BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-84/11

# THE IRS/CENSUS DIRECT MATCH STUDY

# FINAL REPORT

by

Danny R. Childers and Howard R. Hogan
Undercount Research Staff
Statistical Research Division
U.S. Bureau of the Census
Room 3582, F.O.B. #3
Washington, D.C. 20233

(301)763-5926

This series contains research reports, written by or in cooperation with staff members of the Statistical Research Division, whose content may be of interest to the general statistical research community. The views reflected in these reports are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries may be addressed to the author(s) or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington, DC 20233.

Recommended by: Kirk M. Wolter

Report completed: July 31, 1984

Report issued: August 1, 1984

# THE IRS/CENSUS DIRECT MATCH STUDY
# FINAL REPORT

## A. Introduction

This project has two principal aims: to investigate the feasibility of using the Internal Revenue Service Individual Master File (IRS/IMF) as a frame for matching to the census in order to estimate gross undercoverage in the census, and to study the difficulties in tracing individuals to the census using the IRS/IMF address.

There has been much discussion recently about using administrative records as a tool in evaluating the census. A combination of administrative records, such as IRS, Medicare, birth and death records, welfare, and other records, has been termed the megalist approach. The IRS/Census Direct Match Study took a sample of persons who filed 1979 tax returns in April 1980. These joint and single filers were matched to the 1980 Decennial Census. Thus, this study can be considered to be a test of the IRS portion of the megalist or composite list for the working age population 18 to 65 years of age.

There are several possible advantages in using the IRS/IMF as the frame from which to draw a sample which will be independent of the census. Since it is not based on household interviews, it is unlikely to reproduce the same omissions as the census. It is especially good for groups with traditionally poor census coverage, such as young working age males. Sampling can be easily controlled on race and income thus permitting the over sampling of black, hispanic, poor, or other "hard to enumerate" groups.

Tracing is a key activity in the proposals for census coverage evaluation research. However, tracing is expensive and time consuming. Tracing will rely heavily on the use of administrative files, such as the IRS/IMF, which will be used to locate a more recent address. The IRS/Census match uses the IRS/IMF directly to obtain a census day residence address to match to the census. It thus increases the understanding of the IRS/IMF as an important tracing tool.

The match from IRS records to census records is conceptually simple. A sample was drawn from the 1979 IRS tax return file. The listing included the name and address of the taxpayer and spouse. The addresses were then coded to census geography and a search of census records was made to see if the sample persons were enumerated in the 1980 Decennial Census. If a person was not enumerated at the tax return address or if we could not code the address to census geography, the sample person was contacted to obtain a correct address or to determine if there existed another address at which the person was enumerated. The sample percentage unmatched will be used as an estimate of census incompleteness for the working age population. The estimates of gross percent enumerated from this study cannot be compared to the Demographic Analysis estimates or the net percent missed from the Post Enumeration Program.

## B. Results

### B.1 Estimated Percent Not Matched

The percent not matched in the age, race, and sex categories in table 1 have been calculated as described in the noninterview adjustment section. The percent not matched in each category is an estimate of the gross missed rate. This number is high in each category partly because the followup interviewing was completed almost three and a half years after the census. This delay made it difficult to get accurate census day addresses. For example, many college students in the age group 18-24 use their parents' address as their permanent adddress when filing tax returns. During followup we contacted the parents' address and the census day address was reported by the parents as the parents' address. Many parents still consider the children as legal residents of the household, but temporarily away. From the results of followup the sample person was not matched because by census definition the child is counted where he/she resides at school. We tried to get other addresses where the sample person may have lived on census day, but we do not believe that these other addresses were consistently requested by the interviewers. If the followup form had been designed differently to probe more for other possible addresses, the results might be different.

The age group 25 to 34 also had a high percent not matched. This further supports the theory that persons below 35 are missed at a higher rate than persons older than 35. After age 35 people seem to settle down more and a person who has a constant, fixed, and visible address has a better chance of being counted in the census.

The overall percent not matched is 12.6 for all races age 18 to 64. The percent not matched is 11.1 for non Black, non Hispanic; 21.5 for Black, non Hispanic; and 19.3 for Hispanics.

**Table 1: Percent Not Matched for Race, Sex and Age**

| | AGE | | | | | | |
|---|---|---|---|---|---|---|---|
| | 18-24 | 25-34 | 35-44 | 45-54 | 55-64 | blank | total |
| Non Black<br>Non Hispanic | | | | | | | |
| Male | 25.6 | 20.1 | 6.9 | 4.2 | 3.5 | | 13.4 |
| Female | 22.4 | 7.4 | 7.2 | 2.9 | 4.2 | | 8.8 |
| Total | 24.1 | 14.0 | 7.0 | 3.5 | 3.8 | | 11.1 |
| Black<br>Non Hispanic | | | | | | | |
| Male | 33.9 | 36.6 | 20.3 | 15.2 | 11.4 | | 26.6 |
| Female | 24.4 | 20.6 | 10.9 | 14.8 | 3.5 | | 16.7 |
| Total | 29.3 | 28.1 | 15.3 | 15.0 | 8.1 | | 21.5 |
| Hispanic | | | | | | | |
| Male | 17.6 | 19.0 | 13.4 | 18.6 | 21.2 | | 17.6 |
| Female | 30.5 | 22.6 | 12.3 | 8.5 | 18.2 | | 18.3 |
| blank | | | | | | 41.1 | 41.1 |
| Total | 24.5 | 20.9 | 12.8 | 12.3 | 20.0 | 41.1 | 19.3 |
| No<br>Characteristics | | | | | | | |
| blank | | | | | | 25.9 | 25.9 |
| Female | | | | | | 28.8 | 28.8 |
| Total | | | | | | 26.9 | 26.9 |
| Total | | | | | | | 12.6 |

There was a small group in the study that had no age, sex, or race information. This group with no characteristics was also matched to the 1980 census and had 26.9 percent not matched. This group contained primarily immigrants and other adults who recently entered the labor force. It is obviously more difficult to trace and match persons without characteristics.

