Evaluating 2003 NSCG Dual-Frame Estimates in Preparation for the 2010 NSCG

Donsig Jang,[1] David W. Hall[2*]
[1]Mathematica Policy Research, 600 Maryland Avenue, SW, Suite 550,
Washington, DC, 20024
[2]U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD, 20746

ABSTRACT

The National Survey of College Graduates (NSCG), sponsored by the National Science Foundation (NSF), is the nation's leading source of detailed statistics on the science and engineering labor force. Starting with the 2010 survey cycle, NSF plans to use multiple sampling frames to construct the NSCG. NSF had attempted to use a similar dual-frame approach for the 2003 NSCG, but the differing population estimates yielded by each frame led the foundation to switch to a single-frame approach. However, new research on the 2003 NSCG dual-frame design presents an opportunity to re-evaluate this decision. In this paper, we examine the issues associated with the 2003 NSCG dual-frame design, including what may have caused the differing estimates and how the single- and dual-frame estimates compare in terms of key characteristics of interest. The goal of this research is to gain a better understanding of the 2003 dual-frame estimates as NSF prepares for the 2010 NSCG.

Key Words: NSCG, dual-frame estimation

1. INTRODUCTION

Sponsored by the National Science Foundation (NSF), the National Survey of College Graduates (NSCG) is a major component of the Scientists and Engineers Statistical Data System (SESTAT), which captures data on all scientists and engineers in the United States. The SESTAT also includes two other component surveys: the National Survey of Recent College Graduates (NSRCG) and the Survey of Doctorate Recipients (SDR). The current SESTAT was established based on a design recommended by the Committee on National Statistics (CNSTAT) in late 1980's (National Research Council 1989). Since 1993, the U.S. Census Bureau has conducted the NSCG mostly biennially. The Census Bureau selected the initial sample from the long-form respondents to the 1990 Decennial Census, which at the time was ideal for collecting data on all U.S. scientists and engineers. However, because another full sampling frame was not available until the following decade, the NSRCG and SDR were also conducted to capture information on recipients of U.S.-earned bachelor's or higher degrees earned since the baseline survey in 1993. The NSRCG covered recent U.S. graduates with bachelor's and master's degrees with higher sampling rates than did the NSCG, and the SDR covered U.S. doctorates with a higher sampling rate than the NSCG.

[*] Work on this article was supported and funded by the National Science Foundation. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the National Science Foundation or the U.S. Census Bureau.

1.1. The NSCG During the 2000 Decade

Due to the timing of the 2000 Decennial Census and the availability of the long-form file, the first round of the SESTAT (including the NSCG) in the 2000's was not conducted until 2003. Similar to the 1993 NSCG, the 2003 NSCG attained full coverage because the sample was selected from the Decennial long-form respondents, which covered the entire U.S. population as of April 1, 2000. Specifically, this population covers:

- Those who received U.S.- or foreign-earned degrees (bachelor's or higher) in science and engineering (S&E) or S&E-related fields on or before April 1, 2000.
- Those who received U.S. or foreign-earned degrees (bachelor's or higher) in non-S&E fields on or before April 1, 2000, but were working in S&E or S&E-related fields as of the 2003 survey reference week (the week of October 1).

The Census Bureau followed up this initial sample in 2003 with supplemental samples from the 2003, 2006, and 2008 NSRCGs. During this decade, the bureau only followed respondents, and thus serious attrition effects were expected due to:

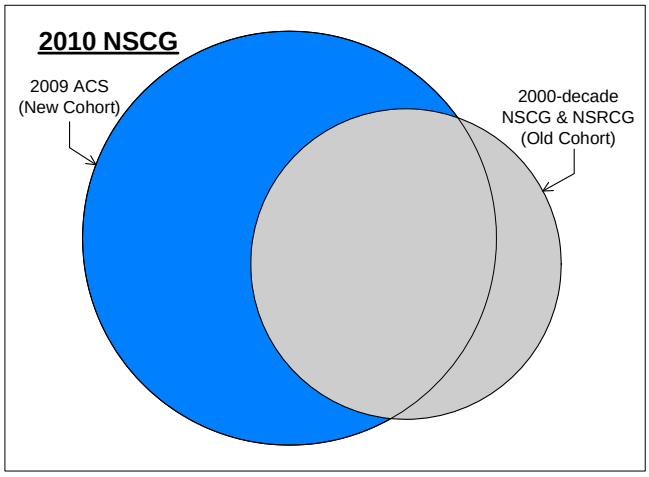
- Temporarily out-of-scope cases (for example, respondents who were abroad or institutionalized), and
- Nonresponse adjustments, which may not have corrected nonresponse bias entirely (cumulative nonresponse bias can be unduly high).

