Effects of Misclassification of Race/Ethnicity Categories in Sampling Stratification on Survey Estimates

Donsig Jang[1], Amang Sukasih[1], Xiaojing Lin[1],
Kelly H. Kang[2], Stephen H. Cohen[2]
[1] Mathematica Policy Research, Inc., 600 Maryland Avenue, SW, Suite 550, Washington, DC 20024
[2] National Science Foundation, Division of Science Resources Statistics, 4201 Wilson Boulevard, Room 965S, Arlington, VA 22230

Abstract
Misclassification of race/ethnicity occurs when there is a discrepancy of classifications based on two different sources, e.g., administrative data and reported values. As a result of this error, a sample designed to meet analytic objectives could, when implemented, result in the loss of effective sample sizes in key domains involving the race/ethnicity group. In assessing misclassification error in the race/ethnicity category, the true values and the misclassified values are established. In practice, the true value is often unknown and can only be assumed. In our study, we assessed whether the misclassification of race/ethnicity occurred during the sampling frame construction, assuming that the data obtained from the respondents is more accurate than the frame and will serve as the true values. This assumption is aligned with survey practice where estimates for race/ethnicity are often derived based on reported values rather than the frame values. We estimate the misclassification matrix in which misclassification parameter/proportion can be calculated with the usual weighted survey estimate. We also investigated the impact of misclassification on survey estimates (weighted totals), where these estimates were produced using the weights that had been raked into three different marginal population totals

Key Words: effective sample size, NSRCG, raking

1. Introduction

1.1. Sources of Survey Errors
Data collection through a sample survey involves several steps—preparation, execution, and dissemination. These include complex processes such as sample frame construction, sample selection, data collection and processing, and estimation. Each process is subject to error. *Total survey error* encompasses all errors occurring in the survey—from frame construction to estimation. Figure 1 summarizes these errors. Suppose that the goal of a data collection is to estimate parameter θ_P from a target population $\mathcal{P}$. A sample is selected from a sampling frame $\mathcal{F}$ that contains all units in the target population. It is desirable to have a good sampling frame so that θ_F is essentially the same as θ_P, where

index *F* indicates the frame and *P* indicates the population. However, the sampling frame is often not perfect in the sense that it fails to cover some portion of the population or lacks information for correct classification of each unit's eligibility status or stratum membership. Errors due to an imperfect sampling frame can be classified, in general, as a coverage and/or misclassification error.

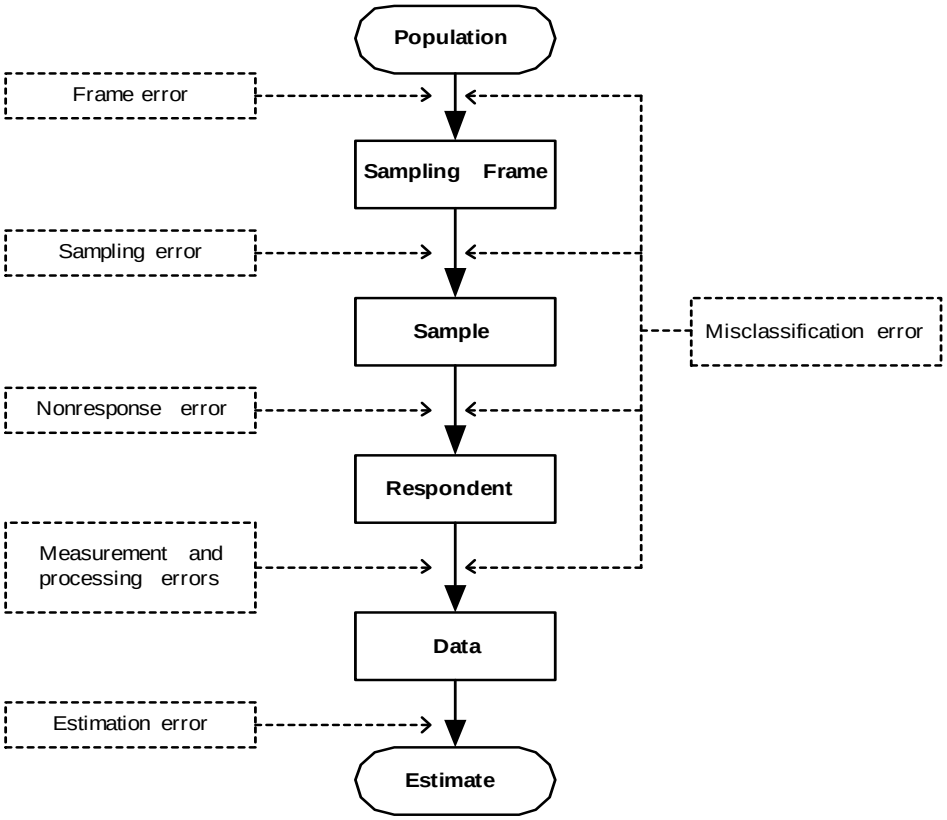

Figure 1: Total survey errors

A survey estimate is derived from the units in a sample rather than the entire population. Thus, the estimate is subject to *sampling error*; the sampling error accounts for uncertainty of the sample selection process due to the fact that the sample chosen is one of many possible realizations. Sampling error can often be controlled during the sample design stage and can be correctly quantified at the estimation stage.

