NONRESPONSE ISSUES OF THE NATIONAL SURVEY OF COLLEGE GRADUATES

Antoinette Tremblay and Thomas F. Moore III

1. Survey Description

The National Survey of College Graduates (NSCG) of the National Science Foundation (NSF) collects data on U.S. scientists and engineers; it attempts to capture and measure their unusual importance to the nation's continued productivity and economic growth. The 1993 NSCG sample design reflects the efforts that were taken to reduce the effects of nonresponse. The sample came from the Sample Edited Detail File (SEDF), which contains data gathered from the 1990 Decennial Census long forms. Persons on the SEDF who were noninstitutionalized, U.S. residents with a Bachelor's Degree or higher, and under 76 years of age as of April 1, 1993 were eligible for the sample (Census 1993). The file was stratified by the following cells: highest degree (3), sex (2), NSF group (8, combination of place of birth, disabled, race, Hispanic origin), and occupation (19). Of the 912 possible cells, 863 were nonempty. Sampling was generally proportional to stratum size, with adjustments for minimum counts needed for publication. The adjustments had the effect of oversampling special groups, such as women and minorities, to increase the accuracy of the data for publication. Thus, the 1993 NSCG sample size is 214,643 persons. When the base weight[1] is applied to these records, the resulting size is 31,809,582. The data was collected in three phases: mail, computer-assisted telephone interviewing (CATI), and personal visit.

2. Research Purpose and Scope

The research described herein is a component of a NSCG Nonresponse and Frame Bias Report, which is to assess the potential for bias in the current and subsequent NSCGs due to survey nonresponse and frame defects. In correspondence from NSF to the Bureau of the Census (Kruytbosch 1994), the following excerpts are provided:

> "In deciding on whether to continue the NSCG panel in future years, an assessment must be made as to the nature of the non-response bias. The problem with longitudinal surveys is the often cumulative nature of non-response... Because of the cumulative nature of the non-response and the need for NSF to assess the potential for bias, we would like the Census Bureau to analyze and prepare a report on non-response. This should include the full spectrum of quantitative and qualitative information available on this topic... Topics to investigate could include...a demographic analysis of non-response based on the 90 Census information..."

In this vein, three main problems/projects are addressed in this research to provide an interpretable picture of a structure for the data. Across all projects, the ultimate goal is to achieve an understanding of which demographic variables, or interactions thereof, drive the phenomenon of nonresponse. First, if nonrespondents of the 1993 NSCG appear to differ substantially from the respondents, a sample of them would have been desirable in the 1995 NSCG sample, either as part of the survey or as a separate sample for analytical purposes. By comparing various 1990 census demographic variables of these nonrespondents and respondents, it can then be determined if a correlation exists between certain demographic variables and nonresponse. Second, nonresponse may be correlated with frame variables or may result from survey procedures. Thus, the demographic comparison/classification analysis is repeated by reason for nonresponse. Third, a discussion of preliminary results across the data collection methods

[1] The Base Weight is the Sampling Take Every multiplied by the Subsampling Take Every.

of mail, CATI, and personal visit is presented. For example, is it worth using only one or two interview modes for various subsets of the sample?

Each project is similar in data requirements and methodology. Final status of the NSCG and fourteen census demographic variables are obtained from the 1993 NSCG data file. First, there is a large amount of background and exploratory data presentation to satisfy NSF's request for a 'full spectrum' of information. This includes response rates by the demographic variables and by data collection mode. Second, considerable effort is expended into providing simple characterizations of the conditions (i.e., profiles) that determine when a sampled person is in one class (i.e., nonresponse) rather than another (i.e., response). This is performed using the classification option of the Classification and Regression Trees (CART) statistical software of the California Statistical Software, Inc. (CSSI 1985, 1993). As stated in *Classification and Regression Trees* (Breiman, Friedman, Olshen, Stone 1984), "An important criterion for a good classification procedure is that it not only produce accurate classifiers, within the limits of the data, but that it also provide insight and understanding into the predictive structure of the data." CART was selected since it has the capabilities to provide invaluable profiles of respondents and nonrespondents.[2] As an illustration, output from CART could indicate, via profiles, that for these groups of sampled persons with these characteristics, respondents are significantly different from nonrespondents. Thus, something different (e.g., extra nonresponse conversion efforts) should possibly done for these special groups. These profiles would also be invaluable in the second project's intentions, as would their contributions to the third project of modes of data collection. For example, while always striving to maximize response, decisions could be made as to which 'type' of persons should undergo which type(s) of interview modes.

More 'traditional' analyses were desired to augment the exploratory and CART analyses. Although the large sample size would render significant results for other analyses, it was decided to conduct chi-square and regression analyses for several portions of this research, if only to obtain an ordering/prioritization of the 'most significant' and 'least significant' variables. Chi-square analysis evaluates the relationships between each demographic variable and response/nonresponse for the entire dataset, by reason for nonresponse for the nonrespondents, and by response/nonresponse for the entire dataset by mode of data collection. Chi-square tests, via contingency tables, are then conducted with the null hypothesis of independence. A design effect of 1.4 is incorporated into the chi-square test statistics and alpha=0.10. SAS logistic regression, using a stepwise procedure, is also conducted. It investigates which demographic variables, if any, can be used to predict response/nonresponse. Because of timing constraints, this regression is only conducted for the demographic comparison of respondents and nonrespondents, with the nonrespondents including the out-of-scope records.

Unweighted and weighted background data and results are presented throughout the research. However, since CART software is unable to incorporate sampling weights, all associated results are unweighted.

3. A Few Definitions and Clarifications

The terms error, misclassification, error improvement, true error rate and total error rate are used throughout the research and are defined as they occur. Let it suffice, here, to state that a (misclassification) error results when CART either classifies an 'actual' respondent as a nonrespondent, or an 'actual' nonrespondent as a respondent. The total error rate is the percent of misclassifications resulting from the specified CART analysis. The true error rate is the percent of misclassifications, void of

[2] Although CART is a 'relatively' new software package to the Census Bureau, it has been successfully used operationally in the classification of farm/not farm records of the 1992 Census of Agriculture mailing list (Ash, Kraus, Peterson 1995).

demographic data, if all respondents (or nonrespondents) are classified to the prominent category. Error improvement or reduction in error results when the use of demographic data by CART decreases the number of misclassifications from when all cases are classified in the prominent category of response or nonresponse.

Regardless of the analysis methodology, three basic types of comparisons can exist throughout the research. Each differs on the definition of nonresponse, but all aim to reduce the cost of misclassification. First, respondents are compared to the combined class of nonrespondents and out-of-scopes; second, respondents are compared to nonrespondents, excluding the out-of-scopes from the analysis; lastly, a comparison is sometimes made across respondents and nonrespondents and out-of-scopes. Therefore, depending on the definition employed, the composition of nonrespondents may or may not include the out-of-scope person records.

Throughout many of the tables presented in this report, components of a total may not sum exactly to the total; this is due to rounding. Also, whenever clarifications are needed for a table, the footnote will appear only on the unweighted data table.

4. Comparison of Results with Literature Review

Before undertaking this research, the analyst was confident that several 1990 census demographic variables would clearly distinguish 1993 NSCG respondents from nonrespondents. Other analysts and managers concurred, even pointing to the support of these differences in past research. However, after an extensive literature review, past research offers little guidance on descriptions of what kinds of persons tend to be survey nonrespondents. Indeed, although certain demographic characteristics exhibit more consistent results than others, there exists a wide variation in results over different studies in the literature. These studies differ in terms of data collection mode(s), longitudinal versus cross-sectional surveys, analysis methodology, and, mostly, where/how the data for the nonrespondents was obtained. For example, DeMaio (1980) observed no differences on Current Population Survey (CPS) between respondents and refusers for race and sex, but substantial differences in age and income. However, in contrast to the research herein where data for the nonrespondents is obtained from a three-year old Decennial Census, data for the CPS refusals was obtained by interviewer completed forms, geographic information, and imputation. The reader is directed to read *Survey Errors and Survey Costs* (Groves 1989), especially the chapter "Empirical Correlates of Survey Participation" for a broad review of sources and findings in terms of the demographic characteristics of refusers.

5. CART and its Restrictions to this Research

CART is a nonparametric statistical analysis program that can automatically find hidden structures in data via binary tree construction. Simplistically, it is a method which selects salient features of the data, discards the background noise, and feeds back understandable summaries of the information. In its analysis of categorical (classification) or continuous (regression) variables, decisions can then be based on that structure.

This report addresses a classification problem, with CART creating learning and test samples for the estimation of the classifier and its accuracy.[3] The learning and test samples are sampled independently

[3] CART developers indicate that although the test sample approach has the drawback that it reduces effective sample size, test sample estimation is honest and efficient if the total data size is large (i.e., greater than 30,000 records). Other estimation methods are the bootstrap, cross-validation, and resubstitution.

from a desired distribution using a pseudo-random number generator. The learning sample is usually two-thirds of the records and the remaining one-third of the data comprises the test sample. Simplistically, CART then constructs binary decision trees from the input variables by performing various nonparametric statistical operations on the learning sample which maximize the homogeneity of the dependent variable within each of the branches. The best split on the best variable(s) at each tree node is then selected. Thus, a classifier is constructed. Using the test sample, a pruning process then selects the most efficient tree having close to the minimum estimated error (i.e., misclassification) rate. Each terminal node of such a tree gives the predicted class or value of the response of a record in the node. In determining to which class each tree node is classified, CART simplistically assigns the 'profile' to that class having the majority of records. CART output provides classification results for both the test and learning samples. (Within rounding error, they are identical in this analysis.)

A restriction of CART is its inability to handle sampling weights; therefore, only unweighted data are used in the analysis. A second restriction of CART involves its creation of the learning sample. CART allows the analyst to input a learning sample size other than the default of 20,000 records. However, internal CART software restricts the size to be less than or equal to 99,999 records. (The test sample size is then automatically generated as 1/2 this input.) As detailed in Section 1, there are 214,643 total person records available for analysis, each assigned a final status of response, out-of-scope, or nonresponse.[4] When comparing nonrespondents to the combined class of nonrespondents and out-of-scopes, for example, it would be desirable if the learning sample consisted of 143,095 records (2/3 of 214,643) and the test sample consisted of 71,548 records (1/3 of 214,643 or 1/2 of learning sample size). When restricted to a learning sample size of 99,999 and then generating a test sample size of 49,509, only 70-77% of the available data is used in any of the three types of classifications, with 47%-51% being allocated to the development of the learning sample and 23-25% to the test sample. Programmers from the California Software, Inc. are currently working on removing this limitation. Regardless, the analyst is confident that the results and conclusions presented in this report are valid.

6. Background Information and Data Input

A related by-product of this research is an extensive amount of background and exploratory data analysis for each of the three projects, for all demographic variables. (Additional background data for the third project related to mode of data collection is contained in Section 7.3.) Most of the tabulations in this section are provided for unweighted and weighted data. *All discussion applies to the weighted tabulations, presented in rounded 000s.*

Table 1 provides the final status codes and their descriptions for the 1993 NSCG. (The breakout of status code 10 is provided for the analysis by nonresponse reason.) Of the 31,810 weighted person records, 22,626 (71.13%) are defined as respondents, 6,435 (20.23%) as nonrespondents, and 2,750 (8.65%) as out-of-scope. With some additions, the variables available for classification are the same as were used in the 1993 NSCG sample selection; Table 2 shows minor regroupings of the possible values. Note that NSFGRP is a combination of the variables RACE, ORIGIN, PBIRTH, CTZN, and the various limitation variables. Table 3 is a cross tabulation of these available demographic variables by the actual final status codes. Cell entries are percentages. Of interest are the differing percentages for PBIRTH and CTZN for emigrants (code 5), as compared to the other final status codes and other variables.

