Calibration Estimation and Longitudinal Survey Weights: Application to the NSF Survey of Doctorate Recipients

Michael D. Larsen * Siyu Qing † Beilei Zhou‡ Mary A. Foulkes§

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 explores calibration estimation for construction of such a longitudinal weight. 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. 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. Third, the weight should be low in variability, because high variability weights are associated with low precision in estimation. Choices of initial weights and calibration targets are compared in a series of analyses.

Key Words: Calibration weighting, longitudinal study, panel study, raking, SESTAT, survey sampling.

1. Introduction

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. The survey is used as the basis for reports such as NSF (2008, 2011). Every survey year the target population is a little bit different because people enter (e.g., new Ph.D. recipients in the U.S.) or leave (e.g., deaths) the population. Numerous variables are included in the data set. 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.

Survey weights adjust for oversampling and nonresponse on a cross-sectional basis. That means that analysis using the survey data with the survey weights in a given year is representative of a corresponding population. The survey weights

*The George Washington University, Department of Statistics, Biostatistics Center, 6110 Executive Blvd., Suite 750, Rockville, MD 20852 email: mlarsen@bsc.gwu.edu

†The George Washington University, Department of Statistics

‡The George Washington University, Biostatistics Center

§The George Washington University, Department of Epidemiology and Biostatistics, Department of Health Policy, Biostatistics Center

are not designed explicitly for longitudinal analysis of data sampled in different survey years. This fact does not mean, however, that no longitudinal analysis is possible. Indeed, 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. Despite this fact, there are no longitudinal weights for the survey that would enable estimation of statistical models or comparison of finite population characteristics using data from multiple survey waves together.

1.1 Longitunidal analysis and the SDR

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. For example, if one wants to estimate the probability of proceeding from PhD in 1991-1992 to postdoc to tenured faculty member within ten years in the biological, agricultural and environmental life sciences, one can examine the recent PhD graduates in this area sampled in the 1993 survey who can be linked over time in the 1995, 1997, 1999, 2001, and 2003 surveys. One then could look at the same question for other years, such as the 1993-1994 PhD graduates appearing first in the 1995 survey and linked through the 2006 survey.

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. For example, there could be graduates in from 1991-1992 who did not enter the survey until a survey year after 1993, such as 1995. Individuals such as these cannot be readily combined with the 1993 survey data, because their 1995 survey weights are designed only for cross-sectional estimation.

Another limitation occurs when estimating statistical models of change over time. Imagine estimating change in salaries over time (years 1991 to 2002, surveys 1993 to 2003) by field of study and demographic characteristics, such as sex, rank, Carnegie ranking of institution, and U.S. citizenship. 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

Before proceeding to describe calibration weighting to create longitudinal survey weights, it should be noted that some surveys directly plan for longitudinal, panel, or time series analysis.

The American Community Survey (ACS; http://www.census.gov/acs/www/) selects five years of household sample cases at once (U. S. Census Bureau 2009; chapter 4). Within each county, the sample for five years is selected all together and then split into five parts. Doing so produces consistent weights for combining sample respondents together. This is particularly important for estimation of characteristics in small places. The ACS is not longitudinal, however, because individuals are included in only one survey year of data collection.

The Current Population Survey (CPS; http://www.bls.gov/cps/; http://www.census.gov/cps/) is designed to measure the level of and changes in employment, unemployment, and labor force participation. The CPS is longitudinal

in that individuals are measured for four months initially and then for another four months after an eight month break (U. S. Census Bureau 2006). Longitudinal weights are discussed in chapter 10 (starting on page 10-14). These weights are constructed for flows based on population controls from the U. S. Census Bureau. As discussed below, such information is not available to researchers utilizing the NSF SDR data.

There are many other surveys – longitudinal surveys and panel surveys – that are designed to measure change over time. Several of these surveys plan survey estimation and weighting with this goal in mind. Examples of these surveys 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/). it is beyond the scope of this article to review the methodology utilized in these and other studies.