The percent not matched in race, return type, and income categories has also been calculated in table 2. These percent not matched were also calculated as described in the noninterview adjustment section. There are three income categories, less than $8,000, $8,000 to $14,999, and $15,000 and over, indicating gross income from the 1979 tax return. The income on the joint return is the total income for both filers and the income on the single return is the income for the single filer. The median household income is estimated by the Bureau from the 1980 census to be $16,841. Thus the $15,000 and over category is approximately median income and above.

The percent not matched for single filers in all three race categories is much higher than for joint filers. The joint filers are generally older and more settled than the single filers. A person who is single, young, and below the median income has a tendency to be more mobile than the remainder of the population. This person has a greater chance to be missed in the census. If that person is also Black or Hispanic, that person has an even greater chance of being missed in the census. Joint filers with a more permanent, fixed, and visible address will be missed at a lower rate.

**Table 2: Percent Not Matched for Race, Return Type and Income**

| | Income | | | |
|---|---|---|---|---|
| | Less than 8,000 | 8,000 - 14,999 | 15,000 or more | Total |
| Non Black<br>Non Hispanic | | | | |
| Joint | 16.3 | 4.5 | 6.0 | 6.7 |
| Single | 24.1 | 20.8 | 7.8 | 20.8 |
| Black<br>Non Hispanic | | | | |
| Joint | 28.4 | 19.7 | 4.2 | 11.6 |
| Single | 29.8 | 30.3 | 36.2 | 30.9 |
| Hispanic | | | | |
| Joint | 25.9 | 27.0 | 10.9 | 18.3 |
| Single | 25.5 | 24.6 | 33.6 | 26.3 |

B.2 Percent Not Traced

One objective of this matching study using the IRS/IMF as a sampling frame was to see what proportion of the sample persons could not be traced to their place of residence for a followup interview. If a high proportion of the persons selected from the 1979 tax return file could not be matched to the IRS/IMF address and could not be traced to their present address during all of the three followup attempts, the IRS/IMF would not be a good source for sampling persons for coverage evaluation of the census. On the other hand, the 1979 tax return was filed before April 15, 1980 and should be a good source for sampling the working age population that is 18 to 64 years of age.

A sample person was coded as "tracing failed" after the mail followup questionnaire was not returned, the telephone interviewer could not locate a telephone number for the sample person or anyone who ever heard of him/her, and the field interviewer was unable to find the sample person or anyone who could give any information about the sample person. The estimated percent not traced for the total working age population was 3.1 percent.

[illegible]

**Table 3: Percent Not Traced**

| | AGE | | | | | |
|---|---|---|---|---|---|---|
| | 18-24 | 25-34 | 35-44 | 45-54 | 55-64 | total |
| Non Black<br>Non Hispanic | | | | | | |
| Male | 4.4 | 6.1 | 2.9 | 1.7 | 0.0 | 3.4 |
| Female | 5.1 | 3.3 | 1.2 | 0.0 | 0.0 | 2.1 |
| Total | 4.7 | 4.7 | 2.0 | 0.8 | 0.0 | 2.8 |
| Black<br>Non Hispanic | | | | | | |
| Male | 4.0 | 6.7 | 10.1 | 0.0 | 0.0 | 5.2 |
| Female | 0.0 | 3.9 | 14.5 | 0.0 | 0.0 | 4.9 |
| Total | 2.0 | 5.2 | 12.5 | 0.0 | 0.0 | 5.0 |
| Hispanic | | | | | | |
| Male | 3.9 | 8.7 | 0.0 | 8.7 | 0.0 | 5.7 |
| Female | 0.0 | 16.0 | 3.0 | 4.3 | 0.0 | 6.5 |
| Total | 1.8 | 12.5 | 1.5 | 6.5 | 0.0 | 5.8 |
| Total | | | | | | |
| Male | 4.3 | 6.3 | 3.4 | 1.9 | 0.0 | 3.7 |
| Female | 4.3 | 4.1 | 2.6 | 0.2 | 0.0 | 2.6 |
| Total | 4.3 | 5.2 | 3.0 | 1.0 | 0.0 | 3.1 |

## C. Sampling and Variances

### C.1 Sampling from IRS tax files

Although a small sample was desired for the experiment, we wanted to ensure that estimates could be made by "race/ethnicity" and by region. Since race does not appear on the IRS return, a double sampling scheme was employed.

The IRS/IMF files consist of approximately 91,000,000 tax returns. The returns are ordered by the social security number of the primary tax filer. Age, race, and sex do not directly appear on the IRS/IMF. However, the Census Bureau maintains a file of Social Security Administration (SSA) records from which age, race, and sex can be obtained for specific IRS cases.

The Census Bureau's SSA record file consists of 20 percent of all social security numbers. Whether a number falls into the sample or not is based on an algorithm which has been shown to be random. It is possible to identify in advance whether a given number should fall into the 20 percent sample. It is possible then to draw a sample of tax returns such that the social security number of the primary taxpayer always falls into the 20 percent stratum. There was one complication. The Census Bureau's records were not updated recently. Thus, there was a small proportion of cases which fall into the 20 percent stratum, but for which the Bureau had no record. For these, we did not code age, race, and sex. They were sampled differently.