[4] For the second project involving reasons for nonresponse, the final status codes of out-of-scope and nonresponse are broken out further.

Table 1a. NSCG Final Status Codes, Unweighted Data

Final Status Code		Description	Total	%
Response	1	Complete	148932	69.39
Out-of-Scope[5]	2	Age Over 75	211	0.10
	3	Deceased	2407	1.12
	4	No Bachelor's Degree	14315	6.67
	5	Emigrant	2132	0.99
	6	Institutionalized	159	0.07
Nonresponse	7	Ill	1833	0.85
	8	Refusal	15082	7.03
	9	Incomplete	759	0.35
	10	Anything Else	28813	13.42
	10a	(PMR[6]) with correction (move)	298	0.14
	10b	(PMR) Jeffersonville correction	4	<0.01
	10c	(PMR) move, no forwarding	19460	9.07
	10d	(PMR) forwarding expired	14	<0.01
	10e	(PMR) temporarily absent,...	2090	0.97
	10f	Not located	5300	2.47
	10g	Wrong person	1333	0.62
	10h	Foreign address, APO	24	0.01
	10i	Not received	290	0.14
Total			214643	100.00

[5] Codes 2, 3, and 4 are permanently out-of-scope; codes 5 and 6 are temporarily out-of-scope.

[6] PMR denotes postmaster returned. There are 6,250 records which began with CATI, and did not go through the mail mode of data collection. For these records that have the final status codes listed here as 10a-10e, the PMR notation should be ignored.

Table 1b. NSCG Final Status Codes, Weighted Data

Final Status Code		Description	Total (000s)	%
Response	1	Complete	22626	71.13
Out-of-Scope	2	Age Over 75	34	0.11
	3	Deceased	315	0.99
	4	No Bachelor's Degree	2164	6.80
	5	Emigrant	217	0.68
	6	Institutionalized	20	0.06
Nonresponse	7	Ill	247	0.78
	8	Refusal	2329	7.32
	9	Incomplete	108	0.34
	10	Anything Else	3751	11.79
	10a	(PMR) with correction (move)	39	0.12
	10b	(PMR) Jeffersonville correction	<1	<0.01
	10c	(PMR) move, no forwarding	2500	7.86
	10d	(PMR) forwarding expired	2	<0.01
	10e	(PMR) temporarily absent,...	271	0.85
	10f	Not located	710	2.23
	10g	Wrong person	187	0.59
	10h	Foreign address, APO	3	<0.01
	10i	Not received	38	0.12
Total			31810	100.00

Table 2. Demographic Variables Available for Classification

Demographic Characteristic (variable name)	Values
Age Group (AGEGRP)	1=[16,29]; 2=[30,59]; 3=60+
Sex (SEX)	1=Male; 2=Female
Race (RACE)	1=White 2=Black 3=Native American 4=Asian/Pacific Islander 5=Other
Spanish/Hispanic Origin (ORIGIN)	1=No; 2=Yes
Place of Birth (PBIRTH)	1=US or Outlying Area; 2=Other
Citizenship (CTZN)	1=Yes; 2=No
Highest Education Degree (EDUC)	1=Bachelor's or Professional 2=Master's 3=Doctorate
Occupation Group (OCCGRP)	1=Physics/Life/Biology Scientists 2=Math/Computer Scientists 3=Social Scientists 4=Engineers, Architects, Surveyors 5=Other
Mobility Limitation Status (MOLIMT)	1=Yes; 2=No
Personal Care Limitation (PCLIMT)	1=Yes; 2=No
Work Limitation Status (WRKLIMT)	1=Yes; 2=No
Work Prevention Status (WRKPVT)	1=Yes; 2=No
NSF Group (NSFGRP)	1=Disabled 2=Hispanic 3=White/Other 4=Black 5=Asian/Pacific Islander 6=Native American 7=Foreign Born, US Citizen 8=Foreign Born, NonUS Citizen
Metropolitan Statistical Area (MSA)	1=Yes; 2=No; 9=Missing

Table 3a. Classification Variables by Final Status Code (%), Unweighted Data

Demographic Variable	Final Status Code										Total
	1	2	3	4	5	6	7	8	9	10[7]	
AGEGRP=1	18	10	5	26	31	16	19	15	17	39	21
AGEGRP=2	72	27	50	64	65	50	65	76	62	57	69
AGEGRP=3	10	62	45	11	4	33	16	10	21	4	10
SEX=1	59	51	73	53	64	69	66	63	53	56	59
SEX=2	41	49	27	47	36	31	34	37	47	44	41
RACE=1	79	80	83	68	57	75	71	79	68	64	76
RACE=2	9	14	10	15	4	16	12	10	12	17	10
RACE=3	1	<1	1	1	<1	3	1	1	1	1	1
RACE=4	10	5	5	11	36	2	14	9	16	14	11
RACE=5	2	1	1	4	4	4	2	1	2	4	2
ORIGIN=1	94	94	95	87	88	90	93	95	89	89	93
ORIGIN=2	6	6	5	13	12	10	7	5	11	11	7
PBIRTH=1	83	80	87	73	31	89	70	81	67	70	80
PBIRTH=2	17	20	13	27	69	11	30	19	33	30	20
CTZN=1	93	91	96	86	45	94	87	93	83	81	91
CTZN=2	7	9	4	14	55	6	13	7	17	19	9
EDUC=1	69	73	71	85	61	73	73	74	78	75	71
EDUC=2	26	22	24	13	28	21	22	22	18	20	24
EDUC=3	5	4	5	2	11	6	5	4	4	4	5
OCCGRP=1	4	2	3	2	4	0	3	3	2	3	3
OCCGRP=2	5	1	3	2	4	3	4	5	3	4	5
OCCGRP=3	3	2	2	1	3	1	3	3	1	3	3
OCCGRP=4	13	4	10	7	10	8	9	12	8	9	12
OCCGRP=5	76	91	83	88	79	89	82	78	86	81	78
MOLIMT=1	1	9	17	3	1	30	6	2	2	1	2
MOLIMT=2	99	91	83	97	99	70	94	98	98	99	99
PCLIMT=1	2	10	12	5	2	24	6	3	4	3	3
PCLIMT=2	98	90	88	95	98	76	94	97	96	97	97
WRKLIMT=1	6	28	41	10	2	50	14	7	11	6	7
WRKLIMT=2	94	72	59	90	98	50	86	93	89	94	93

(continued)

[7] The detailed breakout of status code 10 into 10a-10i is provided in the Results Section, Reasons for Nonresponse.

Table 3a (continued). Classification Variables by Final Status Code (%), Unweighted Data

Demographic Variable	Final Status Code										Total
	1	2	3	4	5	6	7	8	9	10	
WRKPVT=1	2	22	24	4	1	39	8	2	5	2	2
WRKPVT=2	98	78	76	96	99	61	92	98	95	98	98
NSFGRP=1	7	29	41	11	2	51	14	9	11	7	8
NSFGRP=2	4	1	2	6	3	4	4	3	4	6	4
NSFGRP=3	61	38	36	41	22	23	41	5	39	41	56
NSFGRP=4	7	11	6	11	2	9	9	8	9	13	8
NSFGRP=5	2	0	1	2	2	0	1	8	2	2	2
NSFGRP=6	1	<1	1	1	<1	2	1	2	1	1	1
NSFGRP=7	11	11	10	13	15	6	17	1	16	11	11
NSFGRP=8	7	9	4	14	55	6	13	1	17	19	9
								2			
								7			
MSA=1	11	14	13	11	8	11	9	8	9	8	10
MSA=2	89	85	87	88	91	88	91	9	90	92	90
MSA=9	1	1	1	1	<1	1	1	1	1	1	1
								1			
Total	69	<1	1	7	1	<1	1	7	<1	13	

Table 3b. Classification Variables by Final Status Code (%), Weighted Data

Demographic Variable	Final Status Code										Total
	1	2	3	4	5	6	7	8	9	10	
AGEGRP=1	17	11	5	26	33	17	20	14	17	40	20
AGEGRP=2	72	27	50	63	63	49	65	76	62	56	69
AGEGRP=3	11	62	45	11	4	34	16	10	21	4	10
SEX=1	54	47	72	49	59	66	65	60	47	52	54
SEX=2	46	53	28	51	41	34	35	40	53	48	46
RACE=1	90	88	89	82	66	84	83	89	82	78	88
RACE=2	5	8	7	9	3	12	7	6	7	11	6
RACE=3	<1	<1	<1	<1	<1	1	<1	<1	<1	<1	<1
RACE=4	4	3	3	6	28	1	8	4	10	8	5
RACE=5	1	<1	<1	2	3	3	1	1	1	2	1
ORIGIN=1	97	96	97	93	90	94	96	98	94	94	97
ORIGIN=2	3	4	3	7	10	6	4	2	6	6	3
PBIRTH=1	92	89	92	84	47	92	82	90	80	82	89
PBIRTH=2	8	11	8	16	53	8	18	10	20	18	11
CTZN=1	97	95	98	91	58	96	92	96	89	89	95
CTZN=2	3	5	2	9	42	4	8	4	11	11	5
EDUC=1	72	75	73	87	68	73	76	77	82	79	74
EDUC=2	24	22	23	11	24	23	21	20	15	18	22
EDUC=3	4	3	4	2	8	4	3	3	3	3	4
OCCGRP=1	1	1	1	1	2	0	1	1	1	1	1
OCCGRP=2	2	<1	1	1	2	1	1	2	1	2	2
OCCGRP=3	1	1	1	1	1	<1	1	1	1	1	1
OCCGRP=4	4	1	4	2	5	3	3	4	3	3	4
OCCGRP=5	91	97	93	96	90	96	93	92	95	92	92
MOLIMT=1	1	6	12	2	1	23	4	1	1	1	1
MOLIMT=2	99	94	88	98	99	77	96	99	99	99	99
PCLIMT=1	1	6	8	3	2	18	4	2	3	2	2
PCLIMT=2	99	94	92	97	98	82	96	98	97	98	98
WRKLIMT=1	4	18	29	7	2	37	10	4	7	4	4
WRKLIMT=2	96	82	71	93	98	63	90	96	93	96	96

(continued)

Table 3b (continued). Classification Variables by Final Status Code (%), Weighted Data

Demographic Variable	Final Status Code										Total
	1	2	3	4	5	6	7	8	9	10	
WRKPVT=1	1	14	18	3	1	30	6	1	3	1	1
WRKPVT=2	99	86	82	97	99	70	94	99	97	99	99
NSFGRP=1	4	19	29	7	2	38	10	5	7	5	5
NSFGRP=2	2	1	1	3	2	2	2	1	2	3	2
NSFGRP=3	81	63	57	66	40	45	64	78	64	65	77
NSFGRP=4	4	6	4	7	1	7	5	5	5	8	5
NSFGRP=5	1	0	1	1	1	0	1	1	1	1	1
NSFGRP=6	<1	<1	<1	<1	<1	1	<1	<1	<1	<1	<1
NSFGRP=7	5	6	6	8	11	4	10	6	10	7	6
NSFGRP=8	3	5	2	9	42	4	8	4	11	11	5
MSA=1	12	19	13	14	9	14	10	9	12	9	12
MSA=2	87	80	87	86	91	85	89	91	87	91	88
MSA=9	1	<1	1	1	<1	1	1	1	1	1	1
Total	71	<1	1	7	1	<1	1	7	<1	12	

Other related background information includes response rates.[8] They are provided across various demographic variables in Tables 4 through 8 only to depict nonresponse by different categories; they were looked at independently of the CART, chi-square and logistic regression analyses. The low response rate of 64.89% for NSFGRP=(8) Foreign NonUS Citizen (Table 8b) is, again, related to Table 3's values of PBIRTH and CTZN for emigrants.

