only.

	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
Row	U1	U2	U3	U1	U2	U3
1	x			1000	0	0
2	x	x		1000	1000	0
3	x	x	x	6000	6000	6000
4	x	x	x	0	1000	1000
5		x		0	0	1000
6		x	x			
7			x	$N_1 = 8000$	$N_2 = 8000$	$N_3 = 8000$

What analysis would benefit from considering a composite population comprised of individual, overlapping populations from multiple survey years? One analysis that should clearly benefit from using subjects sampled in all years would be a regression of Y on X over the time periods. The composite population sample should have larger sample size and more observations than any one year sample. Discusssion of this analysis can be found in Larsen *et al.* (2011).

3.1 Prototype Population

Table 1 illustrates a prototype scenario for a cross-sectional survey. The populations in years 1, 2, and 3 are U_1, U_2, and U_3, respectively. Within each population is a domain or subpopulation of interest, $d_j \subset U_j$, such as female doctorate recipients, recent graduates, minority doctorate recipients, or graduates with a degree is a specific field of study. Variables measured in the population can be numerous, but for estimation and calibration work they will be divided into two sets in survey year j: X_j are variables used as covariates or control variables, Y_j are outcome variables of interest to the study. Within each population, a sample is selected: $s_j \subset U_j$ in survey year j.

The populations overlap as depicted in left portion of Table 2. The rows are not intended to be proportional to population size. Rows 1-4 denote the population in survey year 1. Rows 2-6 denote the population in survey year 2. Rows 3-4 and 6-7 denote the population in survey year 3. Some elements in the three populations appear in only one survey year: row 1 in year 1, row 5 in year 2, and row 7 in year 3. Other elements appear in two of the three populations: row 2 in years 1 and 2 and row 6 in years 2 and 3. In some applications, such as labor force surveys, elements could appear in years 1 and 3, but not in year 2. Such a scenario is not considered in this work, but should fit within the general framework proposed below. Other elements, represented by rows 3 and 4, exist in all three populations. If the populatoin size each year is $N_1 = N_2 = N_3 = 8000$, each year 1000 individuals enter the population, and each year 1000 leave the population, then the right portion of Table 2 gives population sizes illustrating the sizes of overlaps across years. The rows do not necessarily correspond to rows in previous tables.

The sampling design for the Survey of Doctorate Recipients is described on the National Science Foundation NCSES

Table 3: Prototype sampling design for prototype scenario for longitudinal weighting. x means that the units were not in the population that year. Sample weights computed cross-sectionally within strata in prototype scenario for longitudinal weighting. Weighting formulas can differ by strata. Final column is the composite weight for three survey years together.

Row	Year	Year 1	Year 2	Year 3	**Composite**
	Population	U1	U2	U3	U
1	stratum 1	s_1, w_1	x	x	w
2	stratum 1	s_1, w_1		s_{34}, w_3	w
3	stratum 1	s_1, w_1	s_{21}, w_2	x	w
4	stratum 1	s_1, w_1	s_{21}, w_2	s_{31}, w_3	w
5	stratum 2	x	s_{22}, w_2		w
6	stratum 2	x	s_{22}, w_2	s_{32}, w_3	w
7	stratum 3	x	x	s_{33}, w_3	w

(2011) website. The prototype sampling design is depicted in Table 3. The rows are not intended to be proportional to sample size. The sample in survey year 1 is $s_1 \subset U_1$, which is represented in rows 1-4. The sample in survey year 2 is $s_2 = \{s_{21}, s_{22}\} \subset U_2$ and is represented in rows 3-6. Elements in rows 3 and 4 that were selected in s_1 are included again in s_2. Together they are denoted $s_{21} =\subset s_2$. Other elements in U_2 are selected for the survey year 2 sample from elements in the population in U_2 that were not in the population in year U_1. The subset $s_{22} \subset s_2$ with $s_{22} \subset U_2 \setminus U_1$ is in rows 5 and 6. These elements correspond to new PhD's in the Survey of Doctorate Recipients; they received their degrees and entered the survey target population after the years included in survey year 1.

The x's in the table indicate that the population in the given column (survey year) did not include the elements covered by the rows. For example, rows 5-7 represent elements that were not members of population U_1, rows 1 and 7 were not in population U_2, and rows 1, 3, and 5 were not in population U_3. Not depicted in the table are members of the population there were not sampled. For example, the elements not sampled in survey year 1 are $U_1 \setminus s_1$.

