The 2010 National Survey of College Graduates Weighting Research: Leveraging the 2009 American Community Survey [1]

David Hall, Benjamin M. Reist, Julie Walker, and Michael White
U.S. Census Bureau, U.S. Department of Commerce, Washington, DC 20233

Abstract: The 2010 National Survey of College Graduates (NSCG) used 2009 American Community Survey (ACS) respondents as its sampling frame. This sample design afforded the NSCG the opportunity to use the rich set of data available from the ACS to conduct weighting research. In particular, research on response propensity based locating and nonresponse adjustments. We researched response propensity models using ACS data collection paradata, as well as ACS survey data, to aid predicting NSCG response patterns. The ACS's rich set of information enhanced NSCG's ability to improve our weighting methodologies and therefore produce higher quality survey estimates.

Key words and phrases: Nonresponse Adjustments, Propensity Modeling

1. Introduction

The 2010 National Survey of College Graduates (NSCG) used the 2009 American Community Survey (ACS) as the first phase of a two-phase sample design. This provided the NSCG with person-level and household-level frame data based on the ACS. Additionally, paradata from the ACS data collection was available, mainly mode of collection. These data were leveraged in locating and eligibility/nonresponse weighting adjustments.

Historically, the NSCG has used sampling strata as nonresponse weighting cells and collapsed cells when there were not a sufficient number of cases in a cell. This collapsing was done in an ad hoc manner. The goals for using propensity based nonresponse adjustments was to eliminate the need for ad hoc cell collapsing used in prior rounds and to incorporate paradata from the ACS's data collection into the nonresponse adjustments.

ACS data were used to model propensities and assess how well the weighting adjustments mitigated bias. This was done using full sample and ACS estimates of totals that were highly correlated with key estimates produced by the NSCG.

[1] Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

The rest of the paper is organized as follows. Section 2 gives background on the NSCG. Section 3 discusses 2010 NSCG nonresponse adjustment research. Finally, section 4 offers some discussion.

2. Background

The NSCG is a longitudinal survey conducted every two to three years by the U.S. Census Bureau on behalf of the survey's sponsor, the National Science Foundation (NSF). The NSCG is the largest of three surveys that are combined to form the Scientists and Engineers Statistical Data System (SESTAT). The other two surveys are the National Survey of Recent College Graduates (NSRCG) and the Survey of Doctorate Recipients (SDR). SESTAT is a comprehensive and integrated system of information about the employment, educational, and demographic characteristics of the science and engineering population in the United States. The integrated data from these three surveys serve as the basis for the development of national estimates on the science and engineering (S&E) workforce.

Previously, the NSCG selected its sample from decennial census long form respondents. The long form was a large frame from which to select a sample (approximately 1 in 6 households in the United States, including Puerto Rico and the outlying areas[2]) and provided a wealth of information for sampling purposes. The SESTAT target population is defined to be people with a bachelor's degree or higher, educated or working in an S&E or S&E-related field or occupation, who are age 75 or younger, noninstitutionalized[3], and living in the United States as of the survey reference date. The long form asked for the highest level of education a person completed, so we could identify the college graduates in the population. However, we could not distinguish those with an S&E or S&E-related degree that we were targeting. We could only identify people working in an S&E or S&E-related occupation. Thus, we needed to field a large screener survey in order to capture those with an S&E or S&E-related degree who were working in a non-S&E occupation or not working at all. In addition, the decennial census long form was only available once per decade, so later rounds of the NSCG during the decade were supplemented with respondents from the NSRCG. This method captured new graduates in the population, but did not catch new immigrants coming into the country who earned their degree before entering the U.S or people who entered an S&E occupation.

The 2003 NSCG initially selected a sample of 177,320 persons from the 2000 decennial census long form frame, of which approximately 100,000 responded.

[2] The outlying areas include American Samoa, Guam, Northern Mariana Islands, and the U.S. Virgin Islands.

[3] Institutionalized people include those residing in prisons, correctional facilities, group homes and residential treatment centers for juveniles, nursing homes, hospitals, hospice facilities, and residential schools for people with disabilities.

Of these respondents, only 66,247 had either an S&E or S&E-related degree or occupation. The remaining 111,073 cases were either nonrespondents or otherwise ineligible for the SESTAT target population and were not followed-up with in the 2006 NSCG. Thus, most of this initial sample was screened out of future rounds of the survey.