When an appropriate sample design is used, the resultant sample is expected to be representative of the population from which the sample is drawn. However, during data collection, sampled units may not respond to the survey or may respond only partially, which may introduce a *nonresponse error*. There have been tremendous research efforts made by survey methodologists and statisticians to deal with nonresponse issues, especially during the past three decades (Groves et al. 2002; Little and Rubin 2002).

The other source of survey errors is *measurement error*, which occurs when respondents may provide "inaccurate" answers. These errors may be caused by the interviewer, respondent, survey questionnaire or the mode of data collection used. These errors are equally important as nonresponse errors but have received less attention partly because of the more complex nature of the problem.

Ultimately, in survey practice we want to assess the error associated with the estimator obtained from the data, $\hat{\theta}_P$ compared to the true value θ_P; that is $\hat{\theta}_P - \theta_P$. Thus, it is necessary to understand the entire survey process to estimate total survey error.

1.2. Quantifying Sampling and Nonsampling Errors

Sampling error is often quantified through the variance (or standard error) of the estimator for a population parameter. The variance measures the uncertainty of the estimator because it is derived from a specific set of sample data. Survey statisticians have put tremendous research efforts into variance estimation methods (for example, see Binder 1983; Krewski and Rao 1981; Rao 2003; Shao 1996; Wolter 1985).

Nonsampling error is rather difficult to measure because it has many components that often require additional measurement outside the scope of the survey. For the past two decades, however, a greater appreciation of the effect of nonsampling errors on survey estimates has dominated the discussion of total survey error (Lessler and Kalsbeek 1992; Lyberg et al. 1997; Groves et al. 2004). Individual components of nonsampling error have been studied and best practices on questionnaire development, mode of data collection, measurement techniques and other aspects of data collection have been developed with a goal to minimize nonsampling error in practice.

Our research deals with one aspect of nonsampling error caused by frame error—the *misclassification error,* which is a type of measurement that exists in categorical data when there is a discrepancy between classification of the observed and the true values. It can exist in any step of data collection (see Figure 1) including during the coding process (either during frame construction or survey data processing) or during respondent interviews, when error is caused by misinterpretation of question items or vague definitions of survey variables. Our focus here is only on the misclassification error that can occur during the sampling frame construction process, specifically when variables used for sampling stratification contain imperfect information. This would occur, for example, when race/ethnicity is used to stratify the population and, during the frame construction, some Hispanic males are classified as white males.

Given the true value, misclassification error can be measured. In practice, the true value is often unknown and the value needs to be assumed. An example of this can occur when a sample survey is used to check the quality of a census data collection. Since resources used in the sample survey are often better than those used in the census, data collected from the sample survey is then used as the "true" value to assess misclassification errors. In our survey example, sampling variables used for stratifying the population were available from administrative data during the frame construction. Later on, these variables were also collected from the sampled respondents during the survey interview. To assess misclassification effect in the frame construction, we assume that the data obtained from the respondents is more accurate and treat it as the true values. This assumption is actually aligned with survey practice in that survey estimates are often made based on reported rather than frame values.

1.3. Misclassification Error in Stratification

Sample frame construction is an important but often overlooked function of the survey process. Of errors caused by the sampling frame construction, coverage error is of primary importance. An attempt is usually made to minimize coverage error both at the sampling frame construction and estimation stages. For example, to compensate for the

coverage error at the estimation stage, post-stratification or raking adjustments can be made by conforming the total number of ultimate units, estimated from the sample, to a known population total for some key characteristics from a reliable source, if available.

However, even with a sampling frame with complete coverage, there might be another source of survey error called classification error. This occurs when classification of frame units, with respect to sampling variables such as primary sampling unit and stratum, is made based on erroneous information available on the sampling frame. In general, stratification is often considered to achieve two design objectives: control of sampling variability within homogeneous sampling strata and allocation of sample sizes to analytic domains sufficient enough to meet domain-specific precision requirements. A sample can be efficient to the extent that, based on the information included in the frame, there is very little error in classifying the sampling units into strata; however, because the frame information is often incomplete or inaccurate, stratification variables are often misclassified at the design stage. One possible consequence of this error is that a sample designed as an equal probability of the selection sample in a misclassified sampling stratum would, when implemented, result in a sample with unequal weights within the corresponding analytic domain that is constructed based on the true classification.

2. Methodology

2.1. Customary Misclassification Parameter

Kuha and Skinner (1997) provide a discussion on the effects of misclassification and methods to measure these effects under the survey framework. In particular, they present methods of measuring misclassification on categorical variables in a finite population. Suppose that variable A has m categories; let A be defined as the variable with true values, while A^* is defined as the corresponding variable with misclassified values. Let the parameter θ_{jk} represent a misclassification of a true category k as category j. This is a random process in which the probability of misclassifying a respondent as belonging to category j when in fact his/her true category is k is

$$\theta_{jk} = \text{Prob}(A^* = j \mid A = k), \quad j,k = 1,\cdots,m. \tag{1}$$

That is, given a value of A, say $A = k$ (being treated as fixed), this value may be misclassified into j (recorded in variable A^*) with probability θ_{jk}. The summation of such probability across all values of j will be one for each k; that is, $\sum_{j=1}^{m} \theta_{jk} = 1$. The misclassification matrix Θ is then defined as an $m \times m$ matrix with (θ_{jk} ; j, k = 1, … m), where each column must sum to one. In a finite population context, θ_{jk} may be interpreted as the proportion of finite population units in true category k classified as category j. When there is no misclassification, Θ is an identity matrix. The larger the misclassification error, the smaller the proportion in the diagonal elements of Θ. To construct this matrix, an analyst needs to have true values and misclassified values for each observation. However, we often do not have the true values for each individual in the population, having instead two values/variables only for a subset of the population. In this situation, the analyst can estimate Θ using cases where both variables are available.