The sample in survey year 3 can be found in rows 2, 4, 6, and 7. Elements in row 2 are selected from those that were selected in years 1 and 2 ($s_{31} \subset s_{21} \subset s_1$). Units in row 6 ($s_{32}$) are selected from the elements that were new to the population in survey year 2 and selected in $s_{22} \subset s_2$. Units in row 7 (s_{33}) are selected from the new members of population U_3. Additional units (row 2, s_{34}) are selected from $U_1 \bigcap U_3$ that were selected in year 1, but not in year 2.

The set s_1 is sampled from stratum 1, which is U_1. The set s_{22} is sampled from stratum 2, which is $U_2 \setminus U_1$. The set s_{33} is sampled from stratum 3, which is $U_3 \setminus (U_1 \cup U_2)$. Note that $s_{21} \subset s_1$ and $s_{31} \subset s_{21}$ are taken from stratum 1, s_{32} is taken from stratum 2 ($U_2 \setminus U_1$; $s_{32} \subset U_3 \bigcap U_2 \setminus U_1$), and s_{34} is drawn from stratum 1 ($U1$; $s_{34} \subset s_1$, $s_{34} \cap s_{31} = \emptyset$, $s_{34} \subset U_1 \bigcap U_3$). Sampling rates for the simulation will be determined within strata.

Table 3 presents cross-sectional weights that would be determined for each survey year. Weighting formulas can differ by strata. Each year a subject is included in the sample it receives a weight. The final column of Table 3 illustrates the goal of a composite or single weight for each subject included in one or more of the samples in survey years 1, 2, and 3.

3.2 Calibration Options

Step 1 in the calibration procedure is to choose initial weights. For initial weights, four options are being considered: (1) Equal weighting for elements in $s = s_1 \cup s_{22} \cup s_{33}$. (2) The earliest available weight (w_1 for s_1, w_2 for s_{22}, w_3 for s_{33}). (3) The average of available weights for each case. (4) The latest available weight (w_3 for s_3, w_2 for s_2 excluding s_3, w_1 for the rest). Step 2 in the process of calibrating cross-sectional weights to produce a set of longitudinal weights for analysis of data from combined survey years is to identify targets for calibration. Potential targets that could be used singly or in combination include: (A) Population sizes N_1, N_2, N_3. (B) X total estimates ($\hat{t}_{X1}, \hat{t}_{X2}, \hat{t}_{X3}$). (C) Domain sizes ($N_{d1}, N_{d2}, N_{d3}$). (D) X total estimates in the domain ($\hat{t}_{X1d}, \hat{t}_{X2d}, \hat{t}_{X3d}$). In the simulation reported in Larsen *et al.* (2011), some combinations of calibration control totals were used. The sets of control totals were (1) A, (2) A and B, (3) A and C, (4) A, B, and C, and (5) A through D. Some are known values, such as population sizes, whereas others are estimates themselves. Others, including second moments and interactions among variables, could have been possible.

A difference between this simulation and application to the actual NSF Survey of Doctorate Recipients, or to any other survey for that matter, is that there could potentially be several domains and auxiliary variables to consider. It is an open question as to how many variables can or should be used in survey weight calibration. In general, calibrating on many variables has the potential to increase variability of resulting weights, which could dramatically increase standard errors for some estimates.

Step 3 is to select a calibration method. Only two were considered in Larsen *et al.* (2011): raking and linear regression calibration. Both are implemented in the R package `survey`, which addresses Step 4.

One of the requirements of the calibrated weights is that the the weight needs to be useful for reproducing key cross-sectional analyses. This is given as both a requirement for consistency and an attempt to produce advantages in estimation via correlations. In addition, it is of interest to examine the impact of weighting on a longitudinal analysis. Estimands and corresponding estimators considered for evaluation are listed below. These options were considered in Larsen *et al.* (2011).