These attrition effects motivated the NSF to consider selecting a fresh sample from a sampling frame with full coverage. As described above, in past years the complete frame was available only once a decade from the Decennial Census. However, in 2005, the American Community Survey (ACS), which replaced the Decennial long-form survey in 2010, began producing a sample that represented the entire U.S. population. This led to a new NSCG sample design based on the ACS data, which can be used in each survey cycle as opposed to just once a decade.

The NSF, in consultation with the CNSTAT, is planning to use a rotating-panel sample design with four panels, which will be fully implemented by the 2016 survey round. The 2010 survey will be a transition survey round, with half of the sample from the 2000's NSCG respondents (the "old cohort") and the other half from the 2009 ACS sample data (the "new cohort"). This will make the NSCG 2010 sample a dual-frame sample, which will require dual-frame estimation. See Finamore, et al (2011) for more details on the rotating panel design.

The 2010 NSCG old cohort consists of sample members from several survey components: the majority of the sample from the 2000 Decennial long-form respondents, and supplemental samples from the 2001, 2003, 2006, and 2008 NSRCGs. This cohort has been used to capture up to 10 years of information on the 2003 NSCG sample and is thus a valuable source of longitudinal data.

Figure 1. The 2010 NSCG Sample



1.2. Dual-Frame Estimation

As mentioned above, the 2010 NSCG sample is a dual-frame sample consisting of the old cohort from the 2000's NSCG/NSRCGs and the new cohort from the 2009 ACS. We can therefore partition the entire 2010 NSCG target population into three components:

- $N \cap O^C$: covered by the new cohort only
- $N \cap O$: covered by both the new and old cohorts
- $N^C \cap O$: covered by the old cohort only

In the dual-frame sample design, a population total Y can be expressed as the sum of three "nonoverlapping" component totals: $Y = Y_{N \cap O^C} + Y_{N \cap O} + Y_{N^C \cap O}$. The estimation focus should then be on $Y_{N \cap O}$, which can be estimated by either $\hat{Y}^N_{N \cap O}$ based on the new cohort or $\hat{Y}^O_{N \cap O}$ based on the old cohort. If both estimators are approximately unbiased, the overall estimator can be obtained as a linear combination of two estimators for the overlapping component and single estimators for the other components: $\hat{Y} = \hat{Y}_{N \cap O^C} + \lambda \hat{Y}^N_{N \cap O} + (1-\lambda)\hat{Y}^O_{N \cap O} + \hat{Y}_{N^C \cap O}$, where $0 \le \lambda \le 1$. A dual-frame estimation generally focuses on combining two estimators for the overlapping component by determining λ. Hartly (1962, 1974) proposed a method to determine λ to minimize the variance of $\hat{Y}$, which was later improved by Fuller and Burmeister (1972) by adding a corrected term $\lambda_C(\hat{N}^N_{N \cap O} + \hat{N}^O_{N \cap O})$ to the estimator and determining λ and λ_C to minimize the variance of $\hat{Y}$.

Instead of determining the optimal linear multiplier for each variable or statistic, it is preferable to combine dual samples so that a single weight can be used to estimate all survey variables. Kalton and Anderson (1986) accomplished this by calculating sample inclusion probabilities in the overlapped population. Later, Skinner and Rao (1996) proposed a pseudo-maximum likelihood (PML) estimator, which allows the production of a single set of weights for estimation. Under a dual–random digit dialing (RDD) frame

setting, Brick et al. (2006) identified overlapping cases and divided their weights by two. Mecatti (2007) proposed a similar method for a more general multiframe sample setting. For details on dual-frame estimation in general, see Lohr (2007) and the references cited therein.