The cost of sampling from the IRS/IMF is high, but largely independent of the size of the sample. Therefore, the size of the initial sample was set at over twenty times the size of the desired subsample. This gave plenty of room for subsampling. A 1-in-415 sample of returns was drawn in such a way that the social security number of each primary filer was contained in the 20 percent stratum. That is, initial sampling was based on:

| | |
|---|---|
| Total Returns (approx) | 91,000,000 |
| 20% Stratum (approx) | 18,200,000 |
| Initial sample | |
| SSN located in 20% file | 215,514 |
| SSN not located in 20% file | 3,602 |

The age, race, and sex of the primary tax filer was coded for the 215,514 located cases. In addition, "Hispanic/Non-Hispanic" was inferred from surname, using the Bureau's file of Hispanic surnames. Region was coded based on the Zip code of the tax return. At this point, non U.S. addresses (e.g., Puerto Rico and Canada) were deleted. Since we do not have data directly on the secondary filer, we assumed that age, race, and ethnicity were the same and sex was opposite that of the primary filer. These assumptions will not hold in a small number of cases, but this number should be insignificant.

The main purpose of the subsampling was to ensure that an adequate number of cases was drawn to allow for separate estimates for each racial/ethnic group and for each region. We stratified bv sex, race, age of primary tax filer, and Zip code. When the primary filer is selected on a joint return the secondary filer is also included in the sample.

We defined the following variables:

1) Sex = Male and Female
2) Race/Ethnicity = Hispanic; Black, Non Hispanic; and Non Black, Non Hispanic
3) Age = 18-34 and 35-64
4) Region = Northeast, North Central, South, and West

We wanted to keep the sampling as close to self-weighting as possible. The universe (all tax filers) divides fairly evenly by age, sex, and region, but race/ethnicity presented a different situation. In order to get approximately one thousand of each sex for Black, non Hispanic and Hispanics, a much higher sampling rate was needed than for non Black, non Hispanic. A sample of one thousand in each sex and race/ethnicity category will allow adequate sample size for regional estimates, as well as national estimates. The remainder of the sample was then allocated to non Black, non Hispanic which gave us an adequate allocation across regions. This lead to the following subsampling intervals:

| | |
|---|---|
| Hispanic | 1-in-6 |
| Black, Non Hispanic | 1-in-12 |
| Non Black, Non Hispanic | 1-in-44 |

Since the initial sample was 1-in-415, the following overall sampling intervals resulted:

| | |
|---|---|
| Hispanics | 1-in-2,490 |
| Black, Non Hispanic | 1-in-4,980 |
| Non Black, Non Hispanic | 1-in-18,260 |

Table 4 gives sample size by race, sex, age and region.

**Table 4: Sample Size (People) by Race, Sex, and Region**

| | Total | Males | Females | Blank |
|---|---|---|---|---|
| NE | 2311 | 1023 | 1121 | 167 |
| Non Black, Non Hispanic | 1268 | 636 | 632 | |
| Blacks, Non Hispanic | 410 | 197 | 213 | |
| Hispanic | 416 | 190 | 197 | 29 |
| No Characteristics | 217 | | 79 | 138 |
| South | 3613 | 1740 | 1767 | 106 |
| Non Black, Non Hispanic | 1630 | 831 | 799 | |
| Black, Non Hispanic | 1131 | 559 | 572 | |
| Hispanics | 725 | 350 | 351 | 24 |
| No characteristics | 127 | | 45 | 82 |
| NC | 2332 | 1107 | 1143 | 82 |
| Non Black, Non Hispanic | 1541 | 784 | 757 | |
| Black, Non Hispanic | 452 | 220 | 232 | |
| Hispanic | 227 | 103 | 111 | 13 |
| No characteristics | 112 | | 43 | 69 |
| West | 2631 | 1165 | 1245 | 221 |
| Non Black, Non Hispanic | 1044 | 538 | 506 | |
| Black, Non Hispanic | 220 | 106 | 114 | |
| Hispanic | 1156 | 521 | 539 | 96 |
| No Characteristics | 221 | | 86 | 125 |
| Total | 10,887 | 5035 | 5276 | 576 |
| Non black, Non Hispanic | 5483 | 2789 | 2694 | |
| Black, Non HIspanic | 2213 | 1082 | 1131 | |
| Hispanic | 2524 | 1164 | 1198 | 162 |
| No Characteristics | 667 | | 253 | 414 |

The social security number for some returns could not be located in the SSA sample file and therefore could not be coded for age, race, and sex. This may have occurred in some situations due to an invalid number that was not detected by IRS. However, most of the primary filers on these returns probably received their number after the 20 percent stratum SSA file was last updated. The filers in these cases are most likely children, immigrants, or other adults who recently entered the labor force.

Children are out-of-scope for this study, since IRS records are known to be representative for only the adult population. Immigrants, possibly including unregistered aliens, constitute a group of particular interest for this research. In order to sample this group, we divided the returns not located in SSA sample file into those showing adult and non-adult characteristics. Operationally, these were defined as:

1) "Adult"
 a) Return type other than single,
 b) Dependents claimed,
 c) Earned income credit claimed, or
 d) Adjusted gross income greater than 15,000

2) "Non-adult" - All others.

The results are displayed in table 5.

**Table 5: IRS Sample of Returns Without Characteristics by Region**

| | Total | NE | S | NC | W |
|---|---|---|---|---|---|
| Total | 3523 | 784 | 998 | 834 | 907 |
| "Non-adult" | 2948 | 617 | 892 | 752 | 687 |
| "Adult" | 575 | 167 | 106 | 82 | 220 |

We included all "adults" in the sample. It should be noted that tracing and matching was especially hard for this group since we lacked personal characteristics.

C.2 Variance of proportion not traced

Hansen and Hurwitz[1] used double sampling for nonresponse problems. This technique was first used when the initial attempt was made by mail and the persons who did not respond to the mail questionnaire were subsampled for a more expensive personal interview.