1. Means in year j: estimation using sample s_j and new weights w, $j = 1, 2, 3$. Comparison is made to estimation using sample s_j and weights for sample year j, w_j.
2. Domain means in year j: estimation using sample $s_j \cap d_j$ and new weights w, $j = 1, 2, 3$. Comparison is made to estimation using sample $s_j \cap d_j$ and weights for sample year j, w_j.
3. Change in means: estimation using cases sampled in both years.
4. Change in domain means: estimation using cases sampled in both years.
5. Linear mixed effects model estimate of slope in population U: estimation of regression slope using single stage cluster sample.

3.3 Simulation Study

The simulation study in Larsen *et al.* (2011) was implemented as follows. The population, sample, weighting, and variable details described therein were utilized. Conduct the following steps $b = 1, \ldots, B = 1000$ times:

1. Generate a population in years 1, 2, and 3 from the models given above.
2. Select a sample in years 1, 2 and 3 according to the stated sampling scheme.
3. Compute and estimate control totals.
4. For each combination of starting weights and groups of control totals, compute calibration weights using raking. Raking cannot be used when methods A through D are used together due to the interaction between domain size and domain total.
5. For each combination of starting weights and groups of control totals, compute calibration weights using linear regression calibration. All groups of controls can be used with linear regression calibration.

6. Estimate each estimand and its standard error using each set of calibrated weights.

Results of the simulation were given in Larsen *et al.* (2011). As reported in that article, the proposed estimation methods seem to work well. One suggestion from that article is to consider ways to properly account for uncertainty due to estimated contorl totals in estimation with calibrated weights. Propagation of uncertainty in another scenario, namely, analysis of files created through record linkage, was considered by Lahiri and Larsen (2005). Development of methods for improved variance estimation will be reported in subsequent work.

4 Application to the SDR

Methods were applied to multiple survey years of the NSF Survey of Doctorate Recipients. Longitudinal calibration was implemented for either three survey years or five survey years. The combination of three survey years was 1993, 1995, and 1997. The combination of five survey years added 1999 and 2001 to the trio used previously. The entire SDR sample was used in calibrating weights. The response variable chosen for analysis is the respondent salary. Two domains of interest were females and minorities. Both variables are binary variables in this analysis. Different combinations of calibration factors were used as described below. Computations were performed using the `survey` package (Lumley 2011) in R (2008). Linear regression calibration was used in all cases. No negative weights were encountered. Replication variance estimation methods were not used in this study as the control totals were treated as if they had been known before calibration. This is reasonable in this case, because the population numbers presumably would have been known by those designing the sampling plan for the survey.

Calibration totals were chosen to be population size totals for the population in the chosen survey years and for a domain in the chosen survey years. Three calibration combinations were considered when three surveys were used together in calibration weighting.

1. Calibrate on the population total only in years 1993, 1995, and 1997. The population total in each year was taken to be the sum of the survey (expansion) weights in each year.
2. Calibrate on the population total and the number of females (the size of the female domain group) in years 1993, 1995, and 1997. Implicitly one then calibrates on the number of males (the size of the male domain group) in those years as well.
3. Calibrate on the population total, the number of females (the size of the female domain group), and the number of minorities (the size of the minority domain) in years 1993, 1995, and 1997.

The same three calibration combinations were considered when five surveys were used together in calibration weighting. For the five survey application, however, totals in years 1993, 1995, 1997, 1999, and 2001 were used. Thus, option 1 calibrated to three (five) totals, option 2 calibrated to six (ten) totals, and option 3 calibrated to nine (fifteen) totals in the three (five) survey year application.

Means and standard errors were computed for the average salary overall, for females, and for minorities by survey year. Table 4 reports results for the the average salary overall. Estimated means, standard errors, and percent difference in means in 1993, 1995, 1997, 1999, and 2001 surveys are reported. Results are reported for different combinations of calibration targets. Calibration used data from five surveys together or three surveys together. The original mean estimates and standard errors are based on single surveys.

First, comparing the result of calibrations in the case of three survey years versus the case of five survey years, it is clear that the calibrated means of average salary from the three surveys are much closer to the original means of salary than are the calibrated means of average salary from the five surveys. That is, for the population mean overall, the percent difference between the original means and the calibrated means are smaller then three surveys are used instead of five surveys. This makes sense because with more surveys the weights need to be modified more to match the additional control population size totals.

Second, as the number of calibration totals is increased, in either the three survey or five survey application, the percent difference between the original means and the calibrated means decreases. This result is consistent across years and both numbers of surveys.