[8] Response Rate = (Complete + Out-of-Scope) / Total.

Table 4a. Response Rate by AGEGRP, Unweighted Data (000s)

	Resp.	Out-of-Scope		Non-Response	Total	Response Rate %
		Perm	Temp			
AGEGRP=1	26791	3828	691	13934	45244	69.20
AGEGRP=2	106788	10371	1459	29514	148132	80.08
AGEGRP=3	15353	2734	141	3039	21267	85.71
Total	148932	16933	2291	46487	214643	78.34

Table 4b. Response Rate by AGEGRP, Weighted Data

	Resp.	Out-of-Scope		Non-Response	Total	Response Rate %
		Perm	Temp			
AGEGRP=1	3911	584	75	1915	6484	70.47
AGEGRP=2	16302	1524	146	4075	22047	81.52
AGEGRP=3	2414	404	16	444	3278	86.46
Total	22626	2512	237	6434	31810	79.77

Table 5a. Response Rate by SEX, Unweighted Data

	Resp.	Out-of-Scope		Non-Response	Total	Response Rate %
		Perm	Temp			
SEX=1	87664	9413	1481	27331	125889	78.29
SEX=2	61268	7520	810	19156	88754	78.42
Total	148932	16933	2291	46487	214643	78.34

Table 5b. Response Rate by SEX, Weighted Data (000s)

	Resp.	Out-of-Scope		Non-Response	Total	Response Rate %
		Perm	Temp			
SEX=1	12211	1291	141	3578	17221	79.23
SEX=2	10416	1221	95	2856	14588	80.42
Total	22626	2512	237	6434	31810	79.77

Table 6a. Response Rate by OCCGRP, Unweighted Data

	Resp.	Out-of-Scope		Non-Response	Total	Response Rate %
		Perm	Temp			
OCCGRP=1	5328	295	89	1454	7166	79.71
OCCGRP=2	7324	415	88	2026	9853	79.44
OCCGRP=3	4210	249	67	1366	5892	76.82
OCCGRP=4	18924	1210	219	4491	24844	81.92
OCCGRP=5	113146	14764	1828	37150	166888	77.74
Total	148932	16933	2291	46487	214643	78.34

Table 6b. Response Rate by OCCGRP, Weighted Data (000s)

	Resp.	Out-of-Scope		Non-Response	Total	Response Rate %
		Perm	Temp			
OCCGRP=1	290	15	4	73	382	81.00
OCCGRP=2	391	21	4	105	522	79.95
OCCGRP=3	250	13	3	75	342	77.94
OCCGRP=4	1007	62	10	232	1311	82.30
OCCGRP=5	20688	2401	215	5949	29253	79.66
Total	22626	2512	237	6434	31810	79.77

Table 7a. Response Rate by EDUC, Unweighted Data

	Resp.	Out-of-Scope		Non-Response	Total	Response Rate %
		Perm	Temp			
EDUC=1	102278	14063	1412	34790	152543	77.19
EDUC=2	38589	2406	625	9673	51293	81.14
EDUC=3	8065	464	254	2024	10807	81.27
Total	148932	16933	2291	46487	214643	78.34

Table 7b. Response Rate by EDUC, Weighted Data (000s)

	Resp.	Out-of-Scope		Non-Response	Total	Response Rate %
		Perm	Temp			
EDUC=1	16285	2141	163	5045	23635	78.65
EDUC=2	5487	320	56	1200	7063	83.01
EDUC=3	854	51	18	189	1112	83.02
Total	22626	2512	237	6434	31810	79.77

Table 8a. Response Rate by NSFGRP, Unweighted Data

	Resp.	Out-of-Scope		Non-Response	Total	Response Rate %
		Perm	Temp			
NSFGRP=1	11035	2646	124	3635	17440	79.16
NSFGRP=2	5885	970	77	2337	9269	74.79
NSFGRP=3	90566	6815	504	21641	119526	81.89
NSFGRP=4	10843	1762	51	5202	17858	70.87
NSFGRP=5	3506	309	34	883	4732	81.34
NSFGRP=6	1159	195	7	479	1840	73.97
NSFGRP=7	15995	2155	319	5389	23858	77.41
NSFGRP=8	9943	2081	1175	6921	20120	65.60
Total	148932	16933	2291	46487	214643	78.34

Table 8b. Response Rate by NSFGRP, Weighted Data (000s)

	Resp.	Out-of-Scope		Non-Response	Total	Response Rate %
		Perm	Temp			
NSFGRP=1	953	251	11	326	1541	78.87
NSFGRP=2	350	60	5	142	556	74.55
NSFGRP=3	18262	1625	96	4479	24463	81.69
NSFGRP=4	920	163	5	447	1535	70.88
NSFGRP=5	208	20	2	52	282	81.49
NSFGRP=6	47	8	<1	20	75	73.87
NSFGRP=7	1161	191	25	421	1797	76.58
NSFGRP=8	726	194	92	548	1560	64.89
Total	22626	2512	237	6434	31810	79.77

7. Results

Results via CART, chi-square, logistic regression, and various exploratory analyses of the data are provided, in varying degrees, for each of the three nonresponse projects. Since none of the projects produce distinguishing characteristics for respondents and nonrespondents, detail on the second and third projects is far less than is provided for the first. The extensive detail for the first project is provided to document the basic methodology used for all three. As detailed earlier, all CART input are unweighted data. Also, the reader is reminded that NSFGRP is a combination of the variables RACE, ORIGIN, PBIRTH, CTZN, and the various limitation variables. Thus, all analyses presented were conducted with and without the inclusion of NSFGRP; if it was excluded, its components were included, and if it was included, its components were excluded. Results from analyses that included both NSFGRP and its components were, of course, unreliable.

7.1 Demographic Comparison of Respondents and Nonrespondents

The various classifications performed via CART are based on three different class definitions for the response variable. First, respondents are compared to the combined class of nonrespondents and out-of-scopes; second, respondents are compared to nonrespondents, excluding the out-of-scopes from the analysis; lastly, a comparison across respondents and nonrespondents and out-of-scopes is made.

7.1.1 Respondents vs. (Nonrespondents + Out-of-Scopes)

CART

If no prior information is available, and the out-of-scope records are included with the nonrespondent records,[9] the best that one can do is to classify all records as respondents. One-hundred percent of the actual respondents are correctly classified by CART as respondents, and 100% of the actual (nonrespondents + out-of-scopes) are incorrectly classified as respondents. Then, a true error rate of 30.61% exists. However, because CART uses a learning sample of 99,999 records, the resulting error rate is 30.38%. This error rate is detailed in Table 9.

Table 9. CART Classifications of Resp vs. (NR+OS), No Prior Info

CART	Actual		Total
	Resp	NR+OS	
Resp Total Classification Rate	69615 69.62% 100.00%	30384 30.38% 100.00%	99999 100.00%
NR+OS Total Classification Rate	0 0% 0%	0 0% 0%	0 0%
TOTAL	69615 69.62%	30384 30.38%	99999
Total error rate = (0+30384) / 99999 = 30.38%			

[9] Out-of-Scope and nonrespondent are often abbreviated as OS and NR, respectively, in this paper.

In contrast to Table 9, prior knowledge of CTZN, AGEGRP, RACE, OCCGRP, WRKPVT, EDUC, ORIGIN, and PCLIMT provides a classification tree with an error rate of 29.43%. However, this is an improvement of only 0.95% than that of classifying all records as respondents (i.e., Table 9 vs. Table 10: 30.38%-29.43%). This listing of demographic variables also indicates the prioritized order of each variable's contribution to the reduced total error. It is interesting to relate these results back to the independent analysis provided in Table 8. Several of the components of NSFGRP (i.e., CTZN, RACE, ORIGIN, WRKPVT, and PCLIMT) are involved in the CART classification. (When NSFGRP, and none of its components, was included in the CART input, a total error rate which was higher only in the hundredths resulted.)

As detailed in the first column of Table 10, CART defines a respondent as having one of six combinations of the demographic variables; the remaining six combinations define a (nonrespondent + out-of-scope). Of the actual respondents, 94.08% are correctly classified by CART as respondents and 5.92% are incorrectly classified as (nonrespondents + out-of-scopes); 16.72% of actual (nonrespondents + out-of-scopes) are correctly classified and 83.28% are incorrectly classified as respondents.

Table 10. CART Classifications of Resp vs. (NR+OS), Prior Info

CART Profiles	Actual				Total
	Resp		NR+OS		
Resp		%		%	
CTZN=1 WRKPVT=2 AGEGRP=2,3	52280	74.87	17549	25.13	69829
CTZN=1 WRKPVT=2 AGEGRP=1 RACE=1,4,5	10635	64.30	5904	35.70	16539
CTZN=1 WRKPVT=1 RACE=1,4 PCLIMT=2	910	56.70	695	43.30	1605
CTZN=2 AGEGRP=2 OCCGRP=1,2,4	566	62.82	335	37.18	901
CTZN=2 AGEGRP=2 OCCGRP=3,5 EDUC=3	413	60.74	267	39.26	680
CTZN=2 AGEGRP=2 OCCGRP=3,5 EDUC=2 ORIGIN=1	688	55.44	553	44.56	1241
Total	65492	72.13	25303	27.87	90795
Classification Rate	94.08%		83.28%		90.80%
NR+OS		%		%	
CTZN=1 WRKPVT=2 AGEGRP=1 RACE=2,3	1337	48.60	1414	51.40	2751
CTZN=1 WRKPVT=1 RACE=2,3,5 PCLIMT=2	74	37.95	121	62.05	195
CTZN=1 WRKPVT=1 PCLIMT=1	171	37.83	281	59.88	452
CTZN=2 AGEGRP=2 OCCGRP=3,5 EDUC=1 ORIGIN=1	1221	48.49	1297	62.17	2518
CTZN=2 AGEGRP=2 OCCGRP=3,5 EDUC=1,2 ORIGIN=2	253	38.69	401	61.31	654
CTZN=2 AGEGRP=1,3	1067	40.51	1567	59.49	2634
Total	4123	44.80	5081	55.20	9204
Classification Rate	5.92%		16.72%		9.20%
TOTAL	69615	69.62	30384	30.38	99999

Total error rate = (4123+25303) / 99999 = 29.43%

Chi-Square Analysis

Excluding SEX, values for the chi-square test statistic range from 128.2 for MSA to 2,330.0 for CTZN and 3,976.6 for NSFGRP for the weighted data. Thus, the null hypothesis that the values for the demographic variable and response/nonresponse status are independent is rejected for all these variables

at alpha=0.10. Disregarding NSFGRP since it is a combination of several of the other variables, a listing of the variables with the top eight highest chi-square values are: CTZN, AGEGRP, PBIRTH, RACE, EDUC, ORIGIN, WRKPVT, and MOLIMT. The cells with the largest contributions to the total consistently come from the nonresponse cells. These extremely large chi-square test statistics are caused, greatly, by the large sample size. (Even the unweighted data results in extremely significant chi-square values.) The value for the chi-square test statistic for SEX is 4.4, which is still significant, but has p=0.013.

