# QUALITY OF THE INITIAL NAMES AND ADDRESSES IN THE 1993 NATIONAL SURVEY OF COLLEGE GRADUATES

**Lloyd Hicks and Thomas F. Moore**

This section contains two reports on the quality of the initial names and addresses. Some records in the initial sample could not be mailed or could not be located because of problems with names or addresses. The first report on bad addresses discovered during the Metronet operation is based on the May 4, 1993, Memorandum for the record by Lee H. Giesbrecht. The second report compares demographic characteristics of the bad name cases with the final NSCG sample.

## Address Check

Table 1 shows combinations of the U.S. Postal Service's National Change of Address (NCOA) Return Code and MetroNet's proprietary Change of Address (COA) Return Code cross-tabulated with the presence or absence of a telephone number for the 240,579 NSCG records we sent to MetroNet. We did the Address Search operation on the entire initial sample of about 241,000 cases in February 1993, before reducing the sample to the target size.

To understand the table, it is important to understand the basics of how MetroNet's Address Search works. MetroNet first compares all the addresses (not names) in our file with the addresses in their NCOA file. If an address matches the NCOA file, the NCOA address is returned to us. These types of cases are shown in the second row of the Table 1. If our address matches the NCOA file and a new address is returned to us, no further searching is performed. However, if there is no NCOA match (Row 1 on table), the NCOA match is a NIXIE (Row 3 on table) or the NCOA match is Moved, Left No Forwarding Address (Row 4 on table), MetroNet attempts to match that address against their own COA file. Sixty-nine percent (165,591) of the cases we sent had no NCOA or COA match. We expected these cases to have good addresses.

NCOA NIXIE is a near, but not exact, match. Because the Postal Service has strict rules about giving change of address information, they will not release the information unless the old address we supply is an exact match.

We got telephone numbers for about 46 percent of the cases. Note that there are 22,425 NCOA Nixie (near match) cases for which MetroNet did not have a new address. And 1,253 cases that the NCOA listed as having moved without leaving a forwarding address, for which MetroNet did not have a new address. These 23,678 cases are 9.8 percent of the total 240,579 cases. This means we should have expected 9.8 percent of the sample cases to be ones likely to be difficult to locate. In fact, 9.1 percent of the sample had a final outcome code of Moved, Unable to Locate.

Table 1: National Survey Of College Graduates Crosstab of Change of Address and Presence of Telephone Number

| NCOA Return Code | COA Return Code | With Phone # | Without Phone # | Total |
|---|---|---|---|---|
| 02 Not Found | 02 Not Found | 88,140 | 77,451 | 165,591 |
| | 03 New Address Returned | 1,901 | 2,760 | 4,661 |
| 06 New Address Returned | 07 Search Not Performed | 12,412 | 24,107 | 36,519 |
| 07 NIXIE Found | 02 Not Found | 4,553 | 17,872 | 22,425 |
| | 03 New Address Returned | 4,613 | 5,506 | 10,119 |
| 08 Moved, Left No Forwarding Address | 02 Not Found | 0 | 1,253 | 1,253 |
| | 03 New Address Returned | 4 | 4 | 11 |
| Total | | 111,623 | 128,956 | 240,579 |

## Analysis of Bad Name Cases

The NSCG is a longitudinal survey that spans an entire decade. High response is critical to ensure data remain credible during the duration of the survey, especially for the data collected at the end of the decade. It is important to minimize nonresponse for the initial survey, because it is where the highest nonresponse generally occurs. A large part of this can be attributed to the age of the addresses. The initial sample is drawn from the respondents to the census long form. By the time the initial survey for the 1990 design went into the field in 1993, the addresses were three years old.

Another potential cause of high nonresponse is the bad quality of names and addresses obtained from the census files. The names and addresses may be partially or entirely missing or unreadable. As a result these cases can not be interviewed because there is insufficient information to contact them. This paper looks at these cases for the 1993 survey.