Third, standard errors tend to be larger for the calibrated data than originally. For estimating salary in a given year, as estimating is implemented here, there is no increase in sample size with the calibrated data. An alternative, such a generalized least squares regression (e.g., Breidt and Fuller 1999), might realize an advantage due to correlations over time. The increase in standard errors makes sense, because the calibration weighting tends to make weights more variable, which tends to lead to higher variability of estimators. The effect is seen less for the three survey application than for the five survey application.

Table 5 reports results for the mean salary among females. The percent difference between the original and calibrated mean estimates are small, generally less than one-and-a-half percent. For the female group, in contrast to the situation overall, adding control totals does not seem to appreciably impact calibrated standard errors. It also does not seem to impact the percent difference in means. As with the overall mean, standard errors tend to be larger for the calibrated data than originally. The effect is greater for the five survey application than for the three survey application.

Table 6 reports results for the mean salary among minorities. The results for the mean salary among minorities are consistent with those for the overall mean salary reported in Table 4. the calibrated means of average salary for minorities from the three surveys are much closer to the original means of salary than are the calibrated means of average salary from the five surveys. That is, for the minority mean overall, the percent difference between the original means and the calibrated means are smaller then three surveys are used instead of five surveys. As the number of calibration totals is increased, in either the three survey or five survey application, the percent difference between the original means and the calibrated means for minority average salary decreases. Standard errors tend to be larger, more so for the five survey application than for the three survey application, for the calibrated data than originally.

Overall, the calibrated weights do well in the application. The percentage of difference between the calibrated means and the original means are almost all smaller than 1.5%.

5 Discussion

The proposed method for computing longitudinal survey weights from cross sectional survey weights using calibration weighting was applied to NSF SDR data from five years. Initial evidence suggests that calibration can create useful longitudinal weights. Weights preserve means by year and domains without inflating standard errors much in these preliminary applications. It is anticipated that as more control totals, especially estimated control totals, are added to the calibration targets that methods to properly account for variance will make a bigger difference from naive variance estimation methods.

As described in Larsen *et al.* (2011), a critical question is, how should one estimate variance when calibration totals are in fact themselves estimated? The survey estimates used as control totals have their own uncertainty that should be propagated into the standard errors. It is hypothesized that variance estimation with longitudinally calibrated survey weights must take into account the fact that some of the target control values are *estimated* from the separate surveys rather than based on a known population value. The NSF SDR utilizes Generalized Variance Functions (GVFs) for variance estimation (Jang 2001), but replicate weights are available under a restricted use license.

Dever and Valliant (2010) cite examples of surveys in which researchers have estimated control totals and then used post stratification. Dever and Valliant (2010) then compare methods of variance estimation in this context. Elliott et al. (2010) combine samples from two sources in order to improve estimation. In order to combine samples, the authors estimate weights that they refer to as pseudo-weights. In order to incorporate uncertainty due to weight estimation, the authors use a jackknife approach. Breidt and Opsomer (2008) study post stratification where the post strata are formed based on an estimated classification function. They call this endogenous post stratification (ESP). These and other sources could be informative for the issue of variance estimation when control totals are estimated with uncertainty.

Future work will expand the application to the NSF Survey of Doctorate Reciptient data for the puspose of studying career paths of doctoral recipients in Science, Health and Medicine, and Engineering.

References

Ash, S. (2005). Calibration weights for estimators of longitudinal data with an application to the National Long Term Care Survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*. American Statistical Association: Alexandria, VA, 2694–2699.

Breidt, F. J., and Fuller, W. A. (1999). Design of supplemented panel surveys with application to the National Resources Inventory. *Journal of Agricultural, Biological, and Environmental Statistics*. 4(4): 391–403.

Breidt, F. J., and Opsomer, J. D. (2008). Endogeous post-stratification in surveys: Classifying with a sample-fitted model. *Annals of Statistics*. 36(1): 403-427.

Dever, J. A., and Valliant, R. (2010). A comparison of variance estimators for poststratification to estimated control totals. *Survey Methodology*. 36(1): 45–56.

Deville, J. C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418): 376–382.

Deville, J. C., and Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423): 1013–1020.