In applying this technique to this study, stratum 1 consists of persons who are traced before telephone and field subsampling (i.e. persons coded after the initial match and persons who responded to the mail questionnaire). Stratum 2 consists of persons who were untraced after mail followup and were subsampled for telephone and field followup.

$p$ = the proportion untraced

$p_1$= the proportion untraced in stratum 1

$p_2$= the proportion untraced in stratum 2

[1] Hansen, Morris H., Hurwitz, William N., and Madow, William G., 1953 Sample Survey Methods and Theory, Vol. 1, John Wiley and Sons, Inc., New York.

$q$ = the proportion traced

$q_1$= the proportion traced in stratum 1

$q_2$= the proportion traced in stratum 2

$q = 1 - p$

$q_1 = 1 - p_1 = 1$

$q_2 = 1 - p_2$

$n_1$= the number of persons in stratum 1

$n_2$= the number of persons in stratum 2

$n = n_1 + n_2$ = total sample size

$w_1 = \frac{n_1}{n}$ $\qquad$ $w_2 = \frac{n_2}{n}$

$k$ = the inverse of the subsampling rate

$r_2$ = the number of persons who were sub sampled in stratum 2

$n_2 = kr_2$

An estimate of the proportion not traced is

$$p = w_2 \, p_2 \tag{1}$$

since $p_1$ us defined to be zero and an estimate of its variance is

$$v(p) = \frac{pq}{n} + w_2^2 \, (k - 1) \, \frac{p_2 q_2}{n_2} \tag{2}$$

**Table 6: Proportion Not Traced**

| | AGE | | | | |
|---|---|---|---|---|---|
| | 18-24 | 25-34 | 35-44 | 45-54 | 55-64 |
| Non Black, Non Hispanic | | | | | |
| Male | .047 | .060 | .029 | .017 | 0 |
| Female | .054 | .031 | .013 | 0 | 0 |
| Black, Non Hispanic | | | | | |
| Male | .034 | .071 | .166 | 0 | 0 |
| Female | 0 | .039 | .161 | 0 | 0 |
| Hispanic | | | | | |
| Male | .036 | .085 | 0 | .087 | 0 |
| Female | 0 | .173 | .032 | .043 | 0 |

The estimates of the proportion not traced from equation (1) above have been made within each age, race, and sex categorv in table 6. The estimate of the standard deviation of the proportion not traced is the square root of the variance estimate in equation (2) above within each age, race, and sex category. The estimates of the standard deviation of the proportion not traced for the age, race, and sex category are in table 7.

**Table 7: Standard Deviation of Proportion Not Traced**

| | | | AGE | | |
|---|---|---|---|---|---|
| | 18-24 | 25-34 | 35-44 | 45-54 | 55-64 |
| Non Black, Non Hispanic | | | | | |
| Male | .026 | .022 | .020 | .016 | 0 |
| Female | .029 | .017 | .013 | 0 | 0 |
| Black, Non Hispanic | | | | | |
| Male | .033 | .038 | .076 | 0 | 0 |
| Female | 0 | .026 | .045 | 0 | 0 |
| Hispanic | | | | | |
| Male | .034 | .038 | 0 | .050 | 0 |
| Female | 0 | .045 | .029 | .038 | 0 |

C.3 Variance of proportion missed

Expanding the theory of Hansen and Hurwitz to allow for triple sampling for nonresponse, Stratum 1 consists of persons who were traced during the initial match and persons who were traced after mail followup. Stratum 2 consists of persons who were traced in the subsampling for telephone followup. Stratum 3 consists of persons who were not traced in telephone followup.

$p_i$ = the proportion missed in stratum i, i = 1, 2, or 3

$q_i$ = the proportion not missed in stratum i, i = 1, 2, or 3

$q_i = 1 - p_i$, i = 1, 2, or 3

$n_i$ = the number of persons in stratum i, i = 1, 2, or 3

$n = n_1 + n_2 + n_3$

$w_i = \frac{n_i}{n}$, i = 1, 2, or 3

$k_i$ = the inverse of the sub sampling rate, i = 1, 2, or 3

$k_2 = 2$ and $k_3 = 4$

An estimate of the proportion missed is

$$p = w_1 p_1 + w_2 p_2 + w_3 p_3 \tag{3}$$

and an estimate of its variance is

$$v(p) = \frac{p\ q}{n} + w_2^2\ (k_2 - 1)\ \frac{p_2 q_2}{n_2} + w_3^2\ k_2\ (k_3 - 1)\ \frac{p_3 q_3}{n_3} \tag{4}$$

The estimates of the proportion missed from equation (3) above have been made within each age, race, and sex category in table 8. The estimates of the standard deviations of the proportion missed for the age, race, and sex categories are in table 9.

**Table 8: Proportion Missed**

| | AGE | | | | |
|---|---|---|---|---|---|
| | 18-24 | 25-34 | 35-44 | 45-54 | 55-64 |
| Non Black, Non Hispanic | | | | | |
| Male | .205 | .128 | .057 | .044 | .036 |
| Female | .165 | .051 | .059 | .056 | .039 |
| Black, Non Hispanic | | | | | |
| Male | .189 | .228 | .125 | .144 | .063 |
| Female | .177 | .098 | .032 | .100 | .032 |
| Hispanic | | | | | |
| Male | .135 | .124 | .101 | .141 | .133 |
| Female | .209 | .159 | .067 | .069 | .121 |

**Table 9: Standard Deviation of Proportion Missed**

| | AGE | | | | |
|---|---|---|---|---|---|
| | 18-24 | 25-34 | 35-44 | 45-54 | 55-64 |
| Non Black, Non Hispanic | | | | | |
| Male | .029 | .017 | .013 | .012 | .012 |
| Female | .029 | .013 | .016 | .012 | .028 |
| Black, Non Hispanic | | | | | |
| Male | .043 | .026 | .042 | .042 | .031 |
| Female | .043 | .028 | .015 | .037 | .023 |
| Hispanic | | | | | |
| Male | .040 | .031 | .033 | .031 | .049 |
| Female | .034 | .032 | .024 | .029 | .061 |