Surveys in the area of environmental monitoring are intended to enable estimation over time. One development in this area is generalized least squares estimation as in Breidt and Fuller (1999). Within each survey year, one could estimate the outcome for a variable of interest conditional on certain covariate variables. For each year, one then estimates variance based on sample and weights in a given year. One would then estimate the covariance between estimates in pairs of years. There is covariance that depends on the overlap of samples across time. An estimate of change is then computed as a function of these totals. The variance of the change estimate is then a function of the estimated variances and covariances. The USDA's National Resources Inventory is a survey that utilizes this methodology in estimation. Panel surveys and surveys over time are considered by Duncan and Kalton (1987), Fuller (1999), and McDonald (2003) and references therein. Comparison to these and other survey designs will be considered at a later time.

1.3 Outline

This paper explores calibration estimation (Deville and Särndal 1992 and references given below) for construction of longitudinal weights for cross-sectional sample surveys. Section 2 discusses calibration and formation of longitudinal survey weights from cross-sectional weights. Section 3 outlines a simulation study plan. Section 4 gives preliminary results. Section 5 discusses findings, limitations, and future work. The paper ends with references and acknowledgments, which include a disclaimer.

2. Calibration for Longitudinal Weighting

2.1 Calibration 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.

2.2 Longitudinal Calibration

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.

Section 3 presents the prototype scenario that is used in simulations and discusses the steps listed above in this context.

3. Simulated population, simulation parameters, calibration options, and estimands

3.1 Simulated 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

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	$N_1 = 8000$	$N_2 = 8000$	$N_3 = 8000$

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.

In the simulation, the population size in each year is taken to be $N_1 = N_2 = N_3 = 8000$. It is assumed that each year 1000 individuals enter and each year 1000 leave the population. The right portion of Table 2 gives population sizes illustrating the sizes of overlaps across years. Note that 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 (2011) website. The prototype sampling design is depicted in Table tab3. 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

Table 3: Prototype sampling design for prototype scenario for longitudinal weighting. x means that the units were not in the population that year.

Row	Year Population	Year 1 U1	Year 2 U2	Year 3 U3
1	stratum 1	s_1	x	x
2	stratum 1	s_1		s_{34}
3	stratum 1	s_1	s_{21}	x
4	stratum 1	s_1	s_{21}	s_{31}
5	stratum 2	x	s_{22}	
6	stratum 2	x	s_{22}	s_{32}
7	stratum 3	x	x	s_{33}

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 4 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 4 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.

In the simulation, $n_1 = 600$ subjects are randomly sampled (simple random sampling without replacement) in year 1. In year 2, $n_{21} = 400$ are randomly sampled from those in s_1 and $n_{22} = 200$ are selected from the 1000 new members of population U_2. So $n_2 = 600$ as well. In year 3, $n_{31} = 300$ are randomly sampled

Table 4: 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.

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

Table 5: Population simulation conditions. Individuals have their values randomly generated independently.

Domain $Z \sim$ Bernoulli(1/4)	
$X_1 \sim N(97.5 + 10Z, 30^2)$	$Y_1 \sim N(8000 + 10X_1 + 600Z, 140^2)$
$X_2 \sim N(X_1 + 10Z, 30^2)$	$Y_2 \sim N(9000 + 10X_2 + 800Z, 140^2)$
$X_3 \sim N(X_2 + 10Z, 30^2)$	$Y_3 \sim N(10000 + 10X_3 + 1000Z, 140^2)$

from those in s_{21}, $n_{32} = 100$ are randomly sampled from those in s_{22}, $n_{33} = 150$ are randomly sampled from the 1000 new members of population U_3, and $n_{34} = 50$ are selected from the 200 units selected originally in year 1, but not picked in year 2 ($s_{34} \subset U_3 \bigcap (s_1 \setminus s_{21})$). Table 6 illustrates the sample sizes in the simulation study.