Duncan, G. J., and Kalton, G. (1987). Issues of Design and Analysis of Surveys Across Time. *International Statistical Review*, 55, 97–117.

Elliott, M. R., Resler, A., Flannangan, C. A., and Rupp, J. D. (2010). Appropriate analysis of CIREN data: Using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes. *Accident Analysis and Prevention*, 42, 530–539.

Fuller, W. A. (1999). Environmental surveys over time. *Journal of Agricultural, Biological, and Environmental Statistics*. 4(4): 331–345.

Jang, D. S., Cox, B. G., Edson, D., and Satake, M. (2001). Sampling Errors for SESTAT: 1993, 1995, 1997, and 1999. *Mathematica Policy Research Report 8797-410.*
http://www.nsf.gov/statistics/sestat/stderr99.pdf. [Accessed September 26, 2011].

Kim, J. K. (2010). Calibration estimation using exponential tilting in sample surveys. *Survey Methodology*, 36(2): 145–155.

Kim, J. K., and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78(1): 21–39.

Lahiri, P., and Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*. 100(469): 222–230.

Larsen, M.D., Foulkes, M.A., Qing, S., and Zhou, B. (2011). Calibration Estimation and Longitudinal Survey Weights: Application to the NSF Survey of Doctorate Recipients. *Proceedings of the Survey Research Methods Section, ASA*.

Lumley, T. (2011). survey: analysis of complex survey samples. R package version 3.24-1.

McDonald, T. L. (2003). Review of environmental monitoring methods: Survey designs. *Environmental Monitoring and Assessment*. 85(3): 277–292.

Mirel, L. B., Burt, V., Curtin, L. R., and Zhang, C. (2010). Different approaches for non-response adjustments to statistical weights in the continuous NHAHES (2003-04). *Federal Committee on Statistical Methodology Research Conference*.

National Science Foundation, Division of Science Resources Statistics. (2009). *Characteristics of Doctoral Scientists and Engineers in the United States: 2006*. Detailed Statistical Tables NSF 09-317. Arlington, VA. Available at http://www.nsf.gov/statistics/nsf09317/.

National Science Foundation, Division of Science Resources Statistics. (2011). *Unemployment Among Doctoral Scientists and Engineers Remained Below the National Average in 2008.* Arlington, VA (NSF 11-308). http://www.nsf.gov/statistics/infbrief/nsf11308/.

National Science Foundation, National Center for Science and Engineering Statistics (NCSES) [formerly the Division of Science Resources Statistics (SRS)]. (2011). *Survey of Doctorate Recipients.* http://nsf.gov/statistics/srvydoctoratework/. Accessed 2011-09-22.

R Development Core Team (2008) *R: a language and environment for statistical computing*, V2.7.2. Vienna: R Foundation for Statistical Computing.

Research Triangle Institute (2008). *SUDAAN Language Manual, Release 10.0.* Research Triangle Institute: Research Triangle Park, NC.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2): 99–119.

U. S. Census Bureau. (2006). *Current Population Survey, Design and Methodology*. Technical Paper 66. U.S. Government Printing Office, Washington, DC.

U. S. Census Bureau. (2009). *Design and Methodology, American Community Survey*. U.S. Government Printing Office, Washington, DC.

Zhang, L. C. (2000). Post-stratification and calibration - A synthesis. *American Statistician*, 54(3): 178–184.

Table 4: Estimated means, standard errors, and percent difference in means in 1993, 1995, 1997, 1999, and 2001 surveys. Results are reported for different combinations of calibration targets. Calibration used data from five surveys together or three surveys together. The original mean estimates and standard errors are based on single surveys.

Calibration on population

Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	64135	219.07	63203	181.03	1.47
1995	64444	235.46	63517	203.61	1.46
1997	67869	241.34	67426	229.49	0.66
1999	72081	247.71	71502	220.98	0.81
2001	78244	259.61	77786	234.24	0.59

Calibration on population and female size

Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	63839	216.92	63203	181.03	1.01
1995	64094	233.75	63517	203.61	0.91
1997	67633	239.54	67426	229.49	0.31
1999	71629	246.36	71502	220.98	0.18
2001	77736	258.19	77786	234.24	-0.06