## D. Tracing and Matching

### D.1 Initial Matching at 1979 IRS/IMF address

An initial screening removed the IRS cases with addresses that were post office boxes, rural routes, military addresses, and other nonstandard addresses that could not be easily geocoded. An attempt was made to code the remaining cases to 1980 census geography. The cases that could not be coded to census geography were removed and combined with the ones removed during the initial screening resulting in 1751 cases that were not assigned census geography. The remaining 5685 cases with census geography were matched to the 1980 census. Of the 5685 cases searched at the IRS/IMF address the single filer or both filers on the joint return were matched 78.2 percent of the time to the 1980 Decennial Census at the address reported on the 1979 IRS return. An additional 82 cases had one of the filers on a joint return matched at the IRS address, (i.e., partially matched). Also, 41 cases were determined to be ineligible to be included in the 1980 Decennial Census, because they were deceased, living out of the country, or had an APO or FPO address (military stationed overseas). They were coded as such and removed from further processing. Thus, there were 4485 returns completely coded after this phase of procesing or 60.3 percent of the IRS sample of single and joint returns. (See table 10).

**Table 10: Processing Results**

| | Total cases | Total coded | Cumulative Coded | Remaining not coded |
|---|---|---|---|---|
| IRS cases | 7436 | | | |
| Not Geocodable | 1751<br>(23.5) | | | |
| Searched at IRS/IMF address | 5685<br>(76.5) | 4485<br>(60.3) | 4485<br>(60.3) | 2951<br>(39.7) |
| After mail followup | 2951<br>(39.7) | 705<br>(9.5) | 5190<br>(69.8) | 2246<br>(30.2) |

The 4485 joint and single returns that were coded during the initial match contained 6826 sample persons. Both filers on a joint return were in the sample. The number and percent of the resolved persons in each final enumeration category are in table 11. Almost all of the resolved cases were matched, because only a few were allowed to be anything else before followup.

**Table 11: Results of Initial Match at the 1979 IRS/IMF Address**

| | Persons | Percent of Resolved Persons |
|---|---|---|
| Resolved Persons | 6826 | |
| Matched | 6749 | .989 |
| Not Matched | 4 | .0006 |
| Non Interview | 2 | .0003 |
| Out of Scope | 71 | .010 |

### D.2 Prefollowup Sorting

After the initial match, all cases were assigned to one of three groups:

A. Cases that could be assigned a final match status without followuo.
B. Cases where the address could not be located in the census.
C. Cases where the address was found in the census, but the sample people were not found.

A complete listing of the codes associated with the three groups above is located in Appendix A. Group A needed no followup because the final match status could be assigned after matching the sample persons at the 1980 tax return address to the census questionnaire for that address. Group B went to followuo because the housing unit reported to IRS either could not be converted to census geography or no attempt was made to geocode the address because it was rural or vague. These cases were sent to followuo, first, to determine the census day address and second, to get a location description that would enable us to convert the address to census geography. Group C was geocoded and matched to the census questionnaire for the address reported on the 1980 IRS tax return, but the sample persons were not listed on the 1980 census questionnaire. These cases were sent to followup to get the 1980 census day address.

The reasoning behind creating the above three groups was that the nonmatch rate would be different in the three groups. The noninterview adjustment was conducted separately within the three groups in each demograohic subgroup.

The number of sample persons in each prefollowuo category and in each region are in table 12. The numbers in oarentheses are the percent of persons in each region or total assigned each prefollowup code. Thus, 62.5 percent of the sample persons were assigned a final match status without followuo. Also a total of 21.0 percent of the sample persons were followed up because the address could not be located in the census or was rural or vague and 16.6 percent were followed up because the address was found in the census, but without the sample persons listed on the census questionnaire for the address.

The South had a higher percent of sample persons with prefollowup code B, because the South had more rural addresses and more addresses that were hard to assign census geography. The West contained slightly more sample persons where the address was matched, but the sample persons were not, indicating that there were more movers in the West than in the other three regions.

**Table 12: Prefollowup Code by Region**

| Prefollowup Code | NE | S | NC | W | TOTAL |
|---|---|---|---|---|---|
| | Region | | | | |
| A | 1542<br>(.667) | 1959<br>(.543) | 1543<br>(.662) | 1753<br>(.666) | 6797<br>(.625) |
| B | 371<br>(.161) | 1111<br>(.308) | 424<br>(.182) | 379<br>(.144) | 2285<br>(.210) |
| C | 398<br>(.172) | 543<br>(.150) | 365<br>(.157) | 499<br>(.190) | 1805<br>(.166) |
| TOTAL | 2311 | 3613 | 2332 | 2631 | 10887 |

D.3 Mail Followup

The 2951 unresolved cases after the initial match (39.7 percent of the sample returns) were sent mail followup questionnaires to obtain the sample person's address of residence on April 1, 1980. The 2951 cases involved in mail followup included the 1200 cases where one or more filers were unmatched after matching to the IRS address on the 1979 tax return and the 1751 cases that could not be easily coded to census geography. The cases initially classified as unable to code to census geography were sent a questionnaire designed to obtain the exact 1980 address before additional money, time, and effort were used to code these addresses to census geography. Also, for the post office boxes and rural addresses, a location description and neighboring addresses were requested to make the location of the 1980 residence on a map easier or possible in some cases. There was also a question on the form asking if two names were for the same person in cases where the filer's name was similar to the name listed on the census questionnaire.