The 1993 NSCG sample was selected in two phases in an attempt to minimize nonresponse due to the bad quality of names and addresses in a cost-effective manner. In the initial phase about 245,000 persons were selected for sample. Cases from this sample deemed unmailable, those with bad names and/or addresses, were removed. The remaining sample was then subsampled to the desired sample size of about 214,600 persons. We minimized nonresponse by dropping sample cases we knew we would never interview because of insufficient information to contact them. We saved costs by checking names and addresses on a subset of the universe instead of the whole universe. There were about 5,260 unmailable cases, about 2.1 percent of the 245,000 initial sample persons. About 4,300 of these had bad or missing names, and about 960 had good names but bad or missing addresses.

Because the number of unmailable cases was relatively low, it should not have any impact on the NSCG estimates unless a disproportionately large number of cases came from a small subdomain of the population. If this occurred, the underrepresented subpopulation could incur a bias too large to be effectively reduced by weighting procedures.

In this paper, we look at several demographic variables to determine if there exist any subdomains that experienced a disproportionately large number of unmailable cases. We compare the distribution of the bad name cases to that of the final NSCG sample for five variables identified as important to the survey: highest degree, occupation, age, sex, and NSF group which includes race, ethnicity, disability status, citizenship status, and place of birth. (We used only the bad name cases as opposed to both bad name and address cases because the bad address cases were unavailable at the time of this analysis, and we lacked the resources to get them.)

We created five tables, one for each variable, comparing the distributions of the bad name cases and the 1993 NSCG sample. For each category, the tables show the difference between the two groups and whether the difference is statistically significant at the 10 percent significance level, a bureau standard. In conducting the t-tests, we computed the variances assuming simple random. Differences that are NOT statistically significant are marked with an asterisk (*). Those that are significant are not marked.

In addition, we performed chi-square tests on each pair of distributions. The chi-square test (or goodness of fit test) determines whether the distributions of bad name cases and the 1993 NSCG sample are statistically different. The chi-square test statistic ($\chi^2$) is shown at the bottom of the difference column in each table. The chi-square score ($\chi^2_{.05}$) that it's tested against is given in parentheses beneath the chi-square statistic. The chi-square test is conducted at the 0.05 percent significance level, and the degrees of freedom is equal to the number of cells in both distributions combined minus one. If the chi-square test statistic is larger than the chi-square score than the distributions are considered statistically different.

The last column in each table shows the distribution of the bad names and 1993 NSCG sample conbined. It gives an indication of what the 1993 NSCG sample distribution would have been if there had been no bad or missing names.

Table 2: Distributions of the Bad Name Cases and 1993 NSCG Sample
Highest Degree

| Highest Degree | Bad Names | 1993 NSCG Sample | Difference | Combined Sample |
|---|---|---|---|---|
| Bachelors or Professional | 73.8% | 71.1% | 2.7% | 71.1% |
| Masters | 22.3% | 23.9% | -1.6% | 23.9% |
| Doctorate | 3.9% | 5.0% | -1.1% | 5.0% |
| | | | $\chi^2 = 20.1$ (11.1) | |

Table 3: Distributions of the Bad Name Cases and 1993 NSCG Sample
Occupation Group

| Occupation Group | Bad Names | 1993 NSCG Sample | Difference | Combined Sample |
|---|---|---|---|---|
| Physical/Life Science | 3.6% | 3.3% | 0.3% * | 3.3% |
| Math/Computer Science | 2.5% | 4.6% | -2.1% | 4.5% |
| Social Science | 2.7% | 2.7% | -0.1% * | 2.7% |
| Engineers | 7.4% | 11.6% | -4.2% | 11.5% |
| Other | 83.8% | 77.8% | 6.0% | 77.9% |
| | | | $\chi^2 = 129.3$ (16.9) | |

* The difference is not significant at the 10% significance level.