Survey weights in the simulation are computed as the inverse of the probability of selection within strata. That is, the weight is N/n where N is a relevant population size and n is the sample size. Table 7 gives initial survey weights. The lower weights among new additions to the populations in years 2 and 3 reflect oversampling.

It remains to give details of how the population variables are to be simulated. A domain of interest will be determined by a variable Z with $Z = 1$ meaning that the subject is a member of the domain and $Z = 0$ indicating non membership. Variable Z is generated as a Bernoulli random variable with probability $\pi = 0.25$. Auxiliary variables X_1, X_2, and X_3 and outcome variables Y_1, Y_2, and Y_3 are generated from univariate normal distributions as given in Table 5.

Future work simulations could consider a number of modifications. Of interest would be a smaller domain ($\pi < 0.25$), correlations between X's, between Y's, and between X's and Y's that are weaker or stronger, and non normal data.

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 ($N/n = 10000/800$) 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

Table 6: Prototype sampling design for prototype scenario for longitudinal weighting: sample sizes for simulation study. Samples in years 1 and 2 are listed multiple times in order to illustrate the relationship to later years.

Year	Year 1	Year 2	Year 3
Population	U1	U2	U3
stratum 1	s_1, n_1=600	0	0
stratum 1	s_1, n_1=600	0	s_{34}, n_{34}=50
stratum 1	s_1, n_1=600	s_{21}, n_{21}=400	0
stratum 1	s_1, n_1=600	s_{21}, n_{21}=400	s_{31}, n_{31}=300
stratum 2	0	s_{22}, n_{22}=200	0
stratum 2	0	s_{22}, n_{22}=200	s_{32}, n_{32}=100
stratum 3	0	0	s_{33}, n_{33}=150
Total	$n_1 = 600$	$n_2 = 600$	$n_3 = 600$

Table 7: Prototype sampling design for prototype scenario for longitudinal weighting: initial sample design weights for simulation study.

Year	Year 1	Year 2	Year 3
Population	U1	U2	U3
stratum 1	s_1, $w_1 = \frac{8000}{600} = 13.3$	0	0
stratum 1	s_1, $w_1 = \frac{8000}{600} = 13.3$	0	s_{34}, $w_{34} = \frac{1000}{50} = 20$
stratum 1	s_1, $w_1 = \frac{8000}{600} = 13.3$	s_{21}, $w_{21} = \frac{7000}{400} = 17.5$	0
stratum 1	s_1, $w_1 = \frac{8000}{600} = 13.3$	s_{21}, $w_{21} = \frac{7000}{400} = 17.5$	s_{31}, $w_{31} = \frac{6000}{300} = 20$
stratum 2	0	s_{22}, $w_{22} = \frac{1000}{200} = 5$	0
stratum 2	0	s_{22}, $w_{22} = \frac{1000}{200} = 5$	s_{32}, $w_{32} = \frac{1000}{100} = 10$
stratum 3	0	0	s_{33}, $w_{33} = \frac{1000}{150} = 6.7$
Total of Weights	$w_1 = 8000$	$w_2 = 8000$	$w_3 = 8000$

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, some combinations of calibration control totals are used. The sets of control totals are (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 be 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 are being considered in this work: raking and linear regression calibration. Both are implemented in the R package `survey`, which addresses Step 4. Ash (2005) considered one-step or two-step calibration. Here we consider calibration in a single step. A further option that could be considered in later work is calibration after a logarithmic transformation of the weights. Such as transformation ensures that all weights are positive. Calibration methods receive further comment in the discussion.

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. The last option that is listed below is discussed further in the next section.

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 A focal analysis and an associated question about modeling in finite populations

What analysis would benefit from considering population $U = U_1 \cup U_2 \cup U_3$? One analysis that would clearly benefit from using subjects sampled in all three years

would be a regression of Y on X over the three time periods. The composite population sample should have larger sample size and more observations than any one year sample.