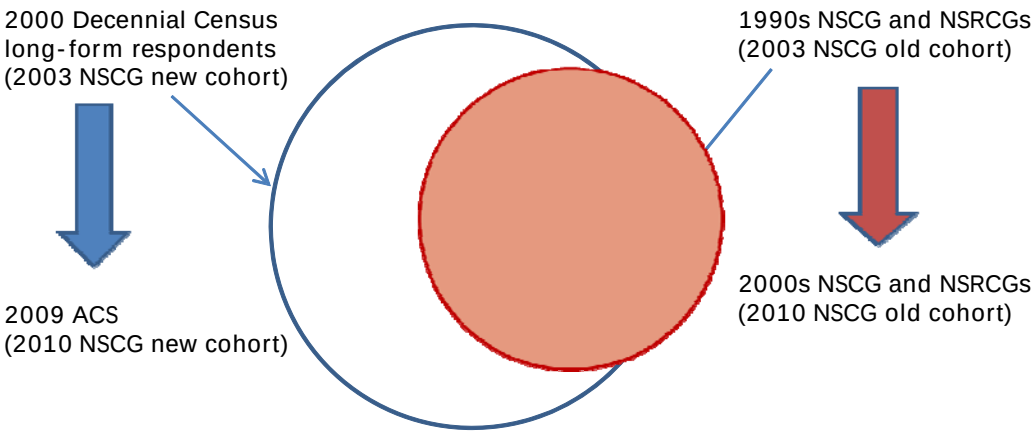
2. USING 2003 NSCG DATA TO PREPARE FOR 2010 NSCG ESTIMATION

One of the most challenging tasks of the 2010 NSCG estimation is identifying each sample unit's eligibility for either sampling frame so that the population can be divided into three components, as described in Section 1.2. NSCG surveys have numerous survey eligibility conditions, so it may be difficult to classify each sample unit's overlap status without misclassification error. In particular, it is the case with one longitudinal sample and the other cross-sectional sample in the dual samples like the 2010 NSCG. The purpose of this paper is to examine the 2003 NSCG sample, which was also dual frame, and to explore any issues that we may encounter in the 2010 NSCG.

2.1. 2003 NSCG Sample

As shown in Figure 2, the 2003 NSCG also used a dual-frame sample design, with a longitudinal sample derived from the baseline survey of the 1990's Decennial Census plus supplemental NSRCGs during the 1990s and a new sample derived from the 2000 Decennial Census. Note that the corresponding new cohort for 2010 was selected from the 2009 ACS annual file.

Figure 2: The 2003 NSCG Sample



As shown in Table 1, the population components for the 2003 NSCG can be defined based on degree and occupation. The components listed in the top three rows (highlighted in red) are covered by both the old and new cohorts. The other parts are mostly covered by the new cohort. Please note that for sample members with a Ph.D., the old cohort only covers foreign degrees earned prior to the 1990 Census, which are also covered by the NSCG new cohort. Also note that in the 1990's, as in the 2000's, the NSCG sample was supplemented by recent college graduates from U.S. institutions.

Table 1. Target Population for the 2003 NSCG

Population Component	New	Old
S&E bachelor's or master's degrees as of 4/1/1990	X	X
Foreign-earned doctoral degrees in S&E fields as of 4/1/1990	X	X
U.S.-earned bachelor's or master's degrees in S&E fields between 4/1/1990 and 4/1/2000	X	X
U.S.-earned bachelor's and master's degrees in S&E fields between 4/1/2000 and 6/30/2000	--	X
Bachelor's or higher degrees in non-S&E fields as of 4/1/2000 but working in S&E or S&E-related fields as of 10/1/2003	X	--
Bachelor's or higher degrees in S&E-related fields as of 4/1/2000	X	--
U.S.-earned doctoral degrees in S&E fields as of 4/1/2000	X	--
Foreign-earned bachelor's or higher degrees in S&E fields between 4/1/1990 and 4/1/2000	X	--

It is relatively easy to identify the NSCG overlap component using degree and occupation data because comprehensive degree information was collected from both the old and new cohorts. This information was mostly captured in the respective baseline surveys for each cohort; the data from the old cohort were updated over the course of the decade. We may encounter discrepancies in degree field coding in the 1990's and 2003, which could lead to misclassification of the overlapped component, but the magnitude of the error is expected to be small.