Table 4: Distributions of the Bad Name Cases and 1993 NSCG Sample
Age

| Age Group | Bad Names | 1993 NSCG Sample | Difference | Combined Sample |
|---|---|---|---|---|
| 16-29 | 24.3% | 21.1% | 3.3% | 21.1% |
| 30-59 | 64.0% | 69.0% | -5.0% | 69.0% |
| 60+ | 11.7% | 9.9% | 1.8% | 9.9% |
| | | | $\chi^2 = 51.8$ (11.1) | |

* The difference is not significant at the 10% significance level.

Table 5: Distributions of the Bad Name Cases and 1993 NSCG Sample
Sex

| Sex | Bad Names | 1993 NSCG Sample | Difference | Combined Sample |
|---|---|---|---|---|
| Male | 59.9% | 58.7% | 1.2% * | 58.7% |
| Female | 40.1% | 41.3% | -1.2% * | 41.3% |
| | | | $\chi^2 = 2.7$ (7.8) | |

* The difference is not significant at the 10% significance level.

Table 6: Distributions of the Bad Name Cases and 1993 NSCG Sample
NSF Group

| NSF Group | Bad Names | 1993 NSCG Sample | Difference | Combined Sample |
|---|---|---|---|---|
| Disabled | 11.5% | 8.1% | 3.4% | 8.2% |
| Hispanic | 3.7% | 4.3% | -0.6% | 4.3% |
| White/Other | 57.4% | 55.7% | 1.7% | 55.7% |
| Black | 8.7% | 8.3% | 0.4% * | 8.3% |
| API | 3.0% | 2.2% | 0.7% | 2.2% |
| Native Americans | 1.0% | 0.9% | 0.1% * | 0.9% |
| Foreign-Born US Citizens | 6.9% | 11.1% | -4.2% | 11.0% |
| Foreign-Born Non-US Citizens | 7.9% | 9.4% | -1.5% | 9.3% |
| | | | $\chi^2 = 164.3$ (25.0) | |

* The difference is not significant at the 10% significance level.

**Analysis**

The results of the chi-square tests show that the distributions of the bad name cases are statistically different from the distributions of the 1993 NSCG final sample for each variable except sex. The chi-square test statistic for the sex variable is 2.7. The chi-square test statistics for the other variables range from 20.1 for highest degree to 164.3 for NSF group. Because the sample size of the bad name cases is relatively large (about 4,300), it is not surprising that values for the chi-square statistic are large.

In addition, we compared percentage estimates of each category between the bad name cases and the NSCG sample for each variable. These comparisons show where the distributions differ the most. Almost all of the differences between the two distributions were statistically significant. These differences can probably be attributed to the large sample sizes of the two groups, in particular the NSCG sample which was about 214,600 cases. The relevant question here is whether these differences are analytically important.

As mentioned earlier, small subdomains that experience a disproportionately large number of bad cases will be affected the most. A positive value in the difference columns in Tables 2 - 6 above indicate that the estimate from the bad cases is larger than estimate from the 1993 NSCG sample. The largest positive difference in the five tables was the "other" category for the occupation group variable in Table 3. The difference was 6.0 percent; but because the percentage of the 1993 NSCG sample cases in this category is so large (77.8 percent), this difference should have no effect on the data. The next largest difference was the "disabled" category for the NSF group variable. The bad name cases in this category was 3.4 percent higher than the 1993 NSCG sample, 11.5 percent vs. 8.1 percent. Even though this category is relatively small, the difference between the two groups is not alarmingly different.

In general, we feel that even though the estimates between the bad name cases and the 1993 NSCG sample are statistically different, they are not analytically important. This is evident when comparing the distributions between the 1993 NSCG sample and the "Combined Sample." The "Combined Sample" is hypothetically what the 1993 NSCG sample would have been if there had been no bad names. Each table shows little or no differences between these distributions. We conclude that the removal of the bad cases from the 1993 NSCG sample had no effect on the 1993 NSCG estimates.