2.2. The Effect of Misclassification on Survey Estimation

In this section we discuss the effect of misclassification on our survey estimates by looking at the bias in estimating a proportion based on A^* (rather than A) and the effect

of misclassification on effective sample size and variance estimates measured by the variance inflation factor due to weight variation.

2.2.1 Effect on Bias

To show the effect of misclassification on the bias, suppose we estimate the population proportion of a specific category *j* in variable *A*; that is, $P(A=j)=P_A(j)$. The customary design-based estimate of $P_A(j)$, using the variable that contains misclassification, can be computed as: $\hat{P}(A^*=j)=\hat{P}_{A^*}(j)=\sum_{i\in s} w_i I(A_i^*=j)/\hat{N}$ where *s* denotes the sample, $\hat{N}$ is a population size estimate, w_i is the survey weight for the *i*th case, and $I(\cdot)$ is the indicator function. If there is no misclassification, then $\hat{P}_{A^*}(j)=\hat{P}_A(j)=\sum_{i\in s} w_i I(A_i=j)/\hat{N}$; and if design-based estimation is used to calculate $\hat{P}_A=\left(\hat{P}_A(1),\cdots,\hat{P}_A(m)\right)^T$ that is an unbiased estimate of $P_A(j)$, then $E\left[\hat{P}_{A^*}(j)\right]=E\left[\hat{P}_A(j)\right]=P_A(j)$ where the expectation is evaluated with respect to the sample design. However, if the variable contains misclassification, since $E\left[I(A^*=j)\right]=\sum_{k=1}^{m}\theta_{jk}I(A=k)$, where the expectation is evaluated with respect to the misclassification model (1), then $E\left[\hat{P}_{A^*}(j)\right]=\sum_{k=1}^{m}\theta_{jk}\hat{P}_A(k)$. Or, using the notation of *m* dimensional vectors and matrix, $E\left[\hat{P}_{A^*}\right]=\Theta\hat{P}_A$ where $\hat{P}_{A^*}=\left(\hat{P}_{A^*}(1),\cdots,\hat{P}_{A^*}(m)\right)^T$ and $\hat{P}_A=\left(\hat{P}_A(1),\cdots,\hat{P}_A(m)\right)^T$. Assuming that $\hat{P}_A$ is an unbiased estimate of P_A, then the bias of $\hat{P}_{A^*}$ can be expressed as $Bias\left[\hat{P}_{A^*}\right]=E\left[\Theta\hat{P}_A\right]-P_A=\Theta P_A-P_A=\left(\Theta-I\right)P_A$, where *I* indicates the identity matrix of order *m*. When the true values are available from the sample, then this bias can be estimated as $\widehat{Bias}\left[\hat{P}_{A^*}\right]=\left(\widehat{\Theta}-I\right)\hat{P}_A$, or in terms of relative bias: $\widehat{\text{Relbias}}\left[\hat{P}_{A^*}\right]=\left(D_{P_A}\right)^{-1}\left(\widehat{\Theta}-I\right)\hat{P}_A$ where D_{P_A} is a diagonal matrix with diagonal elements $\hat{P}_A(j),\ j=1,\cdots,m$, and the elements of $\widehat{\Theta}$ are computed as $\hat{\theta}_{jk}=\left(\hat{N}_{\cdot k}\right)^{-1}\sum_{i\in s} w_i I(A_i^*=j,A_i=k)$, where $\hat{N}_{\cdot k}=\sum_{j=1}^{m}\hat{N}_{jk}$.

Note that the magnitude of bias is a function of two components: the misclassification parameter θ_{jk} and the true parameter $P_A(j)$.

2.2.2 Variance inflation effect

Stratification is usually used to produce better precision in survey estimation. When the sampling strata are constructed in a way that aligns with the analytical domains, an efficient sample can be obtained to produce a *self-weighting* sample within the domain of analysis. Misclassification, however, can occur in the stratification variables. If the domains of analyses are then constructed based on surveyed variables (assumed to be true values) rather than stratification variables (which contain misclassification), the weights within domains will vary. Depending on whether this weight variation is substantial or not, nontrivial weight variation will result in nontrivial loss of efficiency measured by loss in effective sample sizes or an increase in the variance.

At the sample design stage, optimum sample size for each domain has been allocated to meet some precision or cost requirement or both. When there is misclassification, optimal sample size may no longer be attained because the domain sample sizes may have changed. A ratio of domain sample sizes based on the true and misclassified variables defined as $n_d(A)/n_d(A^*)$ may indicate sample size reduction or increase due to misclassification, where the numerator is the sample size for domain *d* based on the true variable *A* and the denominator is the sample size for domain *d* based on the misclassified variable *A**. This sample size change may occur across domains, but the marginal total sample size for a particular variable and the grand total sample size will stay the same. Instead, the effective sample size (Kish 1965)—the sample size required for a simple random sample to have the same precision as the more complex sample design—can be used to reflect the effect of misclassification on survey estimation. The effective sample size for estimation in domain *d* is defined as: $n_{eff,d} = n_d / deff_{w,d}$ where n_d denotes the domain sample size and $deff_{w,d}$ denotes the design effect due to weight variation, i.e., $deff_{w,d} = (\sum_{i \in d} w_i^2)/(n\bar{w}_d^2) = 1 + CV_d^2$. We can then compare the two effective sample sizes using a ratio defined as $\{n_{eff,d}(A)\}/\{n_{eff,d}(A^*)\}$ where the numerator is the effective sample size using the true variable and the denominator is effective sample size specified using the misclassified variable. The ratio less than one indicates that the misclassification adversely affects the precision of the estimates; or vise versa.