We can therefore classify all 2003 NSCG sample cases into one of the three component groups based on overlap status. The population total can then be expressed as $Y = Y_{N \cap O^C} + Y_{N \cap O} + Y_{N^C \cap O}$, and the total for each component can be estimated based on one or two samples from the corresponding component. As shown in Table 2, the total for the overlapping component can be estimated based on either or both of two samples. But notice that the difference between the two estimates is substantial, more than 1.5 million. That is, the new sample-based estimate for the total overlap component is 12.2 percent larger than the old estimate.

Table 2. Total Count Estimates for the 2003 NSCG

	$N \cap O^C$	$N \cap O$	$N^C \cap O$	**Total**
New cohort	9,254,448	13,703,840	--	22,958,288
Old cohort	--	12,034,395	286,895	12,321,290

2.2. Other Conditions in Identifying the Overlapping Component

The large discrepancy between the new and old estimates clearly indicates that the identification of sample units in the overlapping component should be thoroughly reviewed. Besides degree and occupation criteria, other criteria should be considered for

survey eligibility, such as age and living situation. For example, to be eligible, a person must be 75 or younger and living in a noninstitutional setting on the survey reference date. These criteria would produce two kinds of survey-ineligible cases: permanently ineligible and temporarily ineligible. Permanently ineligible people would be those who were 76 or older on the survey reference date; who had severe or terminal illness requiring hospital, hospice, or other long-term care with little or no prognosis for recovery; or who were permanently incarcerated. Once identified, these individuals would be classified as ineligible for all current and future surveys. Because age information is collected from both old and new cohorts with very little missing data, the estimation error due to missing age information, if any, would be negligible. The size of the group that is permanently ineligible for non-age-related reasons is expected to be small.

On the other hand, there are certain conditions that may lead people to temporarily drop out of the workforce or to otherwise leave the scope of the survey. If an individual was temporarily incarcerated or was institutionalized due to a physical or mental impairment with a prognosis for recovery, he or she would be treated as temporarily ineligible. These individuals would shift in and out of the scope of the survey depending on their "institutionalized" status during each survey round: 1993, 1995, 1997, 1999, 2001, and 2003. However, we cannot determine whether the new cohort was institutionalized on one or more survey reference dates during the 1990's because the survey did not capture this information directly from respondents. This issue may have caused misclassification errors, which may in turn have led to the difference in estimates between the two samples. However, we would expect the population size belonging to this case would be smaller than the next case presented below.

2.3. Identification of U.S. Residents on Survey Reference Dates

One of the most significant factors in misclassifying the overlapping component was residence status. Eligible sample members must have lived in the U.S. on the survey reference date. Therefore, to identify new cohort members who would have been eligible for the old cohort, we need to know their U.S. residence status during the 1990's on several reference dates: 4/1/1990, 4/15/1993, 4/15/1995, 4/15/1997, 4/15/1999, and 4/15/2001. However, U.S. residency for the new cohort is only known for 4/1/2000 and 10/1/2003. To gauge the magnitude of potential misclassification due to this lack of data, we first present a table showing the reference dates requiring U.S. residency data for those who meet certain degree requirements (Table 3).

Table 3. Residency Status Needed for the Overlapping Component, by Originating Survey

Survey	4/1/90	4/15/93	4/15/95	4/15/97	4/15/99	4/1/00	4/15/01
1990 Decennial Census	X	X				X	
1993 NSRCG		X		X		X	
1995 NSRCG			X			X	
1997 NSRCG				X		X	
1999 NSRCG					X	X	
2000 Decennial Census						X	
2001 NSRCG						X	X

Based on their degree eligibility, old and new cohort members can be matched to a survey listed in the first column of Table 3. Once paired with a survey, each member must meet the U.S. residence requirements on the dates checked. For example, if a new sample member meets the degree criteria for 1990 Decennial Census, he or she must have lived in the U.S. on at least two of the checked reference dates (4/15/1990, 4/15/1993, and/or 4/1/2000) because only eligible respondents back then were included in the sampling frame for follow-up surveys.