The number of postmaster returns was expected, since many people have moved in the two years and six months since census day. The nonresponse rate was higher than anticipated and was disappointing (See table 13).

**Table 13: Results of Mail Followup**

| | Total Cases | Percent of Total Cases | Total Persons | Percent of Total Persons |
|---|---|---|---|---|
| Sent to Mail Followup | 2951 | 100.0 | 4058 | 100.0 |
| Mail reply | 629 | 21.3 | 1005 | 24.8 |
| Postmaster return | 547 | 18.5 | 578 | 14.2 |
| Nonresponse | 1775 | 60.2 | 2139 | 52.7 |
| No characteristics | | | 336 | 8.3 |

The total cases represents the number of followup questionnaires that were mailed. Total persons represents the number of persons requiring mail followup. If both filers required followup, only one questionnaire was mailed to the address for both filers.

The persons followed up by mail separated by type of address before followup are in table 14. Address type B indicates that the address could not be located in the census and type C indicates that the address was found, but the sample people were not found. More than twice as many persons with address type B returned a completed mail followup questionnaire than type C, because they did live at the tax return address. The type B addresses were difficult to convert to census geography. The post master return (PMR) rate was 16.7 percent for type C and 11.6 percent for type B. There was a higher PMR rate for type C because more of them had moved since 1980 than for the type B addresses.

**Table 14: Persons in Mail Followup by type of address**

| | IRS/IMF Address not found in Census (B) | | IRS/IMF Address found in Census but sample filers not listed (C) | |
|---|---|---|---|---|
| | Persons | Percent of Total | Persons | Percent of Total |
| Mail Reply | 756 | 33.1 | 269 | 15.2 |
| Post Master return | 266 | 11.6 | 296 | 16.7 |
| Non response | 1,182 | 51.7 | 953 | 53.7 |
| No characteristics | 81 | 3.5 | 255 | 14.4 |
| Total | 2,285 | | 1,773 | |

A mail followup questionnaire was returned for 1005 persons and 936 persons were assigned a final enumeration status (see table 15).

**Table 15: Final Match Status for Persons Who Returned the Mail Followup Questionnaire**

| | Persons | | Percent of Coded Persons |
|---|---|---|---|
| Coded | 936 | | |
| Matched | | 729 | .779 |
| Not Matched | | 147 | .157 |
| Non interview | | 40 | .043 |
| Out of Scope | | 20 | .021 |
| Not Coded | 69 | | |
| Total | 1005 | | |

D.4 Telephone Followup

All cases that returned a mail followup questionnaire, but required additional information, were sent to telephone followup. Many of these cases were ones where one filer was matched, but the spouse was not, because of divorce or separation. A response of "divorced" does not help to locate the census questionnaire. The exact census day address is needed. Others needed additional information because the mail followup form was not complete. Mailing addresses, either post office boxes or other rural addresses given by the respondent, were easier to geocode with additional location description and intersecting streets. Many college students or other younger persons without an address they consider as their permanent address will respond that their address on April 1, 1980 is their parents' address even when they did not live there. If a single filer was not listed on a census questionnaire that was obviously their family's, the case was sent to telephone followup. A telephone interviewer is more able to discern the person's true census day address than the respondent on a mail followup form.

The postmaster returns and nonresponse cases were subsampled in order to reduce cost. One fourth of the cases where the characteristics were not available for the primary filer were sent to telephone followup along with one half of the remaining PMR and nonresponse cases (see table 16).

**Table 16: Subsample of Uncoded Persons That Were Sent to Telephone Followup**

| | Total Uncoded Persons | Not in Subsample | Sent For Telephone Followup |
|---|---|---|---|
| PMR, address not geocoded or address not matched (H1) | 283 | 144 | 139 |
| PMR, address match, but filers unmatched (H2) | 295 | 148 | 147 |
| Non response, address not geocoded or address not matched (H3) | 1184 | 592 | 592 |
| Non response, address match, but filers unmatched (H4) | 955 | 469 | 486 |
| PMR or no response, no characteristics for primary filer (H5) | 336 | 250 | 86 |
| Needing additional information (H6) | 69 | 0 | 69 |
| Total | 3122 | 1603 | 1519 |

The telephone followup had an overall success rate of 39.1 percent. The cases that were nonresponses during mail followup had a higher completion rate than the ones that were post master returns. Many of the sample persons who did not respond to the mail followup questionnaire, but still lived at the IRS filing address, would give the necessary information to the telephone interviewer. Table 17 contains the number of persons with completed interviews during telephone followup and the completion rates in parentheses. H1 through H6 are defined in table 16.

**Table 17: Results of Telephone Followup**

| | H1 | H2 | H3 | H4 | H5 | H6 | Total |
|---|---|---|---|---|---|---|---|
| Total | 139 | 147 | 592 | 486 | 86 | 69 | 1519 |
| Complete interview | 32<br>(.230) | 30<br>(.204) | 289<br>(.488) | 171<br>(.352) | 22<br>(.256) | 50<br>(.725) | 594<br>(.391) |
| Incomplete interview or noninterview | 107 | 117 | 303 | 315 | 64 | 19 | 925 |

An interview was classified as complete if the person was determined to be traced and the results of the interview enabled a final match status to be assigned, e.g., as address confirmed, a new address given, partial new address, unknown, refused, deceased before April 1, 1980, in the military, but out of the country on April 1, 1980, or emigrated before April 1, 1980. In an incomplete interview no one could be located by telephone to give us any information about the sample person and no telephone number could be obtained, i.e., tracing failed. A person was considered traced even if the information was not geocodeable or the person was classified an unresolved, because the telephone interviewer talked to the sample person or to someone who knew the sample person. In these instances, no useful information was obtained for locating the sample person, but going to the field would probably not obtain anything more useful. For example, if the sample person said that he moved around a lot in 1980 and did not remember where he was living on April 1, 1980, he was coded as unresolved after telephone followup. No field followup was done for these unresolved cases, since it is not likely that talking to him in person will yield any better information than conducting the interview over the telephone. Thus only untraced cases were eligible for field followup.