To assess the magnitude of the misclassification effect on the survey variance for a specific domain, one can calculate the variance inflation factor (VIF), which is defined as the ratio of the design effects, where the numerator is based on the true variable and the denominator is based on the misclassified variable as follows: $VIF_d = deff_{w,d}(A)/deff_{w,d}(A^*)$. A value greater than one indicates that misclassification increases the variance of the domain estimates, while a value smaller than one indicates decreases.

2.3. Customary Adjustment Methods to Account for Misclassification Effects

Kuha and Skinner (1997) discuss several methods of adjustment that account for a misclassification effect in estimation including:

- Matrix method using validation data: The adjusted cell counts are obtained by multiplying the estimates based on the misclassified variable in the adjustment matrix.
- Model-based method using validation data: The cell counts are modeled on a log-linear model with the true and misclassified variables; their interactions are used as predictors.
- Repeated measurement method: Information about the misclassification parameters comes from repeated measurements of misclassified variables.

In our work, we implemented the adjustment using the matrix method, as described below. Suppose we want to estimate the population proportions $P_A = \left(P_A(1), \cdots, P_A(m)\right)^T$, where *A* may be defined as a variable with *m* categories. If the true *A* is available from the survey, as used in this case, an intuitive estimator for P_A is $\hat{P}_A$. However, this estimator may not be efficient because the sample was designed based on a misclassified variable.

Using the relationship $E\left[\hat{P}_{A^*}\right] = \Theta P_A$, the misclassification-adjusted estimate of P_A can be calculated as $\hat{P}_A^m = \widehat{\Theta}^{-1}\hat{P}_{A^*}$. Note that matrix Θ needs to be nonsingular so that its inverse exists.

For estimation of totals, the misclassification-adjusted estimates can be easily calculated by multiplying the proportion estimates with the population size— $\hat{N}_A^m = \hat{P}_A^m\ N$, where N denotes the population size. Total estimate can be also calculated directly based on totals as follows $\hat{N}_A^m = \widehat{\Theta}^{-1}\hat{N}_{A^*}$. Note, however, that this estimator may produce some negative numbers, which is unrealistic.

3. Application to the NSRCG

The National Survey of Recent College Graduates (NSRCG), sponsored by the National Science Foundation (NSF), collects education, employment, and demographic information from graduates of a U.S. college or university (including U.S. territories) who recently received a bachelor's or master's degree in science, engineering, or health (SE&H) fields. The NSRCG has a two-stage sample design in which schools are selected in the first stage and graduates in the second stage from a list obtained from selected schools. For details about the survey, go to www.nsf.gov/statistics/srvyrecentgrads.

In the first stage, a sample of schools was selected from the sampling frame obtained from the Integrated Postsecondary Education Data System database maintained by the National Center for Education Statistics (http://nces.ed.gov/ipeds/). Then, from each sampled school, lists of graduates were collected along with variables that are essential to determine survey eligibility and to also construct five sampling stratification variables: degree cohort, degree level, field of major, race/ethnicity, and gender. At the second stage, a stratified sampling was used to select a sample of graduates with sampling strata based on a combination of the five variables provided by schools selected at the first stage of sampling (Wilson et al. 2005 and Bandeh et al. 2006).

During the frame construction for the second sampling stage, the information provided by sampled schools was transformed into a standard format. This task included editing, coding, and imputation. Missing values in sampling variables were imputed. (See Jang et al. [2008] for details on coding and imputation and other procedures in the sampling frame construction.) At the end of this process, the race/ethnicity category was determined by the best available information: 1 = Non–Hispanic white, 2 = Non–Hispanic Asian, Pacific Islander/Hawaiian, 3 = Hispanic, Black, American Indian/Alaskan. When the survey was administered, race/ethnicity information was again collected from respondents. When information provided by the respondents differs from the information obtained from the schools, such differences may cause a nontrivial variation in weights and a loss in sample size for critical NSRCG domains, especially associated with specific race/ethnicity categories.

3.1. Misclassification of Race/Ethnicity Categories

In this section we present descriptive statistics on misclassifications that occurred in the race/ethnicity classification in the 2003 and 2006 NSRCG. We assumed that the information provided by graduates during data collection is more accurate than that provided by the sampled schools. We identified all mismatched cases between the school

and the respondent on race/ethnicity variables, presented as a cross-tabulation between two sources (see Table 1). Within each cell, we calculated weighted and unweighted counts.