The real question, therefore, is how to determine the U.S. residence status of sample members for whom this information is not available. This issue is particularly critical because of the steady flow of working people both into and out of the U.S. To identify sample members in the old cohort that belong in the overlapping component, we need to determine their whereabouts on the 2000 Decennial Census date. We can assume that most of the sampled cases in the old cohort would have lived in the U.S. on the 2000 Decennial Census date because they were verified as living in the U.S. at some point during the 1990s and on 10/1/2003, the 2003 survey reference date. More important, most of sampled cases (except for temporarily ineligible cases in the 1999 survey) actually lived in the U.S. on the 1999 survey reference date (4/1/1999). Though it is possible that some people might have left the country after the 1999 survey and come back to the U.S. after the Decennial Census date, that group is likely to be small.

To identify the overlap status of the new cohort, we must first match each sample member to a 1990's survey based on their degree information. As shown in Table 3, each survey component requires data on U.S. residence status.

Some information we found to be useful for predicting a sample person's U.S. residence is a combination of U.S. citizen at birth (USCAB) status and U.S. entry year data, which was collected during the 2003 survey for all respondents except USCABs. Table 4 shows the weighted counts for both the old and new cohorts by survey component, as determined by eligible degrees. As highlighted in red, the weighted counts for those with degrees eligible for the 1993 NSCG but who came to the U.S. after the 1990 Decennial Census date are markedly different between the two samples. More than 300,000 people in the new cohort reported that they had come to the U.S. after 1990, while virtually none from the old cohort did. Although it is possible that a few people visited the U.S. on the Census date, left the country, and then returned to stay, it would be rare. Consequently, only a small fraction of the new cohort cases in this cell should be in the overlapping component.

Table 4. 2003 NSCG Count Estimates, by Survey Component and U.S. Entry Year

	Old Cohort					New Cohort				
	USCAB	**Before**	**During**	**After**	**Total**	**USCAB**	**Before**	**During**	**After**	**Total**
2001 NSRCG	596,066	104,691	0	192	700,949	514,998	99,457	357	312	615,124
1999 NSRCG	801,141	120,655	0	695	922,491	845,896	132,432	184	237	978,749
1997 NSRCG	711,825	101,964	849	1,904	816,542	817,937	123,940	173	734	942,784
1995 NSRCG	640,242	80,165	0	972	721,379	777,661	101,894	0	2,597	882,152
1993 NSRCG	684,895	76,018	0	1,229	762,142	958,002	113,285	278	3,072	1,074,637
1993 NSCG*	7,283,726	816,644	6,182	4,340	8,110,892	7,874,343	966,829	38,948	331,274	9,211,394
Overall	10,717,895	1,300,137	7,031	9,332	12,034,395	11,788,837	1,537,837	39,940	338,226	13,704,840

USCAB = U.S. citizen at birth; before (during, after) = naturalized U.S. citizen or non-U.S. citizen who entered the U.S. before (during, after) the survey reference year.

*The survey reference year for 1993 NSCG is 1990, a Decennial Census year.

However, due to lack of data, we cannot determine the U.S. residence status for most of the new cohort, even approximately. For example, if an immigrant entered the country before the 1990 Census date, or if a U.S. citizen with an eligible degree left the country before the 1990's Census date but returned before the 2000 Census date, he or she should be excluded from the overlapping component. But given the information available, there is no way to identify these individuals. To get a rough idea of the number of temporary emigrants, we calculated the number of cases that were temporarily ineligible due to being out of the U.S. in the 1990's surveys (Table 5).

Table 5. Weighted Counts of Emigrants Excluded from Follow-Up Surveys in the 1990's

Survey Year	1993 NSCG	1993 NSRCG	1995 NSRCG	1997 NSRCG	1999 NSRCG	2001 NSRCG	Total
1993	217,000	38,000					255,000
1995	0	0	34,000				34,000
1997	0	21,000	0	30,000			51,000
1999	0	0	0	0	30,000		30,000
2001						30,000	30,000
Total	**217,000**	**59,000**	**34,000**	**30,000**	**30,000**	**30,000**	**400,000**

In the 1993 baseline survey, a little over 200,000 people were deemed to be outside the U.S. and were thus not included in the follow-up survey. Cumulatively over the course of the 1990's, about 400,000 people are estimated to have been excluded from the NSCG

old cohort for being outside the U.S. However, we cannot determine this figure for the new cohort. That is, up to 400,000 emigrant people might have been eligible for the new cohort.