In 2010, the Census Bureau discontinued the use of the long form. Everybody received what had been known as the short form as part of the 2010 Census. Unfortunately for the NSCG, the short form does not collect information on educational attainment, occupation, or other variables we need to identify persons who are eligible for the NSCG sample. Continuing to supplement the existing NSCG sample with new NSRCG respondents only is not desirable due to the immigrant coverage issue mentioned above. In addition, sample attrition over time can introduce bias to the estimates. The only practical alternative frame identified from which to select a new 2010 NSCG sample is the 2009 ACS.

3. Nonresponse Adjustment

We began our nonresponse research by comparing potential models using data from the 2003 NSCG. The 2003 NSCG selected its sample from respondents to the 2000 decennial census long form. Thus, we had similar frame data as would be available for the 2010 NSCG, which was sampled from respondents to the 2009 ACS.

The 2003 NSCG used combinations of sampling stratification variables to form nonresponse adjustment cells. The variables were demographic group, occupation group, gender, highest degree type, age group, and outlying area (U.S. territory) indicator. Within each cell, weight from the nonrespondents was transferred to the respondents using a ratio adjustment. This weight adjustment is based on the assumption that the responses for the nonrespondents (had they been obtained) would be similar to those of the respondents within the same cell. Small nonresponse adjustment cells and cells with no respondents were collapsed with other cells using complicated criteria to maintain the differential weighting intentionally included in the survey design. While this collapsing reduced the variance of the estimates, it was quite subjective and time-consuming.

We looked at several alternatives to the traditional nonresponse adjustment described above based on propensity models. For each propensity model, a person's response status (respondent or nonrespondent) was modeled against several independent variables in a logistic regression to estimate their propensity to respond. We considered both weighted and unweighted models. The weighted models took into account the underlying sample design information such as clustering, stratification, and sampling weights. We used the individual propensities output from the models directly as weighting adjustment factors, as well as indirectly to form adjustment cells (using propensity score deciles). When using adjustment cells, the weight from the nonrespondents was transferred to the respondents using a ratio adjustment as was done with the traditional approach. For the adjustment factors, we considered the inverse of both the weighted response rate and the unweighted response rate within each cell. We ran one set of models using the six sampling stratification variables that were used to define the traditional nonresponse adjustment cells, and we ran an identical set of models using the same six variables plus three additional variables derived from decennial census paradata to mimic what paradata would be available for the 2010 NSCG. These three additional variables were census completion mode (mail or enumerator), census reference person (self or proxy reported), and residence type (household or group quarter).

When we considered all combinations, we ran twelve propensity models. They consisted of the following components: two types of models (weighted and unweighted); three types of adjustment factors (inverse of the individual propensity score, inverse of the unweighted response rate with adjustment cell defined by propensity deciles, and inverse of the weighted response rate within adjustment cell defined by propensity deciles); and two types of independent variables (with census paradata and without).

The logistic regression models were run in SAS using proc logistic for unweighted models and proc surveylogistic for weighted models. With proc logistic, we used a stepwise selection procedure to identify important variables in the models (the version of SAS we used did not have a stepwise selection option for proc surveylogistic).

We evaluated the twelve models by computing weighted estimates of numerous variables from the decennial census frame and comparing them to estimates produced using the traditional method as well as estimates derived from the 2000 census long form and NSCG base weights. Some of these evaluation variables were related to the sampling stratification variables used to define the nonresponse adjustment (e.g., race, citizenship status, etc.), while others were not related (e.g., marital status, school enrollment, etc.). We evaluated the models by examining the bias, variance, and mean squared error. The research demonstrated the value of using propensity models for making nonresponse adjustments in that we were able to achieve similar results to the traditional method without having to worry about collapsing small adjustment cells.