Logistic Regression

SAS regression analysis is performed using a stepwise procedure to indicate which demographic variables can be used to predict response/nonresponse. Disregarding NSFGRP, all remaining variables are significant in the model, again due to the large sample size. However, a listing of the eight 'most significant' variables in this model of response prediction is valuable: CTZN, AGEGRP, RACE, EDUC, WRKPVT, PBIRTH, OCCGRP, and ORIGIN. As noted in other analyses, although it is significant to the model, SEX is the least significant.

7.1.2 Respondents vs. Nonrespondents, Out-of-Scopes Excluded

CART

If no prior information is available and the out-of-scope records are excluded from the analysis, the lowest error rate is, again, achieved by classifying all records as respondents. Then, a true error rate of 23.85% exists. Within rounding, the resulting error rate using 99,999 records is 23.70%. This error rate is detailed in Table 11.

Table 11. CART Classifications of Resp vs. NR, No Prior Info

CART	Actual		Total
	Resp	NR	
Resp			
Total	76303 76.30%	23696 23.70%	99999
Classification Rate	100.00%	100.00%	100.00%
NR+OS			
Total	0 0%	0 0%	0
Classification Rate	0%	0%	0%
TOTAL	76303 76.30%	23696 23.70%	99999
Total error rate = (0+23696) / 99999 = 23.70%			

Knowledge of prior information does not reduce this total error rate of 23.70%. No classification tree is constructed; the best that one can do is to classify all records as respondents.

Chi-Square Analysis

Values for the chi-square test statistic range from 26.0 for SEX to 2,476.3 for AGEGRP and 2,548.9 for NSFGRP for the weighted data. Thus, the null hypothesis that the values for the demographic variable and response/nonresponse status are independent is rejected for all these variables at alpha=0.10.

Disregarding NSFGRP since it is a combination of several of the other variables, a listing of the variables with the top eight highest chi-square values are: AGEGRP, CTZN, RACE, PBIRTH, EDUC, ORIGIN, MSA, and PCLIMT. The cells with the largest contributions to the total consistently come from the nonresponse cells. These extremely large chi-square test statistics are caused, greatly, by the large sample size. (Even the unweighted data results in extremely significant chi-square values.)

7.1.3 Respondents vs. Nonrespondents vs. Out-of-Scopes

CART

If no prior information is available, the same total error rate of 30.38% depicted in Table 9 results. If prior demographic information is available, no classification tree is constructed; the best that one can do is to classify all records as respondents.

7.2 Demographic Comparison of Respondents and Nonrespondents, by Reason for Nonresponse

The methodology used to provide the results of this project is similar to that used in Section 7.1, with the exception that the demographic comparison is repeated by reason for nonresponse[10]. Initially, some exploratory data analysis relating to the NSCG sample design is presented in an attempt to report on reasons for nonresponse by sampling cell. Then, tabular data similar to that of Tables 4 through 8 follows.

The sample for the 1993 NSCG came from the SEDF, which contains data gathered from the 1990 Decennial Census long forms. There are 863 sampling cells, with information from an expanded EDUC, SEX, NSFGRP, and an expanded OCCGRP creating these cells. Of the 863 nonempty cells, only 767 cells are actually represented by the 214,643 unweighted NSCG person records. Of these, complete records cross 740 cells, and those of nonresponse cross 632 cells. There are 135 sampling cells which define only respondents (291 records); 27 cells define only nonrespondents (29 records); 22 of the 632 cells account for almost 65% of all nonresponse records. For almost all of the 632 cells, the reason for nonresponse is 'Anything Else; PMR move, no forwarding'. The values of Table 1 show that this is no surprise. Of the top ten cells with the lowest response rate, all were sampled to have a Bachelor's Degree, 5 are males and 5 are females, 9 are of 'all other occupation' while the remaining cell is science/engineering related, and they vary across being disabled, white/other, black, Foreign US citizen, and Foreign NonUS citizen. The sampling cell with the lowest response rate of 9.9% (i.e., the highest number of records with a final status of nonresponse or out-of-scope) is defined by: Bachelor's Degree, Male, White/Other, All Other Occupations. This is interesting since the final status code 'No Bachelor's Degree' has the third highest nonresponse rate.

Tables 12 through 16 give the percent of NSCG records by nonresponse reason, for each value of AGEGRP, SEX, OCCGRP, EDUC, and NSFGRP, for both unweighted and weighted data. Figures 1 through 5 depict the weighted data for readers who prefer graphs. Looking at the totals, reason 10c (PMR move, no forwarding) has the largest percentage of all nonresponse reasons; note that this reason is just a component of final status code 10 'Anything Else'. Refusals and persons with no Bachelor's Degree are next in priority. This is in agreement with the background data of Table 1, which includes the complete records. It is interesting that 'No Bachelor's Degree' is prioritized so high; persons eligible for the sample selection of the 1993 NSCG were to have received at least a Bachelor's Degree. In an attempt to validate

[10] Reasons for nonresponse include all out-of-scope and nonresponse final status codes.

this high and unexpected level of 1993 NSCG out-of-scope records, the 1995 NSCG will include a follow-up of the persons who indicated not having a Bachelor's Degree in 1993.

Table 12a. NSCG Records (%) by Nonresponse Reason
For each Value of AGEGRP and Total, Unweighted Data

Status Code	AGEGRP			Total
	1	2	3	
2	0.12	0.14	2.22	0.32
3	0.67	2.89	18.40	3.66
4	19.95	22.05	25.62	21.78
5	3.60	3.34	1.49	3.24
6	0.14	0.19	0.90	0.24
7	1.86	2.89	4.97	2.79
8	12.21	27.55	24.35	22.95
9	0.70	1.14	2.64	1.16
10	60.73	39.81	19.43	43.85
10a	0.68	0.40	0.10	0.45
10b	0	0.01	0	0.01
10c	43.95	25.97	10.40	29.61
10d	0.05	0.01	0	0.02
10e	3.77	3.08	2.05	3.18
10f	9.83	7.88	3.89	8.07
10g	1.73	2.05	2.79	2.03
10h	0.04	0.04	0	0.04
10i	0.68	0.37	0.20	0.44
Total	100.00	100.00	100.00	100.00

Table 12b. NSCG Records (%) by Nonresponse Reason
For each Value of AGEGRP and Total, Weighted Data

Status Code	AGEGRP			Total
	1	2	3	
2	0.14	0.16	2.42	0.37
3	0.61	2.76	16.22	3.43
4	21.94	23.60	28.15	23.56
5	2.77	2.37	1.06	2.36
6	0.13	0.17	0.78	0.22
7	1.88	2.77	4.50	2.68
8	13.11	30.73	26.12	25.36
9	0.72	1.18	2.57	1.18
10	58.69	36.26	18.18	40.84
10a	0.65	0.38	0.09	0.43
10b	0	0.01	0	<0.01
10c	42.26	23.17	9.38	27.22
10d	0.05	0.01	0	0.02
10e	3.74	2.75	1.95	2.96
10f	9.56	7.49	3.94	7.74
10g	1.73	2.07	2.69	2.04
10h	0.03	0.03	0	0.03
10i	0.66	0.34	0.13	0.41
Total	100.00	100.00	100.00	100.00

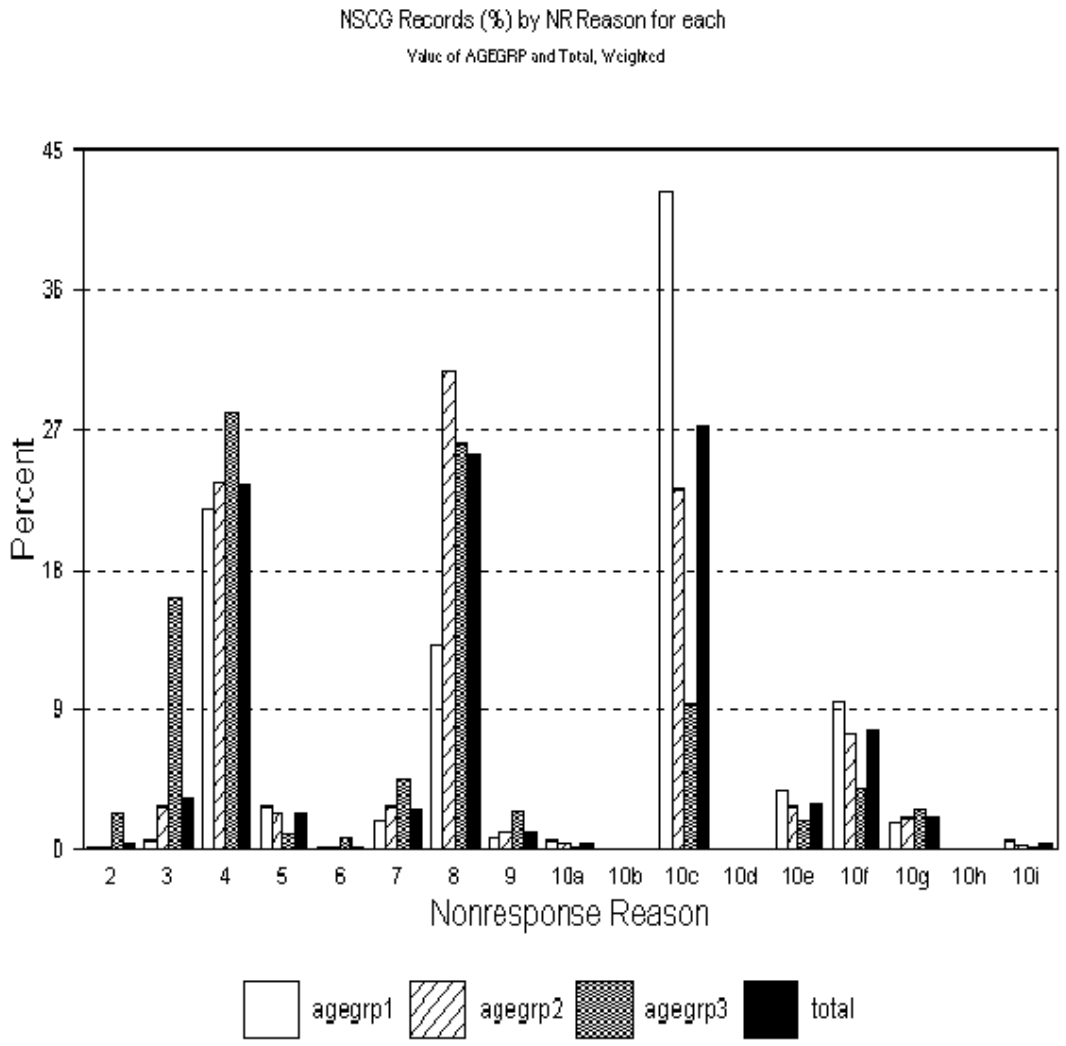
NSCG Records (%) by NR Reason for each
Value of AGEGRP and Total, Weighted
45
36
27
18
9
0
Percent
2
3
4
5
6
7
8
9
10a
10b
10c
10d
10e
10f
10g
10h
10i
Nonresponse Reason
agegrp1
agegrp2
agegrp3
total

Figure 1.