Table 1: Race/Ethnicity Reported in Frame (Misclassified Value) and Survey (True Value): 2003 and 2006 NSRCG

Survey Year	Classification in Stratification	Classification with Reported Values			Total
		White	Asian	Minority	
2003	White	678,516 (5,240)	4,891 (38)	12,586 (87)	695,992 (5,365)
	Asian	136,099 (934)	134,386 (1,148)	26,834 (165)	297,319 (2,247)
	Minority	8,546 (209)	2,659 (69)	149,739 (2,941)	160,943 (3,219)
	Total	823,161 (6,383)	141,936 (1,255)	189,158 (3,193)	1,154,255 (10,831)
2006	White	1,196,301 (8,705)	9,636 (63)	28,473 (181)	1,234,409 (8,949)
	Asian	113,823 (999)	262,197 (2,248)	39,869 (304)	415,889 (3,551)
	Minority	9,841 (207)	4,130 (96)	269,749 (4,728)	283,720 (5,031)
	Total	1,319,964 (9,911)	275,963 (2,407)	338,091 (5,213)	1,934,018 (17,531)

Note: Numbers in the table are weighted counts and (in parentheses) unweighted counts. The race/ethnicity groups in this table are defined above.

3.1.1 Misclassification parameter and bias estimates

The estimates of the misclassification parameter, shown in Table 1, can be calculated as shown in Table 2. Both 2003 and 2006 NSRCG data showed substantial misclassification from white to Asian (16.5 percent in 2003 and 8.6 percent in 2006), and from minority to Asian (14.2 percent in 2003 and 11.8 percent in 2006). Relative bias estimates in race/ethnicity for the 2003 and 2006 data are shown in Table 3. Misclassification of race/ethnicity into Asian led to pronounced relative bias for estimates of the proportion of Asians (109.5 percent in 2003 and 50.7 percent in 2006).

Table 2: Misclassification Matrix for Race/Ethnicity: 2003 and 2006 NSRCG

Survey Year	Classification in Stratification	Classification with Reported Values		
		White	Asian	Minority
2003	White	82.42	3.44	6.65
	Asian	16.53	94.68	14.18
	Minority	1.03	1.87	79.16
2006	White	90.63	3.49	8.42
	Asian	8.62	95.01	11.79
	Minority	0.74	1.49	79.78

There could be several causes contributing to these errors including discrepancies due to inaccurate information in the school administrative data and coding or imputation processes. In the 2003 and 2006 NSRCG frame construction process, missing race/ethnicity values were imputed and/or classified into the three race groups—Asian, white and minority. The imputation used an algorithm that matched last and first name databases and the type of school that served predominantly minority students, when that information was available. Otherwise, all other missing cases were imputed with Asian race/ethnicity to sample them at the same rate as the Asian group.

Table 3: Relative Bias Estimate for Proportions of White, Asian, and Minority: 2003 and 2006 NSRCG

	2003	2006
Relative Bias of P_{White} (White vs. Others)	-15.4%	-6.5%
Relative Bias of P_{Asian} (Asian vs. Others)	109.5%	50.7%
Relative Bias of P_{Minority} (Minority vs. Others)	-14.9%	-16.1%

Further investigation into these misclassified race/ethnicity cases showed that 35 percent (527 out of 1,502) were imputed and of those 510 were imputed with Asian. Among those 510 cases, 396 were classified as "unknown race/ethnicity" by the school but turned out to be white according to survey responses. Table 4 shows misclassification (which is very high) only for imputed race/ethnicity cases. Among the 65 percent (975 out of 1,502) misclassified race cases that were not imputed, 345 cases were classified as "unknown race/ethnicity" by the school but turned out to be white in survey responses. Based on the misclassification for imputed and nonimputed race/ethnicity cases in the 2003 and 2006 NSRCG, classifying the "unknown race/ethnicity" into Asian may lead to a nontrivial misclassification error.

Table 4: Misclassification Matrix for Imputed Race/Ethnicity: 2003 and 2006 NSRCG

Survey Year	**Classification in Stratification**	**Classification with Reported Values**		
		White	Asian	Minority
2003	White	18.38	2.80	11.8
	Asian	81.05	96.76	70.54
	Minority	0.55	0.43	17.65
2006	White	33.81	6.57	10.77
	Asian	65.27	91.96	59.92
	Minority	0.9	1.45	29.3

We also conjectured that the misclassification error might not be ignored in the "nonresident alien" group since, for sampling purposes, these students are being grouped with Asians. To determine if "nonresident alien" cases made a significant contribution to the race/ethnicity misclassification, we counted these cases as follows: In 2003, there were only 671 "nonresident alien" cases (6.2 percent out of a total of 10,831 cases); and in 2006 there were only 1,053 "nonresident alien" cases (6 percent out of a total of 17,531 cases). Below is a comparison of the misclassification matrices for students with temporary U.S. resident visas/nonresident aliens and for U.S. citizens and permanent residents (Table 5).

Table 5: Misclassification Matrix for Race/Ethnicity Among Students with Temporary U.S. Resident Visas (Nonresident Alien): 2003 and 2006 NSRCG

Survey Year	Resident Status	Classification in Stratification	Classification with Reported Values		
			White	Asian	Minority
2003	Temporary U.S. Residents	White	9.2	0.0	1.010
		Asian	82.4	93.2	34.3
		Minority	8.4	6.8	64.6
	U.S. Citizens/Permanent Residents	White	83.48	4.74	2.78
		Asian	13.35	90.52	4.23
		Minority	3.18	4.74	92.99
2006	Temporary U.S. Residents	White	12.3	1.5	1.4
		Asian	84.0	94.6	56.1
		Minority	3.7	3.9	42.6
	U.S. Citizens/Permanent Residents	White	89.09	3.13	3.53
		Asian	8.85	92.85	4.36
		Minority	2.06	4.03	92.10

Table 5 clearly indicates that the misclassification is large among graduates with temporary U.S. resident visa (nonresident alien), especially from white to Asian. We expected that misclassification among U.S. citizens and permanent residents to be minor, but Table 5 shows that the misclassification from white to Asian cannot be ignored.