The question is then how many of these individuals actually returned to the country. The 2003 NSCG included all respondents, as well as temporarily ineligible cases, from the 1999 survey. After four years since the 1999 survey, about 70 percent of the people who had left the U.S. turned out to live in the U.S on the 2003 NSCG survey reference date. This may give us roughly 300,000 (70 percent of 400,000) overestimated by the new cohort due to the lack of sufficient information to detect them. However, adding these and the other immigrants discussed above still leaves us with a gap of about one million between the two cohort samples.

3. SUMMARY AND DISCUSSION

Dual-frame estimation for the NSCG involves a few challenging issues, mainly due to the prominent difference between the two samples: the old cohort is a longitudinal sample, and the new cohort is a sample from a frame with full coverage. One specific challenge is identifying the overlapping population for the NSCG, especially for the new cohort, due to a lack of information on the U.S. residency status of these individuals during the past decade. One possible solution is to include people who leave the U.S. from one survey round to the next in order to maintain full coverage of the population, regardless of their mobility. However, it would be excessively expensive to follow them for multiple rounds of the survey. Instead, we recommend adding a few questions to the survey on each sample person's U.S. residence status during past survey rounds.

Among other challenges not discussed here are attrition and nonresponse bias, which may contribute to potential underestimation of the old cohort.

4. RECOMMENDATIONS FOR THE 2010 DUAL-FRAME ESTIMATION

Identification of the overlapping component. Even with insufficient information to identify sample cases in the overlapping population, we suggest that the NSCG sample be partitioned into overlapping and nonoverlapping parts based on degree. High-quality, compatible degree information should be available for both cohorts. After that, we suggest using the U.S. entry year for non-USCABs to determine U.S. residence status on survey reference dates. Further investigation would still be needed to identify the best predictors for temporary emigrants other than U.S. entry year.

Dual-frame estimation. The new cohort sample for the 2010 NSCG was selected from the 2009 ACS and is thus fully representative of the entire NSCG population. Consequently, we recommend applying a dual-frame estimation (or weighting strategy) to combine the old and new samples, followed by raking adjustments to the combined weight so that the weighted totals conform to the new sample-based totals. The 2003 empirical results showed that the combined sample estimates were compatible with the new sample-based estimates with larger effective sample sizes.

REFERENCES

Brick, J., S. Dipko, S. Presser, C. Tucker, and Y. Yuan. "Nonresponse Bias in a Dual-Frame Survey of Cell and Landline Numbers." *Public Opinion Quarterly,* vol. 70, 2006, pp. 780–793.

Finamore, J., D. Hall, and J. Walker, "NSCG Estimation Issues When Using an ACS-Based Sampling Frame," Presented at the American Statistical Association Joint Statistical Meetings, Miami Beach, Florida, July 30-August 4, 2011.

Fuller, W.A., and L.F. Burmeister. "Estimators for Samples Selected from Two Overlapping Frames." In *Proceedings of the Social Statistical Section,* American Statistical Association, 1972.

Hartley, H.O. "Multiple-Frame Surveys." In *Proceedings of the Social Statistical Section,* American Statistical Association, 1962.

Hartley, H.O. "Multiple-Frame Methodology and Selected Applications." *Sankhya*, vol. 36, 1974, pp. 99–118.

Kalton, G., and D. Anderson. "Sampling Rare Populations." *Journal of the Royal Statistical Society*, vol. 149, 1986, pp. 65–82.

Lohr, S. "Recent Developments in Multiple-Frame Surveys." In *Proceedings of the Social Statistical Section,* American Statistical Association, 2007, pp. 3257–3264.

Mecatti, F. "A Single-Frame Multiplicity Estimator for Multiple-Frame Surveys." *Survey Methodology,* vol. 33, no. 2, 2007, pp. 151–157.

National Research Council. *Surveying the Nation's Scientists and Engineers: A Data System for the 1990s.* Washington, DC: National Academy Press, 1989.

Skinner, C.J., and J.N.K. Rao. "Estimation in Dual-Frame Surveys with Complex Designs." *Journal of the American Statistical Association,* vol. 91, 1996, pp. 349–356.