Table 13a. NSCG Records (%) by Nonresponse Reason
For each Value of SEX and Total, Unweighted Data

Status Code	SEX		Total
	1	2	
2	0.28	0.38	0.32
3	4.62	2.33	3.66
4	19.72	24.65	21.78
5	3.59	2.77	3.24
6	0.29	0.18	0.24
7	3.16	2.27	2.79
8	25.05	20.04	22.95
9	1.06	1.29	1.16
10	42.23	46.10	43.85
10a	0.47	0.43	0.45
10b	0.01	0	0.01
10c	28.28	31.47	29.61
10d	0.01	0.04	0.02
10e	3.08	3.33	3.18
10f	7.57	8.75	8.07
10g	2.34	1.59	2.03
10h	0.03	0.05	0.04
10i	0.44	0.44	0.44
Total	100.00	100.00	100.00

Table 13b. NSCG Records (%) by Nonresponse Reason
For each Value of SEX and Total, Weighted Data

Status Code	SEX		Total
	1	2	
2	0.32	0.42	0.37
3	4.50	2.13	3.43
4	20.95	26.70	23.56
5	2.56	2.12	2.36
6	0.26	0.16	0.22
7	3.18	2.09	2.68
8	28.01	22.17	25.36
9	1.01	1.38	1.18
10	39.20	42.81	40.84
10a	0.43	0.42	0.43
10b	0.01	0	<0.01
10c	25.84	28.88	27.22
10d	0.01	0.03	0.02
10e	2.78	3.16	2.96
10f	7.27	8.29	7.74
10g	2.43	1.57	2.04
10h	0.01	0.05	0.03
10i	0.42	0.40	0.41
Total	100.00	100.00	100.00

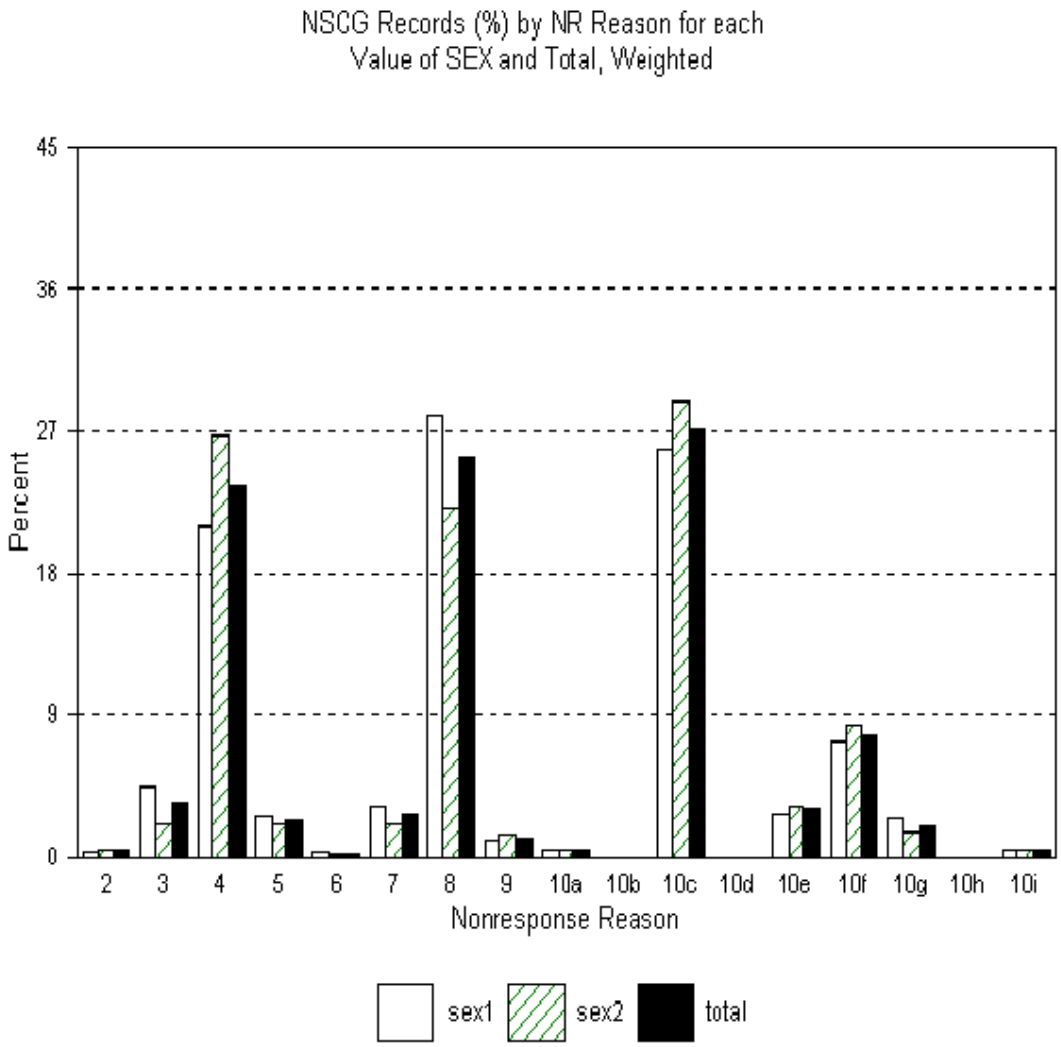
NSCG Records (%) by NR Reason for each
Value of SEX and Total, Weighted
45
36
27
18
9
0
Percent
2
3
4
5
6
7
8
9
10a
10b
10c
10d
10e
10f
10g
10h
10i
Nonresponse Reason
sex1
sex2
total

Figure 2.

Table 14a. NSCG Records (%) by Nonresponse Reason
For each Value of OCCGRP and Total, Unweighted Data

Status Code	OCCGRP					Total
	1	2	3	4	5	
2	0.22	0.08	0.24	0.15	0.36	0.32
3	3.59	2.45	2.26	4.00	3.73	3.66
4	12.24	13.88	12.31	16.28	23.39	21.78
5	4.84	3.32	3.86	3.50	3.14	3.24
6	0	0.16	0.12	0.20	0.26	0.24
7	2.56	2.57	2.73	2.82	2.81	2.79
8	22.25	27.48	27.94	29.70	21.86	22.95
9	0.76	1.03	0.54	1.03	1.21	1.16
10	53.54	49.03	50.00	42.31	43.25	43.85
10a	0.76	0.55	0.42	0.39	0.45	0.45
10b	0	0.04	0	0	0.01	0.01
10c	35.53	33.93	33.89	28.38	29.21	29.61
10d	0.05	0	0.06	0.02	0.02	0.02
10e	2.88	3.32	3.80	2.91	3.19	3.18
10f	11.10	8.94	9.57	7.57	7.93	8.07
10g	2.56	1.90	1.61	2.60	1.97	2.03
10h	0.05	0.08	0	0.07	0.03	0.04
10i	0.60	0.28	0.65	0.39	0.44	0.44
Total	100.00	100.00	100.00	100.00	100.00	100.00

Table 14b. NSCG Records (%) by Nonresponse Reason
For each Value of OCCGRP and Total, Weighted Data

Status Code	OCCGRP					Total
	1	2	3	4	5	
2	0.22	0.08	0.21	0.15	0.38	0.37
3	3.96	2.43	2.17	3.96	3.43	3.43
4	12.06	13.77	12.22	16.27	24.22	23.56
5	4.67	3.21	3.48	3.22	2.28	2.36
6	0	0.13	0.10	0.19	0.22	0.22
7	2.57	2.51	2.61	2.75	2.69	2.68
8	23.21	28.51	30.00	30.69	25.09	25.36
9	0.97	1.05	0.62	0.99	1.20	1.18
10	52.34	48.32	48.57	41.78	40.49	40.84
10a	0.67	0.58	0.33	0.38	0.42	0.43
10b	0	0.03	0	0	<0.01	<0.01
10c	34.88	33.45	33.55	27.76	29.96	27.22
10d	0.06	0	0.05	0.02	0.02	0.02
10e	2.80	3.28	3.53	2.92	2.95	2.96
10f	10.77	8.71	8.85	7.55	7.68	7.74
10g	2.51	1.93	1.54	2.64	2.02	2.04
10h	0.05	0.07	0	0.07	0.03	0.03
10i	0.60	0.27	0.73	0.43	0.41	0.41
Total	100.00	100.00	100.00	100.00	100.00	100.00

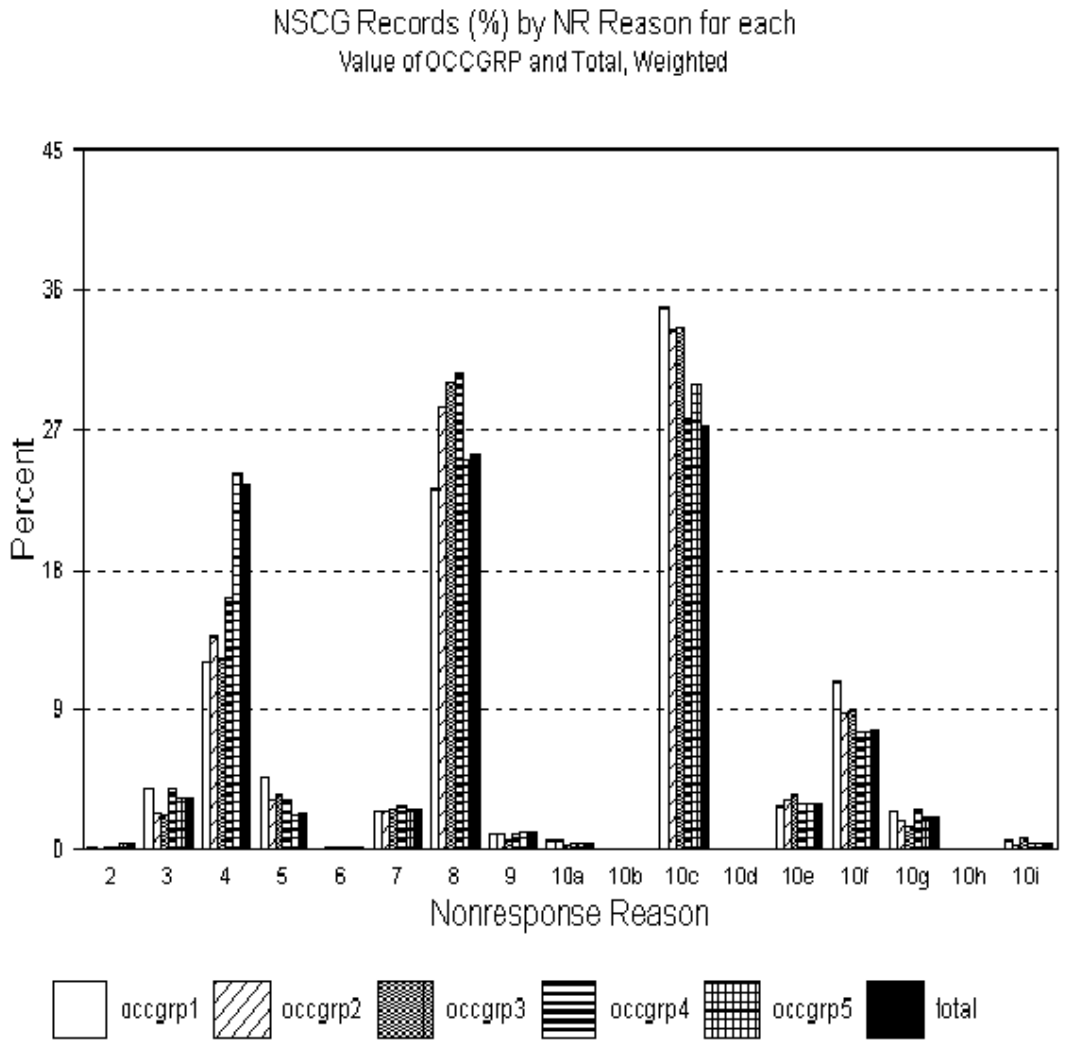
NSCG Records (%) by NR Reason for each
Value of OCCGRP and Total, Weighted
Percent
45
36
27
18
9
0
2
3
4
5
6
7
8
9
10a
10b
10c
10d
10e
10f
10g
10h
10i
Nonresponse Reason
occgrp1
occgrp2
occgrp3
occgrp4
occgrp5
total

Figure 3.