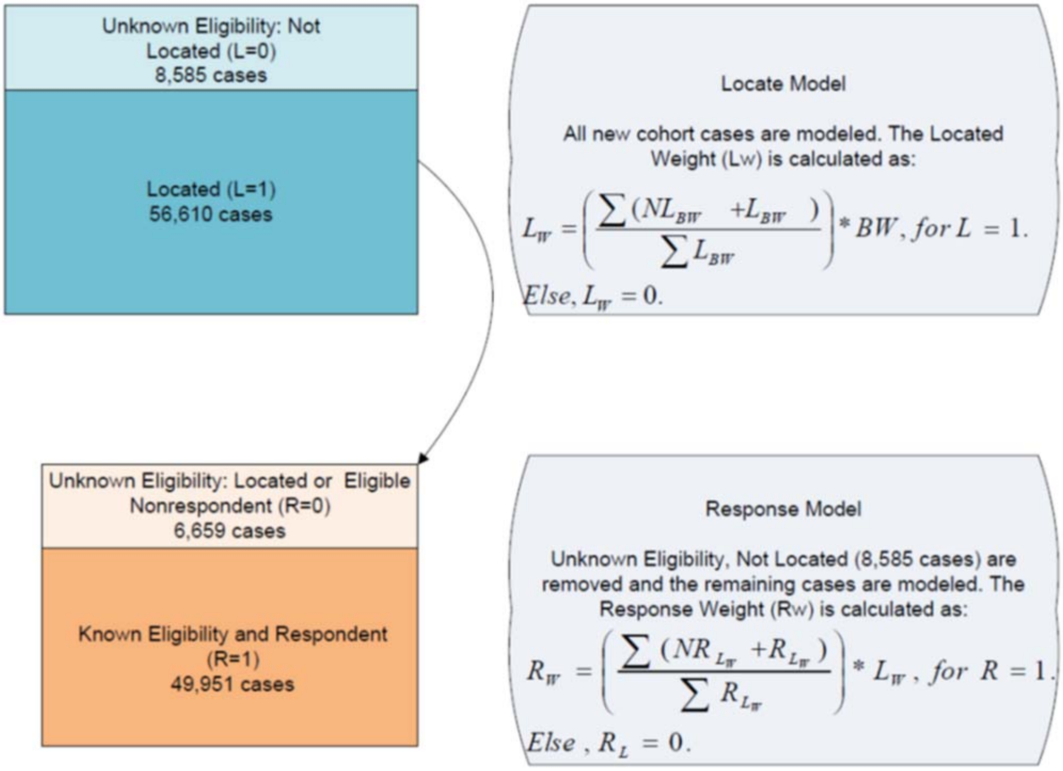
Of the twelve propensity models considered, we selected the one that used an unweighted propensity model with decennial census paradata using the inverse of the weighted response rate as the adjustment factor within adjustment cells defined by propensity deciles as the most promising for our purposes. We chose this model for the following reasons: (a) the estimates had relatively small bias, (b) it did not inflate the weights as much as some of the other methods, (c) the stepwise function in proc logistic allowed for a more parsimonious model, (d) we felt using the individual propensity scores was too dependent on the model fit, and

(e) we believe the paradata (especially response mode) will be useful in explaining people's propensity to respond to the NSCG.

Since the model we selected to use was unweighted, it did not reflect the sample design. Therefore, we tried one additional model in which we added the NSCG base weight as a potential independent variable in the model. However, the stepwise selection procedure in proc logistic did not identify the base weight as important in the model, so that gave us more confidence in choosing an unweighted model.

Once we were able to develop a promising model using 2003 NSCG and 2000 decennial long form data, we turned our attention to developing models for the 2010 NSCG leveraging 2009 ACS response data and paradata. Since the 2010 NSCG used different sampling methods and vastly different sampling rates for S&E cases and non-S&E cases (as determined by their degree field and occupation as reported in the 2009 ACS), we considered performing the nonresponse adjustment separately for the two groups. In addition, we also considered breaking our nonresponse adjustment up into two components, since a person may be a nonrespondent for more than one reason. We first made an adjustment for unlocated cases (who may have responded if they could have been located), and then we made an adjustment for nonresponse among the cases that had been located (refusal, death, etc. – see Figure 1).