Calibration on population total, female and minority

Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	63832	216.9	63203	181.03	1.00
1995	64087	233.76	63517	203.61	0.90
1997	67640	239.62	67426	229.49	0.32
1999	71627	246.4	71502	220.98	0.17
2001	77732	258.23	77786	234.24	-0.07

Calibration on population total only

Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	63758	198.93	63203	181.03	0.88
1995	64213	215.2	63517	203.61	1.10
1997	67766	222.16	67426	229.49	0.50

Calibration on population and female

Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	63594	197.87	63203	181.03	0.62
1995	63984	214.34	63517	203.61	0.74
1997	67554	221.17	67426	229.49	0.19

Calibration on population, female and minority

Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	63585	197.81	63203	181.03	0.60
1995	63976	214.27	63517	203.61	0.72
1997	67554	221.13	67426	229.49	0.19

Table 5: Estimated means, standard errors, and percent difference in means for FEMALES in 1993, 1995, 1997, 1999, and 2001 surveys. Results are reported for different combinations of Calibration targets. Calibration used data from five surveys together or three surveys together. The original mean estimates and standard errors are based on single surveys.

Calibration on population					
Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	51342	342.81	51487	274.15	-0.28
1995	50314	392.09	51079	362.28	-1.50
1997	53592	371	54134	347.04	-1.00
1999	56795	394.36	57492	376.82	-1.21
2001	62825	408.86	63427	394.88	-0.95

Calibration on population and female size					
Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	51333	342.18	51487	274.15	0.30
1995	50396	392.78	51079	362.28	-1.34
1997	53875	378	54134	347.04	-0.48
1999	56803	395.01	57492	376.82	-1.20
2001	62800	407.58	63427	394.88	-0.99

Calibration on population total, female and minority					
Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	51333	342.26	51487	274.15	-0.30
1995	50396	392.78	51079	362.28	-1.34
1997	53875	378	54134	347.04	-0.48
1999	56803	395.01	57492	376.82	-1.20
2001	62800	407.58	63427	394.88	-0.99

Calibration on population total only					
Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	51432	309.4	51487	274.15	-0.11
1995	50515	353.93	51079	362.28	-1.10
1997	53735	347.97	54134	347.04	-0.74

Calibration on population and female					
Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	51428	310.54	51487	274.15	-0.11
1995	50501	353.69	51079	362.28	-1.13
1997	53779	348.81	54134	347.04	-0.66

Calibration on population, female and minority					
Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	51426	310.33	51487	274.15	-0.12
1995	50498	353.43	51079	362.28	-1.14
1997	53777	348.78	54134	347.04	-0.66

Table 6: Estimated means, standard errors, and percent difference in means for MINORITIES in 1993, 1995, 1997, 1999, and 2001 surveys. Results are reported for different combinations of Calibration targets. Calibration used data from five surveys together or three surveys together. The original mean estimates and standard errors are based on single surveys.

Calibration on population

Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	57312	821.33	56701	654.04	1.08
1995	58193	849.59	57581	699.65	1.06
1997	61923	853.16	61997	806.74	-0.12
1999	65210	807.99	64279	745.38	1.45
2001	71284	821.5	70044	751.09	1.77

Calibration on population and female size

Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	57030	804.15	56701	654.04	0.58
1995	57938	838.94	57581	699.65	0.62
1997	61739	847.8	61997	806.74	-0.42
1999	64821	799.41	64279	745.38	0.84
2001	70804	814.6	70044	751.09	1.09

Calibration on population total, female and minority

Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	56910	787.21	56701	654.04	0.37
1995	57856	827.63	57581	699.65	0.48
1997	62024	851.19	61997	806.74	0.04
1999	64846	797.84	64279	745.38	0.88
2001	70756	811.42	70044	751.09	1.02

Calibration on population total only

Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	57578	765.06	56701	654.04	1.55
1995	58605	789.82	57581	699.65	1.78
1997	61967	783.81	61997	806.74	-0.05

Calibration on population and female

Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	57417	755.42	56701	654.04	1.26
1995	58411	783.42	57581	699.65	1.44
1997	61779	778.48	61997	806.74	-0.35

Calibration on population, female and minority

Survey	calibrated		original		% difference
Year	mean	SE	mean	SE	in means
1993	57350	745.55	56701	654.04	1.14
1995	58365	775.73	57581	699.65	1.36
1997	61969	778.8	61997	806.74	-0.05