Table 15a. NSCG Records (%) by Nonresponse Reason
For each Value of EDUC and Total, Unweighted Data

Status Code	EDUC			Total
	1	2	3	
2	0.31	0.37	0.33	0.32
3	3.45	4.46	4.34	3.66
4	24.25	14.11	12.25	21.78
5	2.58	4.65	8.94	3.24
6	0.23	0.27	0.33	0.24
7	2.65	3.19	3.57	2.79
8	22.16	25.76	24.54	22.95
9	1.18	1.09	1.09	1.16
10	43.23	46.11	44.60	43.85
10a	0.45	0.46	0.47	0.45
10b	<0.01	0.02	0.04	0.01
10c	29.12	31.34	30.74	29.61
10d	0.03	0.01	0	0.02
10e	3.14	3.25	3.54	3.18
10f	8.14	8.08	6.56	8.07
10g	1.89	2.47	2.48	2.03
10h	0.04	0.05	0	0.04
10i	0.43	0.43	0.77	0.44
Total	100.00	100.00	100.00	100.00

Table 15b. NSCG Records (%) by Nonresponse Reason
For each Value of EDUC and Total, Weighted Data

Status Code	EDUC			Total
	1	2	3	
2	0.35	0.47	0.34	0.37
3	3.13	4.54	5.05	3.43
4	25.65	15.30	14.50	23.56
5	2.02	3.25	6.67	2.36
6	0.20	0.29	0.30	0.22
7	2.54	3.25	3.23	2.68
8	24.43	29.25	27.99	25.36
9	1.20	1.05	1.25	1.18
10	40.47	42.59	40.67	40.84
10a	0.43	0.42	0.41	0.43
10b	<0.01	0.01	0.02	<0.01
10c	26.97	28.36	27.41	27.22
10d	0.02	<0.01	0	0.02
10e	2.91	3.14	3.10	2.96
10f	7.76	7.85	6.45	7.74
10g	1.94	2.40	2.64	2.04
10h	0.03	0.04	0	0.03
10i	0.41	0.37	0.63	0.41
Total	100.00	100.00	100.00	100.00

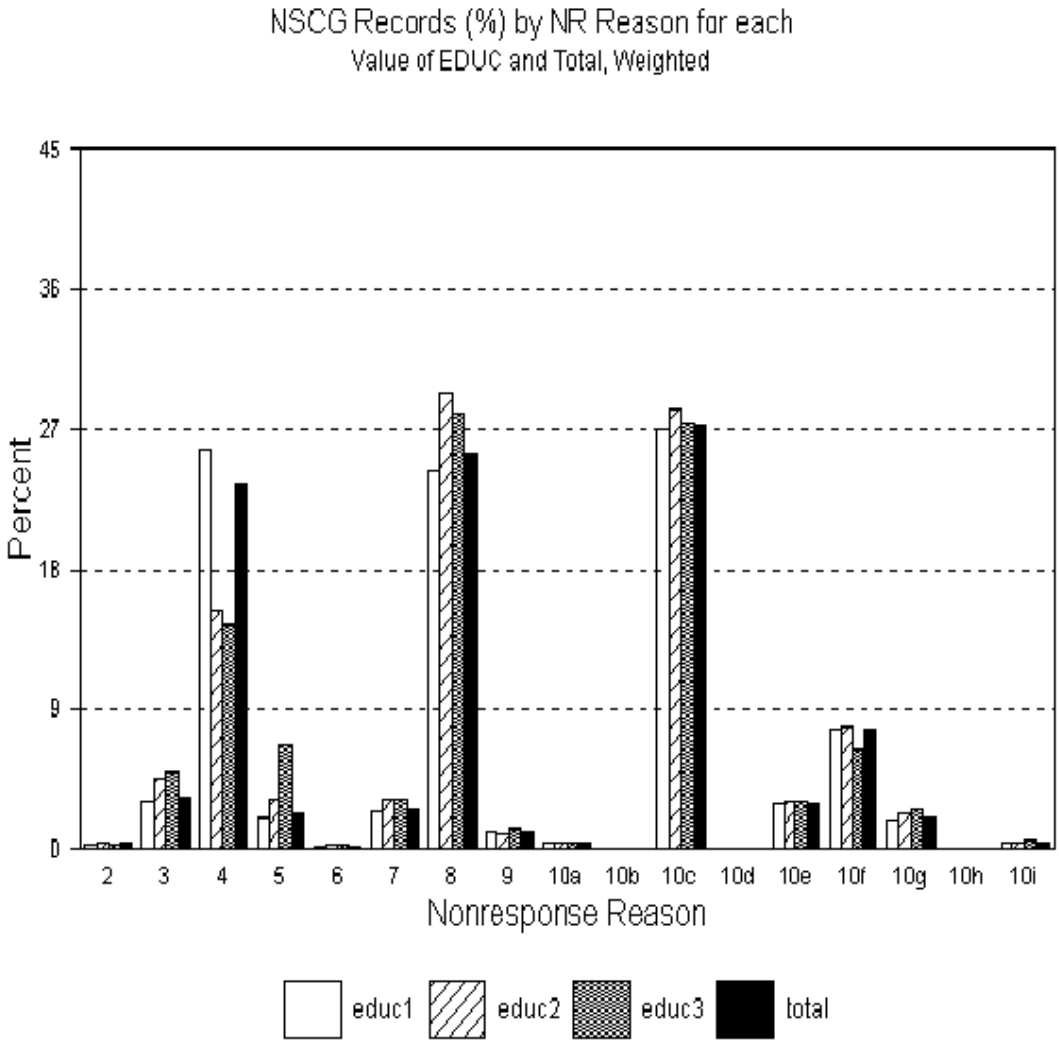
NSCG Records (%) by NR Reason for each
Value of EDUC and Total, Weighted
Percent
45
36
27
18
9
0
2
3
4
5
6
7
8
9
10a
10b
10c
10d
10e
10f
10g
10h
10i
Nonresponse Reason
educ1
educ2
educ3
total

Figure 4.

Table 16a. NSCG Records (%) by Nonresponse Reason
For each Value of NSFGRP and Total, Unweighted Data

Status Code	NSFGRP								Total
	1	2	3	4	5	6	7	8	
2	0.95	0.09	0.28	0.33	0	0.15	0.31	0.18	0.32
3	15.30	1.57	2.98	2.11	2.12	1.91	2.95	0.89	3.66
4	25.06	27.01	20.27	22.68	23.08	26.58	24.15	19.38	21.78
5	0.67	2.10	1.61	0.53	2.77	0.59	3.94	11.46	3.24
6	1.26	0.18	0.13	0.20	0	0.44	0.11	0.09	0.24
7	4.09	2.10	2.57	2.24	2.20	2.20	3.94	2.42	2.79
8	20.47	14.39	30.15	17.90	21.53	17.33	22.75	11.05	22.95
9	1.36	0.89	1.02	0.98	1.39	1.17	1.58	1.28	1.16
10	30.84	51.68	40.99	53.03	46.90	49.63	40.26	53.26	43.85
10a	0.31	0.56	0.45	0.51	0.73	0.29	0.37	0.52	0.45
10b	0	0	0	0	0	0	0.04	0.01	0.01
10c	19.73	36.50	26.95	34.77	31.48	28.78	26.30	39.95	29.61
10d	0	0.06	0.02	0.03	0.16	0.15	0	0.01	0.02
10e	2.15	3.43	2.84	4.03	3.59	2.94	3.40	3.92	3.18
10f	6.21	9.31	8.02	11.32	8.16	15.27	7.36	6.74	8.07
10g	2.05	1.39	2.25	1.80	2.28	1.76	2.33	1.50	2.03
10h	0.02	0.15	0.02	0.04	0.08	0	0.06	0.02	0.04
10i	0.36	0.30	0.42	0.53	0.41	0.44	0.41	0.58	0.44
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 16b. NSCG Records (%) by Nonresponse Reason
For each Value of NSFGRP and Total, Weighted Data

Status Code	NSFGRP								Total
	1	2	3	4	5	6	7	8	
2	1.07	0.09	0.34	0.35	0	0.14	0.33	0.20	0.37
3	15.30	1.52	2.90	2.23	2.26	1.88	2.81	0.92	3.43
4	26.35	27.46	22.96	23.98	24.36	26.11	26.78	22.19	23.56
5	0.63	2.16	1.41	0.52	2.78	0.58	3.86	10.95	2.36
6	1.28	0.17	0.14	0.23	0	0.43	0.13	0.09	0.22
7	4.12	2.11	2.55	2.19	2.22	2.19	3.83	2.39	2.68
8	20.00	14.26	29.47	17.81	21.15	17.88	21.18	10.70	25.36
9	1.36	0.86	1.12	0.91	1.37	1.15	1.64	1.39	1.18
10	29.89	51.38	39.10	51.78	45.87	49.63	39.45	51.19	40.84
10a	0.33	0.49	0.43	0.46	0.76	0.29	0.29	0.50	0.43
10b	0	0	0	0	0	0	0.03	0.01	<0.01
10c	18.99	36.38	25.62	33.80	30.62	29.09	26.06	38.36	27.22
10d	0	0.05	0.02	0.03	0.19	0.14	0	0.01	0.02
10e	2.13	3.41	2.74	4.02	3.51	2.88	3.44	3.79	2.96
10f	6.13	9.24	7.72	11.04	7.98	15.03	7.04	6.48	7.74
10g	1.93	1.40	2.16	1.84	2.31	1.76	2.10	1.47	2.04
10h	0.01	0.12	0.02	0.03	0.06	0	0.07	0.01	0.03
10i	0.38	0.29	0.39	0.55	0.43	0.44	0.42	0.56	0.41
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