If the data were generated from a longitudinal study without reference to a finite population, one likely would use a linear mixed effects model with random effects for individuals. Each individual could have up to three measurements over time. How should such a modeling endeavor be presented in the context of a finite population framework?

One possibility is to consider a generalized least squares estimation as in Breidt and Fuller (1999). Within each stratum or subsample, one could estimate the average salary in a given year conditional on certain variables, such as demographics and field of degree. For each estimated average, one then estimates variance based on sample and weights in a given year. One would then estimate the covariance between estimates in pairs of years. An estimate of change is then computed as a function of these totals. The variance of the change estimate is then a function of the estimated variances and covariances. See also Duncan and Kalton (1987), Fuller (1999), and McDonald (2003) and references therein. Future work will explore recent references and this approach.

Another option would be to consider each subject to be a cluster and each cluster to consist of measurements over time. One then could use a cluster analysis variance formula with the `svyglm` function in the R `survey` package to estimate a linear model with time and other variables as predictors. Is a linear mixed effects growth model with random effect for subject really equivalent to a single stage cluster sample with each subject being a cluster and the model including a time predictor? In other words, how should one implement repeated measures or growth curve models in a finite population survey context? Future work will examine the correspondence and possible lack thereof between linear mixed effects growth models and cluster sampling with a linear model with time covariates.

4. Simulation Results

4.1 Simulation Study Implementation

The simulation study was implemented as follows. The population, sample, weighting, and variable details describe above 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.

The results are summarized by computing the average of estimates, the standard deviation of estimates, the average of estimated standard errors, the standard deviation of estimated standard deviations, and the percent of simulations in which a confidence interval for the estimand covers the true value in the composite population.

4.2 Summary of Results

Results are given in summary fashion only in this paper. Future presentations will present numerical and graphical summaries as well as extensive tables. Estimation in general seems to work well, but there is one dominant issue that is being addressed in ongoing work. In short, it appears that it is very important for variance estimation to take into account the fact that some control totals are estimated from survey data. Estimated control totals have their own uncertainty, which needs to be propagated in analyses. Some literature on this topic is reviewed in the discussion section below. It will make more sense to present more extensive results once methods for properly accounting for uncertainty in estimation with calibrated weights is incorporated into analyses. Propagation of uncertainty in another scenario was considered by Lahiri and Larsen (2005).

For the calibration methods (raking and linear regression) and the initial weight options (the four listed above) considered, very similar results were obtained. That is, estimates and estimated standard errors differed in a minor amount across the method-weight combinations. There are two differences to mention in comparing raking and linear regression. Raking does not produce negative weights, but it was possible for linear regression calibration to produce negative weights. Negative weights can be used in estimation, but in general they are not desirable. One cannot interpret calibration weights in the same way as one tends to interpret survey design weights or nonresponse adjusted weights; namely, as indicating the portion of the population that the observation associated with the weight represents. Calibration weights are supposed to be close to the initial weights but also satisfy the calibration constraints. The raking option, however, could not handle the full combination of options A-D; the R program ran into problems with the implied interaction among the X and domain indicator variables. Negative weights and choices for weight restrictions are mentioned in the discussion.

After calibration using data and targets from three years, estimates in a given survey year using the new weights accurately reproduce estimates from a single survey year using the original sampling weights for that year. Estimates of change (1 versus 2, 1 versus 3, or 2 versus 3) also are preserved.

When population size and domain size are used as calibration targets, estimated standard errors for yearly totals and change between years are approximately the same as before. When calibration targets include the X-variable total or both the overall and the domain X-variable totals, standard errors are estimated to be much smaller (e.g., 60-80% of the value) than the original estimated standard errors. In general smaller standard errors is desirable. In this case, however, coverage of the known population values by confidence intervals based on the calibrated standard error estimates is lower than the nominal 95% level (e.g., 70-85% coverage). A reduction in coverage below the nominal level is not desirable. This drop in coverage is discussed in the next section.