3.1.2 Effective sample size changes and Variance inflation factor

The NSRCG sample was designed to allocate sample size across key domains. In this section we evaluate effective sample size change as well as variance inflation due to misclassification in domains associated with race/ethnicity categories.

Figure 2 shows the ratios between the sample sizes of the true values to that of misclassified values for domains: race/ethnicity, degree level, major field, and gender, for 2003 and 2006 NSRCGs, respectively. The horizontal line at point 1.0 indicates a reference line with ratio equal to one. A point below this line indicates a sample size reduction for estimation within the particular domain due to misclassification. On the other hand, a point above this line indicates that there is a sample size increase for estimation within the particular domain due to misclassification. In Figure 2, ratios for Asians are less than one. This is consistent with the analyses of misclassification matrices in the previous section. Note that the sample size reduction is quite pronounced in some domains within the Asian group.

Figure 3 presents ratios of the effective sample sizes for domains defined by race/ethnicity, degree level, major field, and gender for the 2003 and 2006 NSRCG. Effective sample size for the estimation of Asian and minority groups decreased, as the ratios are smaller than one for a majority of estimation domains. Comparing white and minority groups for some domains in Figures 2 and 3, the increment in sample sizes in Figure 2 did not necessarily increase the effective sample sizes at the same rate as shown in Figure 3. This can be seen clearly for the minority group because misclassification could lead to weight variation within a domain so that the design effect is greater than one and, hence, reduces the effective sample size for domain estimation.

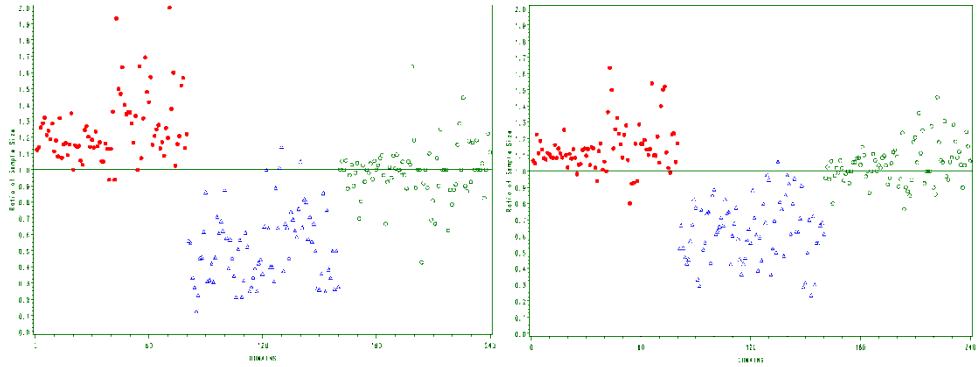

Figure 2: Ratio of domain sample size of the true value to domain sample size of the misclassified value for domains defined by race/ethnicity, degree level, major field, and gender: 2003 and 2006 NSRCG (● = White, △= Asian, ○ = Minority)

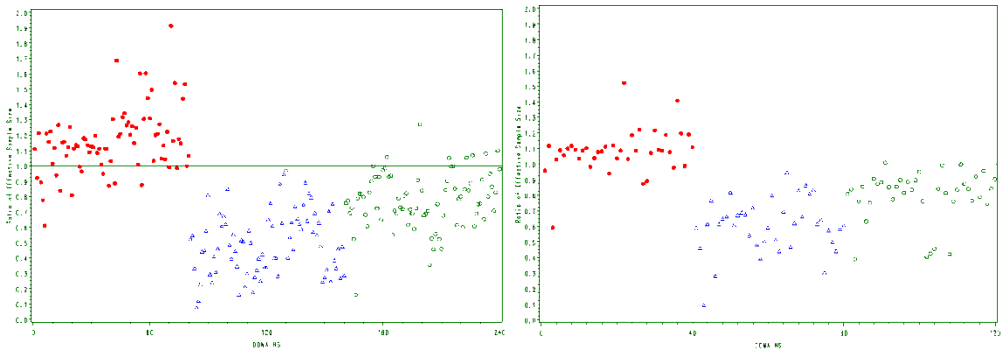

Figure 3: Ratio of domain effective sample size of the true value to domain sample size of the misclassified value for domains defined by race/ethnicity, degree level, major field, and gender: 2003 and 2006 NSRCGs (● = White, △= Asian, ○ = Minority)

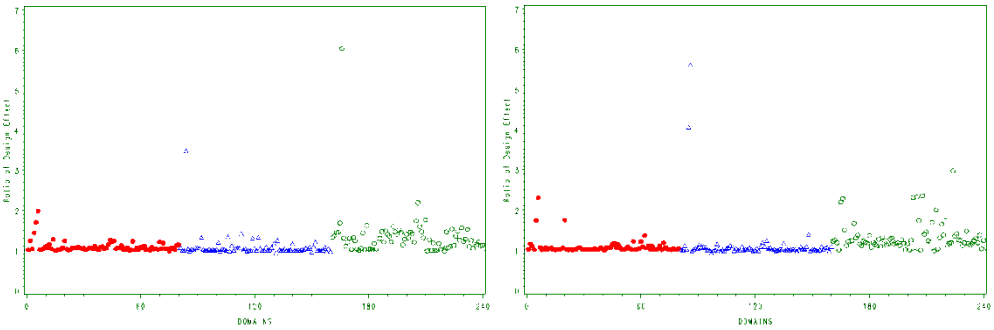

Figure 4: Variance inflation factor due to misclassification for domains defined by race/ethnicity. degree level. major field. and gender: 2003 and 2006 NSRCGs (● = White,△= Asian, ○ = Minority)