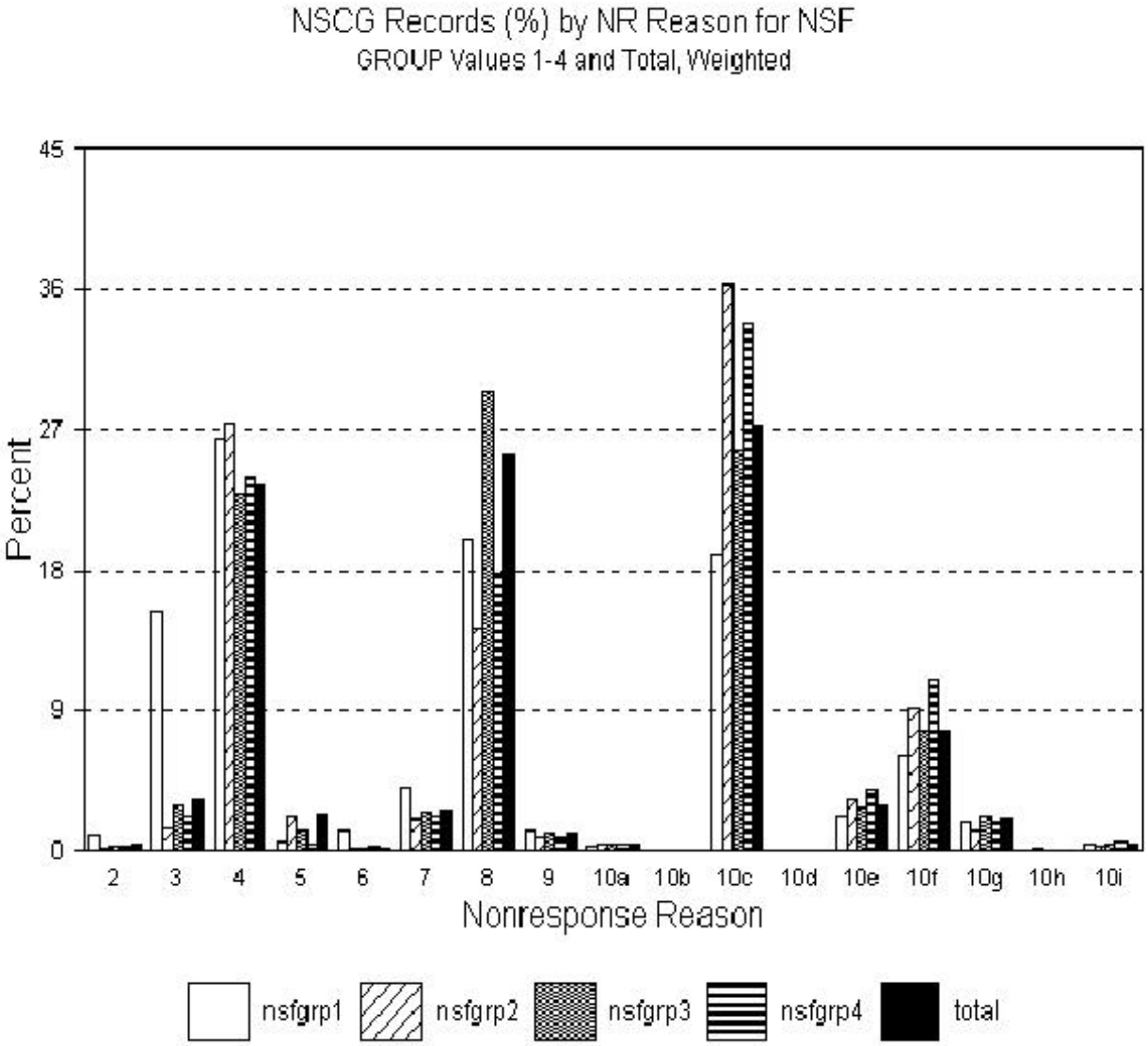
NSCG Records (%) by NR Reason for NSF
GROUP Values 1-4 and Total, Weighted
45
36
27
18
9
0
Percent
2
3
4
5
6
7
8
9
10a
10b
10c
10d
10e
10f
10g
10h
10i
Nonresponse Reason
nsfgrp1
nsfgrp2
nsfgrp3
nsfgrp4
total

Figure 5.1.

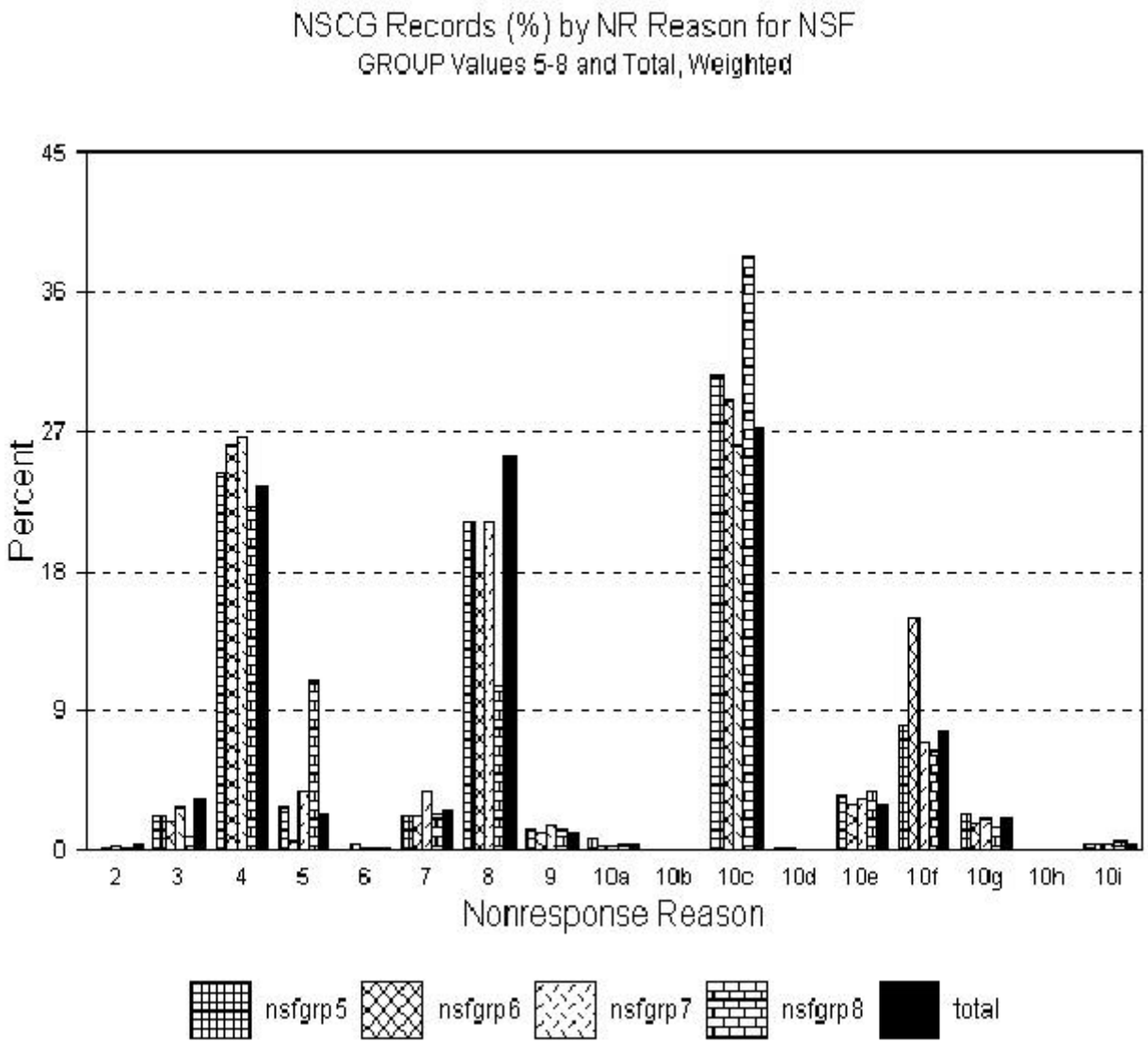

Figure 5.2.

CART

Various CART analyses are performed to try to discover one or more demographic characteristics that are consistently correlated with nonresponse. Because of very few records for many of the nonresponse categories (Table 1), only the top 9, in percent occurrence, of the nonresponse reasons are used: Anything Else (and its subcomponents of PMR move, no forwarding; Not located; and PMR temporarily absent); Refusal; No Bachelor's Degree; Deceased; Emigrant; and Ill. For each of these nonresponse reasons, a first result of CART is that improvement in the total error rate is no greater than 0.67% when a three-way classification is made across respondents, the nonresponse reason, and all other nonresponse reasons. The variables of CTZN, AGEGRP, RACE, OCCGRP, EDUC, PBIRTH, and ORIGIN enter into the classifications in varying degrees. (Note that these are the same variables that we have seen in results throughout this paper.) Second, when the respondents are classified against each nonresponse reason, prior demographic information is unnecessary since no classification tree is created. Regardless of the reason for nonresponse then, one can, simplistically, do no better than to designate all records as respondents. Third, the respondents are removed from the data set and each of the seven nonresponse

reasons are classified against all others. Prior demographic information is again unnecessary. Lastly, a classification across all nonresponse reasons is done. The operational implications are vague, if any, but the result is of interest. Since the nonresponse reason of "PMR move, no forwarding" is the largest, CART analysis indicates that all nonrespondents should be classified as having this reason, with a 7.63% improvement in total error rate and with all fourteen demographic variables entering into the classification.

Chi-Square Analysis

When a table is constructed with the top 9 reasons for nonresponse as the columns and the fourteen demographic variables as the rows, there are 112 cells for chi-square values, all testing the null hypothesis of independence. All but 6 cells are significant[11]. Again, these extremely large chi-square test statistics are caused, greatly, by the large sample size.

Not wanting to provide a prioritized list of the highest chi-square values for each nonresponse reason, let it be noted that those for AGEGRP, OCCGRP, CTZN, EDUC, ORIGIN, PBIRTH and RACE are consistently high across all reasons.

CART Analysis Comparing Refusals and Noncontacts

Perhaps complementary to current research by Groves and Couper in "Theoretical Motivation for Post-Survey Nonresponse Adjustment in Household Surveys," this additional section includes a CART classification analysis that attempts to distinguish refusals from noncontacts. That is, records with status code 8 (Refusal) are classified against records having status codes 7 (Ill), 9 (Incomplete) and 10 (Anything Else). Several iterations are performed: including and excluding the variable NSFGRP, including and excluding from the noncontact class those records having the status code of Incomplete, and performing additional classifications against the respondent class. Regardless, the improvement in error when prior demographic information is known is, at most, 1.33%.

7.3 Differences in Results Across Data Collection Modes

It is possible to compare respondents and nonrespondents across the three data collection modes of mail, CATI, and personal visit since the NSCG data set provides the 'intermediate' status codes for each record, after each data collection effort. (Hereafter, these intermediate status codes are designated outcome codes). Provided as background, the percent of NSCG person records undergoing each type (and combination) of collection mode is given in Table 17.

[11] Values for these six chi-square statistics are: WRKPVT vs. Anything Else=2.03; WRKPVT vs. PMR move, no forwarding=0.60; WRKPVT vs. Temporarily absent=1.27; WRKLIMT vs. Not located=0.20; WRKLIMT vs. Temporarily absent=2.14; PCLIMT vs. Emigrant=0.22.

Table 17a. Distribution by Data Collection Mode, Unweighted Data

Data Collection Mode	Records	Percentage
Mail only	117531	54.76
CATI only	3709	1.73
Personal Visit only	855	0.40
Mail and CATI only	39374	18.34
Mail and Personal Visit only	16704	7.78
CATI and Personal Visit only	1686	0.79
Mail and CATI and Personal Visit	34784	16.21
Total	214643	100.00

Table 17b. Distribution by Data Collection Mode, Weighted Data (000s)

Data Collection Mode	Records	Percentage
Mail only	18371	57.75
CATI only	288	0.90
Personal Visit only	66	0.21
Mail and CATI only	6044	19.00
Mail and Personal Visit only	2224	6.99
CATI and Personal Visit only	126	0.40
Mail and CATI and Personal Visit	4692	14.75
Total	31810	100.00

Various analyses (respondents vs. nonrespondents, with and without the out-of-scopes) are then conducted on five data sets which differ in terms of which outcome code is used as the final status code. It is assumed that the progression of data collection is mail, CATI and personal visit. The first data set consists of only those records which have a mail outcome code, regardless of the presence of CATI and personal visit outcome codes. The second data set consists of only those records which have a CATI outcome code, regardless of the presence of mail and personal visit outcome codes. Likewise, the third consists of those records having an outcome code resulting from a personal visit. Addressing just CATI, there are 135,090 unweighted records which never underwent this data collection method; therefore, the fourth data set substitutes the mail outcome code, if present, for these records. Addressing just personal visits, the fifth data set first substitutes any present CATI outcome code for records with missing personal visit outcome code, and then substitutes the mail outcomes code for those records which did not undergo

either CATI or personal visit.[12] Table 18 gives the sizes of these various data sets. Also, this table provides a breakout of the records by response, out-of-scope, and nonresponse.