All sample persons who were traced during telephone followup were assigned a match status. The results of the match to the census are in table 18.

### Table 18: Final Match Status for Persons who were Traced During Telephone Followup

| | Persons | | Percent of Persons Traced |
|---|---|---|---|
| Traced | 594 | | |
| Matched | | 286 | .481 |
| Not Matched | | 194 | .327 |
| Non Interview | | 107 | .180 |
| Out of Scope | | 7 | .012 |
| Not Traced | 925 | | |
| Total | 1519 | | |

D.5 Field Followup

One fourth of the untraced persons after telephone followup were sent to the field for a personal interview. The persons in field followup in each of three race categories are in Table 19.

### Table 19: Field Followup by Race

| | H1 | H2 | H3 | H4 | H5 | H6 | Total |
|---|---|---|---|---|---|---|---|
| Total | 27 | 29 | 82 | 80 | 18 | 4 | 240 |
| Non Black, Non Hispanic | 17 | 11 | 43 | 16 | 10 | 3 | 100 |
| Black, Non Hispanic | 4 | 10 | 23 | 30 | 0 | 0 | 67 |
| Hispanic | 6 | 8 | 16 | 34 | 8 | 1 | 73 |

All persons who were involved in field followup were searched in the census and final match codes were assigned (see table 20). There were 187 persons traced and 53 persons not traced of the 240 persons in field followup.

At each phase of the tracing and matching operations the final enumeration status was resolved for some sample persons. For others, the enumeration status could not be determined without additional followup. The percent matched and not matched of the resolved cases during the initial match of the sample persons at the 1979 IRS/IMF address and during each of the followup operations has been calculated in table 21. As expected, the percent matched decreased with each additional operation. The percent matched is not constant because only the unresolved and untraced cases went to followup and as the followup progressed from mail to telephone to personal visit, the cases become increasingly more difficult and a higher percentage is truly not matched to the census.

**Table 20: Final Match Status for Persons in Field Followup**

| | Persons | Percent of persons followed up in the field |
|---|---|---|
| Total | 240 | |
| Matched | 75 | .312 |
| Not Matched | 86 | .358 |
| Non Interview | 23 | .096 |
| Out of Scope | 3 | .012 |
| Tracing failed | 53 | .221 |

?

**Table 21: Percent Matched and Not Matched of Traced or Resolved Cases at Each Phase**

| | Initial Match | Mail Followup | Telephone Followup | Field Followup |
|---|---|---|---|---|
| Percent Matched | 98.9 | 78.0 | 48.1 | 40.1 |
| Percent Not Matched | 0.06 | 15.7 | 32.7 | 46.0 |

## E. Noninterview Adjustment

Noninterview adjustment is normally based upon variables such as age, race, sex, and size of place. This study was designed to use the status of the housing unit after the initial match as another variable for nonresponse adjustment.

Also, the noninterview adjustment was done separately for each stage of followup within each cell group. In this study we tried to separate the cases into homogeneous groups for whom the percent not matched that were interviewed would be used as the estimate of the percent not matched for the noninterview cases. This resulted is an estimate of the percent not matched that is larger, but is believed to be closer to the actual percent not matched. For example, the noninterviews in group B after field followup who were Hispanic males 25 to 34 years of age were assigned to be not matched based on the nonmatch rate of all completed interviews for Hispanic males 25 to 34 years of age who were followed up by a personal visit in the field. Thus we allocated the noninterviews to matched or not matched based on more homogeneous subgroups.

The percent matched, not matched, and not interviewed for Hispanic males age 25-34 in group A, B, and C is in table 22. The not matched rate for group C is higher than for group B. Thus imputing the noninterviews to matched or not matched within the groups B and C separately will result in a rate of nonmatches that is a truer reflection of the actual miss rate in the census.

### Table 22: Hispanic Males age 25 to 34 Group

| | A | B | C |
|---|---|---|---|
| Percent Matched | 100.0 | 60.0 | 27.6 |
| Percent Not Matched | 0.0 | 14.3 | 49.4 |
| Percent Noninterview | 0.0 | 25.7 | 23.0 |

The result of the noninterview adjustment is in table 23. Group B resulted in 20.6 percent not matched and group C resulted in 63.4 percent not matched. After combining groups A, B, and C, the percent not matched was 19.0 percent for Hispanic males 25 to 34.

### Table 23: After Imputation for Hispanic Males age 25 to 34

| | A | B | C | Total |
|---|---|---|---|---|
| Matched | 522,900 | 138,409<br>(.794) | 79,206<br>(.366) | 740,515<br>(.810) |
| Not Matched | 0 | 35,891<br>(.206) | 137,424<br>(.634) | 173,315<br>(.190) |

If the noninterview adjustment is done without the prefollowup and followup code classifications, the resulting percent not matched in each age, race, and sex category is in table 24. The percent not matched is lower when ignoring the prefollowup and followup codes, but may not be as accurate.