3.1.3 Variance Inflation Factor

Figure 4 shows the variance inflation factors due to weight variation of the true values compared to that of misclassified values. The horizontal line at point 1.0 indicates a reference line where the variance inflation factor is equal to one. A point above this line

indicates that estimated variances for domain estimates increase due to misclassification. On the other hand, a point below this line indicates that estimated variances for domain estimates decrease due to misclassification. Variance inflation factors for most domains are greater than one—a little more than one for white and up to two for the minority group—which means that misclassification affected efficiency of the estimates in terms of accuracy as measured by their variances.

3.2. Misclassification-Adjusted Estimates

We compared the estimates of the total number of graduates for domains defined by degree level, major, and race/ethnicity. These estimates were calculated using three different weights:

1. Final NSRCG survey weights. These weights are available from the NSRCG public-use file. They have been raked to the domain population totals for the five sampling variables and were constructed using the true/survey response variables.
2. Weights raked to the misclassified estimates. These weights have been raked to the domain population totals for the five sampling variables and were constructed using the misclassified/frame variables.
3. Weights raked to the misclassification-adjusted totals. These have been raked to the domain totals that have been adjusted for misclassification and were constructed using the true/survey response variables.

The resulting estimates are available from the authors upon request. Relative differences were also calculated along with estimates where the NSRCG estimates/standard errors (based on current/published weights) are used as the baseline; that is:

$$\text{Relative difference } (\hat{T}_i) = \frac{\hat{T}_i - \hat{T}_1}{\hat{T}_1} \times 100\%, \quad i = 2,3$$

$$\text{Relative difference } \left[se(\hat{T}_i) \right] = \frac{se(\hat{T}_i) - se(\hat{T}_1)}{se(\hat{T}_1)} \times 100\%, \quad i = 2,3$$

where

$\hat{T}_1$ = Estimates of domain counts (weighted) where weights are raked to the known domain totals and the domains are based on the true variables

$\hat{T}_2$ = Estimates of domain counts (weighted) where weights are raked to the known domain totals and the domains are based on the misclassified variables

$\hat{T}_3$ = Estimates of domain counts (weighted) where weights are raked to the misclassification-adjusted domain totals and the domains are based on the true variables.

A positive relative difference indicates that the current NSRCG estimate/standard error ($\hat{T}_1$ or $se(\hat{T}_1)$) is smaller than the alternative estimate/standard error ($\hat{T}_2$ or $se(\hat{T}_2)$ or $\hat{T}_3$ or $se(\hat{T}_3)$). On the other hand, a negative relative difference indicates that the current NSRCG estimate/standard error is greater than the alternative estimate/standard error. Note that the standard errors calculated in these tables are based on the Taylor series method, instead of the replication method.

There is a clear pattern of differences between the three estimates in the race/ethnicity domain. For the Asian and other race/ethnicity groups, the comparison of $\hat{T}_1$ and $\hat{T}_2$ indicated that there was serious misclassification affecting the precision of the estimates; the current NSRCG estimates are heavily overestimated in all majors, for both bachelor's and master's degrees, and both years of the survey. On the other hand, total estimates for the white group were underestimated in all majors, for both degree types. These results are consistent with the analyses of misclassification matrices (shown in section 3.2.1) where the proportion of misclassification of white and other races misclassified into Asian was high. This indicates that misclassification in the race/ethnicity variable cannot be ignored.

When comparing $\hat{T}_1$ and $\hat{T}_3$ of Asian bachelor's recipients in 2003, total estimates decreased by more than 100,000 cases while the total for white bachelor's recipients increased by more than 100,000. In the 2006 data, such differences are around 70,000 cases. The total estimates for Asian master's recipients decreased by about 25,000 cases while the total estimates for white master's recipients increased by about 23,000 for both years. This brings into question whether the misclassification-adjusted estimates $\hat{T}_3$ are realistic numbers, as shown in the following distribution of bachelor's recipients in 2003 (based on the adjusted estimates):

Race Group	Total Estimate
Asian	16,421
Black, non-Hispanic	80,222
Hispanic	79,348
White, non-Hispanic	738,515
Other race/ethnicity	15,964

The relative differences decreased across time when comparing the 2003 and 2006 estimates for domains by race/ethnicity. This could indicate improvement in information provided by the schools and/or coding process. This is consistent with the misclassification matrix where the proportion of misclassification also decreased from 2003 to 2006.

4. Summary and Discussion

We conducted an investigation on coding discrepancies between information provided by sampled schools and that provided by responding sampled students on race/ethnicity variables used for stratification, sample allocation, and graduate sample selection based on the 2003 and 2006 NSRCG data. Assuming that the information provided by the graduate during the survey was more accurate, or the true response, then such discrepancies can be viewed as a misclassification problem.

Misclassification was non-ignorable in the race/ethnicity variables, especially for white and other race groups to Asian. The effective sample size difference in the white group is 311 cases in 2003 and 430 in 2006, which means the estimation for the white domains used more sample size by 311 cases in the 2003 NSRCG and 430 in the 2006). On the other hand, estimation in the Asian group used 383 less sample size in the 2003 NSRCG

and 624 less in 2006. These large misclassification and severe effective sample size differences affected domain estimation in the NSRCG, especially for those involving the Asian group. In these domains, the current NSRCG estimates seemed to overestimate the population of Asian graduates.