Table 18a. NSCG Records by Outcome, by Mode of Data Collection, Unweighted Data

Data Set	Resp.	OS	NR	Total	Response Rate (%)
Mail	111432	8026	88935	208393	57.32
CATI	22873	4866	51814	79553	34.87
Personal Visit	17228	6141	30660	54029	43.35
CATI, w/Mail repl.	131300	12693	69795	213788	67.35
PV, w/CATI & Mail repl.	148470	18821	47352	214643	77.94

Table 18b. NSCG Records by Outcome, by Mode of Data Collection, Weighted Data (000s)

Data Set	Resp.	OS	NR	Total	Response Rate (%)
Mail	17403	1242	12685	31330	59.51
CATI	3281	694	7174	11149	35.65
Personal Visit	2312	790	4005	7107	43.65
CATI, w/Mail repl.	20241	1908	9594	31744	69.77
PV, w/CATI & Mail repl.	22547	2696	6566	31810	79.36[13]

It was hoped that results from the following analyses would help answer questions of the following type: Is it worth doing only one or two interview modes for various subsets of the sample? For example, perhaps mail could be eliminated for some 'profiles' that exhibit strong dependence with nonresponse, and the data could then be collected initially via CATI or personal visit. Cost data is vital in this analysis and will be discussed at the end of this section.

[12] The fifth data set should be identical to the basic data set used in the "Demographic Comparison of Respondents and Nonrespondents" section. However, although the progression of data collection is mail, CATI, and personal visit, there are instances when, operationally, outcome codes of prior efforts are selected as the final status code over the outcome code of subsequent efforts.

[13] The response rate for 'PV w/CATI & Mail repl.' is not equal to the final response rate for the entire data set (79.77% weighted) since there are instances when outcome codes of prior efforts are selected as the final status code over the outcome code of subsequent efforts. Also, for the data set, there were over 1,000 records where the final status code was changed, without editing of the mode outcome code.

CART

Table 19 reveals that results via ten CART runs for this project are more noticeable than the other two projects. When looking at mail vs. CATI vs. personal visit, mail is the only mode which reveals any possibility for error improvement when information on various demographics is available. It exhibits error improvements of 8.20% and 6.64%, when out-of-scopes are included and excluded from nonrespondents, respectively. However, the profiles of these CART respondents and nonrespondents, it is felt, are too cumbersome to present in this research, let alone incorporate operationally. Data from all fourteen demographic characteristics are involved in at least one of the classifications.

Table 19. Results of CART Classifications, by Mode of Data Collection (%)

Data Set	No Prior Info Error Rate	CART Error Rate	Error Improvement
Mail			
1. Resp vs. (NR+OS)	46.17	37.97	8.20
2. Resp vs. NR; OS excluded	44.15	37.51	6.64
CATI			
3. Resp vs. (NR+OS)	28.69	28.69	0
4. Resp vs. NR; OS excluded	30.57	30.57	0
Personal Visit			
5. Resp vs. (NR+OS)	31.91	31.91	0
6. Resp vs. NR; OS excluded	36.03	36.03	0
CATI, w/Mail replacements			
7. Resp vs. (NR+OS)	37.38	34.29	3.09
8. Resp vs. NR; OS excluded	33.58	31.80	1.78
Personal Visit, w/CATI & Mail repl.			
9. Resp vs. (NR+OS)	30.56	29.83	0.73
10. Resp vs. NR; OS excluded	24.15	24.15	0

For the other three CART analyses which exhibit a non-zero error improvement, all nonrespondents should be classified as respondents if no prior information is available, However, prior information could decrease the error rates, ranging from 0.73% to 3.09%. At the lower extreme, the profiles of the respondents and nonrespondents for run 9 requires data from OCCGRP, EDUC, ORIGIN, PBIRTH, AGEGRP and NSFGRP; the number of demographics increases for the other two.

No classification tree is constructed for five of the ten CART analyses. For these, data on none of the fourteen variables is beneficial and the best that one can do is to classify all nonresponse records to whichever outcome (response or nonresponse) occurs more frequently.

Chi-Square Analysis

Examining only the mail outcome codes, all fourteen demographic variables have significant values for the chi-square statistic, both when the out-of-scope records are included and excluded from the nonresponse class. Again disregarding NSFGRP (its value is 7,806 when the out-of-scopes are included),

a prioritized listing of the variables with the top eight highest chi-square values are: AGEGRP, RACE, CTZN, PBIRTH, EDUC, OCCGRP, ORIGIN, and PCLIMT.

Examining only the CATI outcome codes, all demographic variables except WRKLIMT[14] have significant values for the chi-square statistic, both when the out-of-scope records are included and excluded from the nonresponse class. Again disregarding NSFGRP (its value is 1,334 when the out-of-scopes are included), a prioritized listing of the variables with the top eight highest chi-square values are: OCCGRP, RACE, PBIRTH, CTZN, ORIGIN, AGEGRP, SEX, and EDUC.

Examining only the outcome codes from personal visit, all demographic variables except SEX[15] have significant values for the chi-square statistic, both when the out-of-scope records are included and excluded from the nonresponse class. Again disregarding NSFGRP (its value is 689 when the out-of-scopes are included), a prioritized listing of the variables with the top eight highest chi-square values are: CTZN, AGEGRP, PBIRTH, OCCGRP, RACE, MSA, EDUC, and WRKPVT. Across all three modes, these extremely large chi-square test statistics are caused, greatly, by the large sample size.

Cost Discussion[16]

A complementary and vital issue in this discussion is the tie-in of costs across the three data collection modes. The cost per case[17] is $9.18 for mail, $10.69 for CATI, and $93.45 for personal visit. Knowledge of cost and response rates for each of the modes could result in some desirable future tradeoffs. In order to do a (separate) study of mode effects, 6,250 records did not undergo the mode of mail; their data collection began with CATI, and proceeded to personal visit, if necessary. Their presence provided a rare opportunity to study the effects of mode on response rates in this report. The most noticeable demographic differences between these 'no mail' cases and the 'mail' cases is the exclusion of EDUC=3 (Doctorate) persons and their 20% split across all OCCGRP groups, whereas 80% of the 'mail' group has OCCGRP=5 (other). The following table presents the weighted response rates, by mode outcome, for the 'no mail' records (6,250 unweighted records), the 'mail' records (208,393 unweighted records) and the total dataset (214,643 unweighted records). (As before, the response rate for the entire data set does not equal that for 'PV w/CATI & Mail repl.' since there are instances when previous outcome codes, rather than subsequent codes, are used as the final status code.) There is a 15% improvement in response rate for those records that do undergo mail. When the very similar costs per case for mail and CATI are also considered, and one compares the response rate by mode for the two groups, some very important

[14] When the out-of-scope records are excluded from nonresponse, the observed chi-square value for WRKLIMT is 2.53, as compared to the critical value of 3.79, alpha=0.10, 1 degree of freedom.

[15] When the out-of-scope records are excluded from nonresponse, the observed chi-square value for SEX is 0.01, as compared to the critical value of 3.79 for alpha=0.10, 1 degree of freedom; when these records are included, the value is 0.57.

[16] Because of timing, only an initial and general presentation of results are provided in this section; because of the possible implications, further and more detailed results on this topic should be forthcoming.

[17] These costs were approximated by the Decennial Support Division of the Bureau of the Census. They include, where appropriate, postage, printing, data collection and keying. They exclude certain Bureau salaries, computer processing, and sampling related activities (e.g., locating and keying of decennial census records.)

implications can be surmised. However, several clarifications need to be stated as to the quality of data collection between the 'mail' and 'no mail' groups. These should be forthcoming.

Table 20. Response Rates (%) of Records that undergo Mail and those that do not, Weighted Data

Data Collection Mode	'No Mail'	'Mail'	Total
Mail	-	59.51	59.51
CATI	45.77	34.99	35.66
Personal Visit	50.10	43.47	43.65
CATI w/Mail repl.	45.77	70.00	69.78
PV w/CATI & Mail repl.	65.81	79.57	79.36
Final Status	64.94	80.00	79.77

8. Final Remarks

Simplistically, the goal of this research was to provide an interpretable picture of a structure for the 1993 NSCG data and to determine if any of fourteen 1990 Decennial Census demographic variables could reliably distinguish the survey's respondents from their nonrespondents. The relationships between each demographic variable and response/nonresponse for the whole dataset, between each variable and reason for nonresponse for the nonrespondents, and between each variable and response/nonresponse for the whole dataset by mode of data collection were examined. This was achieved through an extensive amount of background and exploratory data analysis which focused on response rates, CART classification analysis, chi-square analysis, and regression. Although results could not provide 'worthwhile' characterizations of the conditions that determine when a sample person is a respondent rather than a nonrespondent (e.g., a maximum CART error improvement rate of 0.95%), a few of the insights into the data need to be repeated in this final section.

First, the results via chi-square testing and regression do not contradict the results of CART. CART is more 'strenuous' in that it associates response/nonresponse to various demographics depending on where a majority of the records fall. Second, the variables of AGEGRP, RACE, ORIGIN, PBIRTH, CTZN, EDUC, and OCCGRP appear over and over when overall contributions to response are discussed, *especially CTZN*. This ties in beautifully with the lower response rate of 64.89% for NSFGRP=8 (Foreign Born, NonUS Citizen). Conversely, SEX and MSA are at the 'bottom of the totem pole of significance'. Third, in terms of reasons for nonresponse, the implications of the large majority of records having the nonresponse reason of "PMR move, no forwarding" should be explored for the goal of nonresponse reduction. Lastly, results from a comparison across data collection mode hold considerable promise. Although the analysis must be expanded further, the presentation of costs and response rates by modes indicate that there may be groups of the population where a mode is not worth the effort. Although the costs for mail and CATI are approximately the same, the final response rate without mail is 15% lower than that when records progress through mail, CATI and personal visit.

A cautionary note: Previous recommendations from this research indicated that a sample of nonrespondents for future NSCG surveys not be made; it would prove more beneficial to spend survey money on respondents than on the follow-up of nonrespondents. However, after reviewing comments

received at the American Statistical Association meetings, this conclusion is not accepted; nonresponse needs to be evaluated over time and there should be a followup on each cycle's nonrespondents.

9. References and Supporting Materials

Ash, S., Kraus, M., and Peterson, A. (1995), "Evaluation of Classification and Regression Tree Generalized Model Groups for the 1992 Census of Agriculture," Bureau of the Census, Washington, D.C.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification And Regression Trees*, Belmont, California: Wadsworth International Group.

California Statistical Software, Inc. (1985), "An Introduction to CART Methodology."

California Statistical Software, Inc. (1993), "Using the CART Programs, Version 1.3."

DeMaio, T. J. (1980), "Refusals: Who, Where, and Why?," *Public Opinion Quarterly*, pp. 223-233.

Groves, R. M. (1989), *Survey Errors and Survey Costs*, New York: John Wiley & Sons.

Groves, R. M. and Couper, M. P., "Theoretical Motivation for Post-Survey Nonresponse Adjustment in Household Surveys."

Kruytbosch, C. (1994), written correspondence between Kruytbosch, NSF Program Director, Personnel Programs, and Ron Dopkowski, Chief, Consumer Expenditures Branch, Bureau of the Census, April 11, 1994.

Mendenhall, W. and Scheaffer, R. L. (1973), *Mathematical Statistics with Applications*, North Scituate, Massachusetts: Duxbury Press.

U.S. Bureau of the Census (1993), "Sample Selection Specifications for the 1993 National Survey of College Graduates (NSCG) - Revised," memorandum from Preston Jay Waite to Thomas C. Walsh, June 15, 1993.

Various phone discussions with Charles Stone (CART developer) and CART software programmers regarding the drawbacks/restrictions of the learning sample. California Statistical Software, Inc.: 510-283-3392.