### Table 24: Percent Not Matched (ignoring Prefollowup and Followup codes)

| | Age | | | | | |
|---|---|---|---|---|---|---|
| | 18-24 | 25-34 | 35-44 | 45-54 | 55-64 | Total |
| Non Black, Non Hispanic | | | | | | |
| Male | 21.0 | 14.4 | 5.3 | 4.0 | 3.3 | 9.8 |
| Female | 16.8 | 6.1 | 6.2 | 2.7 | 3.7 | 6.9 |
| Total | | | | | | 8.4 |
| Black, Non Hispanic | | | | | | |
| Male | 30.1 | 32.1 | 11.5 | 15.0 | 10.3 | 22.3 |
| Female | 24.7 | 16.1 | 2.6 | 13.9 | 3.6 | 13.3 |
| Total | | | | | | 17.6 |
| Hispanic | | | | | | |
| Male | 16.0 | 16.1 | 13.4 | 11.3 | 21.2 | 15.2 |
| Female | 30.0 | 16.0 | 10.5 | 9.5 | 18.2 | 16.7 |
| Total | | | | | | 16.0 |

## F. Cost

One objective of the study was to get cost estimates for the various operations involved in matching the IRS sample to the census and tracing the unmatched persons to their residence for a followup interview.

Converting the addresses to census geography or geocoding was a costly operation. The addresses that were street name and house numbers were relatively inexpensive compared to the rural and vague addresses. Many of the rural and vague addresses required a telephone call to the post office in order to get a location description to locate the housing unit on a census map.

**Table 25: Geocoding**

| | Cost | Forms | Hours | Cost/form | Hrs/form |
|---|---|---|---|---|---|
| FY 82 | $ 24,809 | 6,412 | 3686 | $ 3.87 | .57 |
| FY 83 | $ 10,076 | 797 | 1317 | $12.64 | 1.65 |

A total of $3,750 was spent on direct labor costs to key the data in fiscal years 1982 and 1983. Direct labor costs do not include overhead and machine costs. There were 10,887 persons in the sample. Therefore, the average cost per sample person was 34 cents (see table 26).

**Table 26: Data Keying**

| | Cost | Persons | Cost/Person |
|---|---|---|---|
| Total<br>(FY 82 and 83) | $3,750 | 10,887 | $ .34 |

The telephone interviewing was done in the Jeffersonville Processing Office by the Current Projects Branch. The cost in table 27 is direct labor cost, which does not include overhead or the cost of the telephone calls. A total of 1331 hours at a cost of $8,689 was spent on telephone followup. A telephone interview was attempted for 1519 sample persons. The average direct labor cost per sample person was $5.72 and an average of .87 hours or 52.6 minutes was spent attempting each telephone interview.

**Table 27: Telephone Followup**

| Cost | Hours | Persons | Cost per person | Hours per person |
|---|---|---|---|---|
| $8,689 | 1331 | 1519 | $5.72 | .87 |

The field followup interview was more costly than the telephone followup. The cost in table 28 is the total cost, which includes direct labor, overhead, travel, and field interviewing costs. A total of $10,500 was spent to conduct a personal interview for the 240 sample persons at an average of $43.75 per sample person.

**Table 28: Field Followup**

| Cost | Persons | Cost per person |
|---|---|---|
| $10,500 | 240 | $43.75 |

## G. Comparison with the 1980 Post Enumeration Program

This research project was designed to study the IRS records as a source for sampling persons in the working age population. These persons were compared to the 1980 census using the address on the tax return that is filed in April 1980. If the joint and single filers were not found in the census at the tax return address, they were traced to their present address to obtain the address on April 1, 1980. Estimates of the gross percent missed were made for this study.

This study was not meant to be an alternate source of estimates of miss rates for evaluating the 1980 Decennial Census. But for comparison purposes, the estimates of gross percent missed from the Post Enumeration Program are in table 29. These figures are also subject to biases and sampling errors which in some cases may be quite large. They are given here merely to indicate general size.

**Table 29: PEP Estimates of Gross Percent Missed**

| | Percent |
|---|---|
| White | 5 |
| Black | 12 |
| Not Spanish | 5 |
| Spanish | 10 |

The estimates of gross percent missed in this study were higher, but the mail followup was done two and one half years after the census, the telephone followup was conducted in the spring of 1983, and field followup was attempted in August 1983. The recall bias would have been less if the study was done closer to census day. Also, since this research project was only for the working age population (i.e. 18-64), young persons and older persons were not included in the estimates of gross percent missed in this study.

The Bureau intends to continue research into the use of nonhousehold sources for coverage evaluation. This study has demonstrated that the problems of post office boxes, rural routes, and business addresses can be overcome with proper followup procedures. The ease of taking a large and diverse sample including many non traditional addresses was impressive. We believe that the potential of this sampling frame is immense.

Appendix A

## Pre-followup Status Code

**Group** **Description**

A To Edit - Completed Cases

This category includes cases classified as:

A1 Final Match during initial matching

A2 Final, Deceased

A3 Final, APO/FPO - APO/FPO Zip include 09001 to 09899; 96200 to 96699 and 98700 to 98799

B Addresses Not Geocoded

This category inlcudes cases classified as:

B1 Rural, Vague Address
These cases were not searched during initial matching.

B2 Not Geocodeable (Before Searching), Complete Address
Cases returned from Geography Branch marked "Not Geocodeable".

B3 Not Geocodeable (After Searching)
Cases in which Searching was attempted but evidence from AR's indicated that we did not have the correct ED.

B4 Business Address, No Living Quarters (LQ)

B5 Correct Geocode, Address Not Found

C Address Matched, But Person Not Matched

In these cases geocoding information appeared correct, but the person may have moved or may have had another address.

C1 Other Filer Matched
One of two tax filers in a joint return have been matched and information has been entered in Section 2 on the trace form. If this filer is a possible match, the case is categorized as C3.

C2 Other Members Found
The tax filer(s) has not been found but the addresses and probably even the correct household have been located. Many of these will be college students living away from home.

C3 Possible Matches
Based on the information at hand, a match cannot be assigned, but enough information is available to indicate a possible match (PM).

C4 Census Questionnaire Blank or Not Found

C5 Other Household Found
The address was located, but another household with a different last name was living there.

C6 Special Places
Address was found, but was that of a special place.
People do not match.