FIGURE 1. PROPENSITY MODEL FOR LOCATED CASES, KNOWN ELIGIBILITY AND RESPONSENT CASES


Unknown Eligibility: Not Located (L=0) 8,585 cases
Located (L=1) 56,610 cases
Unknown Eligibility: Located or Eligible Nonrespondent (R=0) 6,659 cases
Known Eligibility and Respondent (R=1) 49,951 cases
Locate Model
Response Model

All new cohort cases are modeled. The Located Weight (Lw) is calculated as:

$$L_W = \left(\frac{\sum (NL_{BW} \quad + L_{BW} \quad)}{\sum L_{BW}} \right) * BW, for\, L = 1.$$

$$Else, L_W = 0.$$

Unknown Eligibility, Not Located (8,585 cases) are removed and the remaining cases are modeled. The Response Weight (Rw) is calculated as:

$$R_W = \left(\frac{\sum (NR_{L_W} + R_{L_W})}{\sum R_{L_W}} \right) * L_W\, ,\ for\ R = 1.$$

$$Else\ , R_L = 0.$$

Different ACS frame variables were used in each of these models since there was evidence that the mechanisms for locating a case and for a case responding were different.

Finally, since we selected a method that utilized adjustment cells, we also considered how many cells we should create (in our initial 2003 NSCG research, we divided the propensity scores into deciles exclusively). We created 5, 10, 50, 100, 150, 200, 250, and 300 approximately equal-sized cells by grouping similar propensity scores (when creating cells for non-S&E cases separately, we only created 5, 10, and 50 cells due to the small sample size).

We identified several ACS variables and paradata that might be potentially useful in the propensity models, along with their two-way interactions: gender, demographic group, age group, educational attainment, S&E degree status, occupation group, personal income, tenure, mobility status (moved in previous 12 months), urban/rural status, census division (geographic area), number of persons in the household, relationship to householder (to possibly identify proxy responses), ACS tabulation date (interview month), and ACS data collection mode (mail, telephone, personal interview, group quarters personal visit). See Table 1 for which variables and interaction terms were significant in one particular set of nonresponse adjustment models.

TABLE 1. VARIABLES USED IN S&E LOCATING AND UNKNOWN ELIGIBILITY PROPENSITY MODELS

Variable	**Locating**	**Eligibility**
ACS Mode	X	X
Age	X	X
Census Division	X	X
Demographic Group	X	X
Gender	X	X
Highest Degree	X	X
Mobility Status	X	
Number of People in Household	X	
Occupation	X	X
Relationship to Householder	X	X
Science & Engineering Degree	X	X
Tenure	X	X
ACS Mode*Census Division	X	
ACS Mode*Relationship to Householder		X
Age*Highest Degree	X	X
Census Division*Occupation		X
Gender*Tenure		X

As with our initial 2003 NSCG propensity model research, we evaluated the various models by comparing several weighted estimates of ACS frame variables to estimates derived from the 2010 NSCG base weights and ACS weights. These included:

- Race / Ethnicity
- U.S. Citizenship at Birth
- Age Group
- Demographic Group
- S&E Degree
- Census Division
- Citizenship
- Detailed Occupation Group
- Gender
- Educational Attainment
- Grandchildren in House
- Health Insurance Coverage
- Minutes Spent Commuting to Work
- Fertility

Some of these evaluation variables were related to our sample design (e.g., age group, census division, etc.), while others were not (e.g., presence of grandchildren in the household, health insurance coverage, minutes spent commuting to work, and fertility).

For the 2010 NSCG, we decided to perform the nonresponse weighting adjustment separately for S&E cases and non-S&E cases. We further decided to do the nonresponse adjustment sequentially for unlocated cases and then nonrespondent cases within S&E status. To form the adjustment cells, we decided to use 300 cells for S&E cases and 50 cells for non-S&E cases by grouping similar propensity scores into approximately equal-sized cells. This method offered the best combination of low bias, low design effect, and lowest maximum weight of the various propensity models considered.

4 Discussion

The previous nonresponse adjustment technique used by the NSCG used sampling strata as nonresponse adjustment cells. Weighting cells were collapsed when there were an insufficient number of cases in a cell. Within cell, the weight from the nonrespondents was transferred to the respondents using a ratio adjustment. This method is somewhat ad hoc and makes the assumption that nonrespondents and respondents are inherently similar.

This paper serves as an overview of the 2010 NSCG research that uses the rich set of covariates available from the 2009 ACS to improve nonresponse adjustments and ultimately improve survey estimates. We developed propensity models using the rich set of ACS frame data to inform nonresponse adjustments. These models were evaluated by examining bias, design effect and weights. When final replicate weights, we will be able to further evaluate the models using estimates of variance and mean square error. For future research, we plan to leverage available ACS data to improve other methodology including weight trimming, raking, editing, imputation, and calibration.