There are several potential improvements that could be made to address these misclassification problems:

1. ***Increase Sample Size for Affected Domains.*** During the sample allocation stage, optimum sample sizes were allocated to each estimation domain. The goal was to have an efficient sample (i.e., equal weights for cases within domains). Misclassification, however, adds variability to the weights so that effective sample sizes for the affected domains will change. To account for this, sample size in the affected domains—especially the ones reduced due to misclassification—may be increased during the sampling design stage. For example, suppose a large number of white graduates were misclassified as Asian during the sample allocation stage. After data collection, sample size for the Asian group will decrease since misclassified cases will actually go into the white group. Thus, in the next NSRCG, this sample size reduction may be accounted for by allocating a larger sample size to the Asian group.

2. Coding for the Asian group in the 2003 and 2006 NSRCG included cases of "nonresident alien" as well as "unknown race/ethnicity" that cannot be imputed based on the prespecified algorithm. An improvement in coding can be attained by improving the collection of race flags for everyone (including nonresident aliens, unknown race/ethnicity, and multiple race) and then implementing a coding scheme based on these flags.

3. ***Adjustment through Weighting.*** When severe misclassification is detected, adjustments may be needed. In survey work, it is customary to carry out adjustments for errors through a weighting technique. For misclassification adjustment, this can be done by poststratification, or raking, correctly specifying the control (marginal) totals. In the previous sections, we have shown that information on misclassification parameters can be used to estimate misclassification-adjusted totals to be used in raking. Note, in this study we encountered a large impact of race/ethnicity misclassification compared to the survey estimates. However, we have not verified whether these estimates are reasonable or precise.

Further work may still be needed in the following areas:

- The second recommendation above has been implemented to improve coding of race/ethnicity in the 2008 NSRCG. It has been changed to focus on a different classification of race/ethnicity from the one used by schools and in prior NSRCG (Jang et al. 2008). Once survey responses are obtained, a similar misclassification analysis using the 2008 NSRCG data can be done to determine whether this change leads to a better result.

- The third recommendation above can take into account misclassification that exists in the frame data. However, we have not showed analytical justification for

the proposed method that adjusts the proportions to account for misclassification as follows:

$$\hat{P}_A^m = \widehat{\Theta}^{-1}\hat{P}_{A^*}$$

.

We need to show analytically that this method still provides unbiased and efficient estimates of survey outcomes.

Acknowledgements and Disclaimer

Work on this article was supported and funded by the National Science Foundation, contract SRS-0739949. The views expressed here are those of the authors and not necessarily those of the National Science Foundation.

REFERENCES

Bandeh, L.J., J. Donsig, D. Nancy, S. Miki, E.B. Mary, and L. Xiaojing. "2006 National Survey of Recent College Graduates: Sample Frame Development, Sampling, and Location Procedures, Final Report." Washington, DC: Mathematica Policy Research, Inc., August 2006.

Binder, D.A. "On the Variances of Asymptotically Normal Estimators from Complex Survey." *International Statistical Review*, vol. 51, 1983, pp. 279-292.

Groves, R.M., D. Dillman, J. Eltinge, and R. Little. *Survey Nonresponse*. New York, NY: John Wiley & Sons, 2002.

Groves, R.M., F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. *Survey Methodology*. New York, NY: John Wiley & Sons, 2004.

Jang, D., A. Sukasih, X. Lin, and S. Rahman. "Sample Design for the 2008 National Survey of Recent College Graduates." Washington, DC: Mathematica Policy Research, Inc., August 2008.

Kish, L. *Survey Sampling*. New York, NY: John Wiley & Sons, 1965.

Krewski, D., and J.N.K. Rao. "Inference from Stratified Samples: Properties of the Linearization, Jackknife, and Balanced Repeated Replication Methods." *The Annals of Statistics*, vol. 9, 1981, pp. 1010-1019.

Kuha, J., and C.J. Skinner. "Categorical Data Analysis and Misclassification." In *Survey Measurement and Process Quality*, edited by Lars E. Lyberg, Paul Biemer, Martin Collins, Edith De Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. New York, NY: John Wiley & Sons, 1997.

Lessler, J.T., and W.D. Kalsbeek. *Nonsampling Errors in Surveys*. New York, NY: John Wiley & Sons, 1992.

Little, R.J.A., and D.B. Rubin. *Statistical Analysis with Missing Data*. New York, NY: John Wiley & Sons, 2002.

Lyberg, L.E., P. Biemer, M. Collins, E.D. De Leeuw, C. Dippo, N. Schwarz, and Trewin, *Survey Measurement and Process Quality* . New York, NY: John Wiley & Sons, 1997.

Rao, J.N.K. *Small Area Estimation*. New York, NY: John Wiley & Sons, 2003.

Shao, J. "Resampling Methods in Sample Survey." *Statistics*, vol. 27, 1996, pp. 203-254.

Wilson, C., D. Jang, T. Barton, M. Pierzchala, K. Kang, and J. Tsapogas. "2003 National Survey of Recent College Graduates: Methodology Report, Preliminary Draft." Washington, DC: Mathematica Policy Research, Inc., November 2005.

Wolter, K.M. *Introduction to Variance Estimation*. New York, NY: Springer-Verlag, 1985.