For the focal longitudinal analysis, a mixed effects model was fit to the population data. The fixed effects slope parameter estimates were compared to estimates from the sample data, where estimation was implemented as described above in Section 3.3. The model being estimated has a slope which is multiplied by year (1, 2, or 3). There is a clear benefit to using data from all three survey years instead of data only from year one. Data for subjects collected first in survey year 2 have data in years 2 and 3 or in year 2 only. If they have data in years 2 and 3, then their data is informative about the slope. If they are collected only in one year, then their data is informative about the intercept. The estimate of the slope is nearly the same as with only year one data with a much lower standard error. There is adequate confidence interval coverage when calibrating on population or domain size. Calibration on X totals reduces standard error further and reduces coverage a little bit.

5. Discussion

A critical question is, why was there a drop in confidence interval coverage with calibration on X totals? The likely reason is that calibration is being implemented to control weights to a survey estimate rather than to a population total. As mentioned, the survey estimates 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. Resampling methods have been considered by some authors in similar contexts. There are replicate weights for the restricted use NSF SDR data; the replicate weights must be requested separately from the usual restricted use data.

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 study the estimated-control post stratified estimatior (ECPE). The authors present a linearization variance estimator and three delete-one jackknife variance estimators. The results of their work support the need for development of theory and methods in this area.

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. For each jackknife sample, the authors re-estimate the pseudo-weights. In their simulation, the jackknife approach reduces or eliminates undercoverage of 95% confidence intervals.

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). Under certain conditions including a fixed number of parameters in the classification model, the authors demonstrate consistency of estimation and a central limit theorem. In simulations, they show scenarios in which the estimated classification aspect of ESP estimation (EPSE) has a small effect. The authors simulate mean squared errors (MSE) under three methods, but do not discuss variance estimation or confidence interval coverage.

The calibration ideas were applied to a few variables for three years (1993, 1995, and 1997) from the NSF SDR public use data files. Initial evidence suggests that calibration can create useful longitudinal weights. Weights preserve means and

group sizes by year 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.

Three further issues can be mentioned for consideration. First, although the populations from different survey years are being combined, there are some underlying true population exclusions. For example, some individuals, such as recent Ph.D. recipients, are not members of the population until they obtain their Ph.D. Analyses might logically exclude some individuals for some relationships due to this fact or due to their leaving the population.

Second, a more serious complication is variance estimation. The NSF SDR utilizes Generalized Variance Functions (GVFs) for variance estimation (Jang 2001). GVFs are functional relationships, which, in this case, are specific to the given survey year (by gender and major field). A multi-year analysis will need to consider what to do with the existence of multiple GVFs. An alternative is to use the NSF SDR replicate weights. One then must determine whether it is necessary to calibrate separately each set of replicate weights. This can be considered in the context of variance estimation methodologies such as Dever and Valliant (2010). GVFs were not an issue in the simulation.

Only two calibration methods were considered in this work: raking and linear regression calibration. Both were implemented in a single step as opposed to two or more steps (Ash 2005). It is sometimes possible with calibration to also restrict the range of weights. With certain distance functions, calibration equations, and weight restrictions, it is possible that there is no exact solution to the calibration problem. Some programs then seek the solution that minimizes discrepancies from the target controls. Presumably these solutions encounter one or more of the restrictions on weights. Another option that could be considered in later work is calibration after a logarithmic transformation of the weights. Such as transformation is used to ensure all positive weights. It is possible to place bounds on the weights after log transformation as well. Methods for choosing a distance function, a transformation, and weight bounds need to be developed in general and in the problem of longitudinal weighting in particular. Some distance functions and weight bounds were considered in Deville and Särndal's (1992) paper. The Research Triangle Institute (RTI 2008) implements a general methodology that enables this form of calibration. Of course, inherent in the use of calibration is the challenge of selecting variables and subgroups to define the control targets.

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.

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.

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.

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.