CENSUS EXPERIMENTAL MATCH STUDIES

Danny R. Childers and Howard Hogan, Bureau of the Census

Introduction

The U.S. Bureau of the Census is interested in investigating different methods of evaluating the coverage of the census of population in the United States. In 1980 the U.S. Bureau of the Census used a procedure similar to the Post Enumeration Survey (PES) to estimate net coverage error. Two months of the Current Population Survey (CPS) were matched to the 1980 Decennial Census to estimate the gross undercoverage rate. In addition, a sample was selected from the 1980 Decennial Census to estimate the rate of erroneous enumerations, i.e. nonexistent persons, persons counted in the wrong place, and multiple enumerations. The difference in the gross undercoverage rate and the rate of erroneous enumerations is the net undercoverage rate. The PES estimates normally assume that the survey conducted after the census is independent of the census. If this assumption is not true, there will be a bias associated with the estimate of undercoverage.

A Reverse Record Check (RRC) is an evaluation program in which a sample of the population is drawn from a frame created prior to the census, traced forward to the time of the census, and matched to the census. The proportion of the sample which is unmatched provides an estimate of the proportion of the population which was missed in the census. The term reverse record check originated from a procedure called a record check in which a sample from the census was matched to an administrative record file. Thus, when the procedure was reversed and used to estimate the completeness of the census, it was called a reverse record check.

The use of the term reverse record check in the present paper differs from the above definition in two ways: the sample need not be selected from an administrative record file, and the sample was selected from a frame created at a point in time before the census. As discussed in the previous section, a major difficulty with the PES is the bias brought about by the probability that those missed in the census will also be missed in the survey. In theory, an RRC will overcome all or part of this difficulty by selecting the sample to be matched to the census at a greater distance in time from the census than the PES sample. Groups of people who are difficult to enumerate on census day should be easier to enumerate or to include in a sample selected several years before the census. Even if this assumption is correct (i.e., even if hard-to-enumerate groups are easier to sample several years before the census), it will be offset to some extent by another type of bias. In an RRC it is inevitable that some fraction of the sample will not be traced successfully to census day. Hence, a bias arises when those people who were traced successfully are more likely to be counted in the census than those who were not traceable.

Purposes of the Match Studies

The IRS/Census Match Study 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.

Tracing is a key activity in the proposals for coverage estimation research. If the Canadian experience is applicable, tracing in conjunction with the Reverse Record Check technique holds great promise for reducing bias in coverage evaluation. However, tracing is expensive and time-consuming. Tracing relies heavily on the use of administrative files, especially the IRS/IMF which will be used to locate a more recent address. The IRS/Census match uses the IRS/IMF directly to obtain and match to a census day residence address. It thus increases the understanding of the IRS/IMF as an important tracing tool.

There are several possible advantages in using the IRS/IMF as the frame from which to draw a sample that is 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. Samples can be easily controlled on race and income thus permitting the oversampling of black, hispanic, poor, or other "hard to enumerate" groups. Since it is a list sample, a smaller sample is necessary than in an interview sample where clustering is usually required.

The match from IRS records to census records is conceptually simple. A sample was drawn from the 1979 IRS tax-return file, which included name and address of the taxpayer and spouse. The addresses were then coded with census geography (i.e., geocoded) and a search was made of census records to see if the people were enumerated in the 1980 census. If a person was not enumerated at the tax return address or if the address was not geocodeable, direct followup was used to obtain a correct address or to determine if another address existed at which the person could have been enumerated. The percentage unmatched will be used as an estimate of census incompleteness for the working age population.

The purpose of the CPS/Census Retrospective Match Study is to test the procedure by tracing and matching persons and households in the March 1977 Current Population Survey (CPS) to the 1980 Decennial Census. It is believed that the bias resulting from lack of independence in an RRC is less for a sample selected several years before the census than for one selected near the time of the census. However, this reduction in bias may be outweighed by a bias resulting from failure to trace the sample persons from their 1977 residence to a 1980 census questionnaire. This is particularly true for hard to enumerate groups.

Tracing and Matching

Tracing is even more difficult because of the additional time after 1980 during which the sample persons must be traced. Both projects began in 1982 with the initial match to the 1980 Decennial Census. The unmatched cases were then sent a mail followup questionnaire in the fall of 1982. If the mail followup questionnaire was returned by the post office as a postmaster return (PMR) or if there was no response, a telephone followup was attempted in the spring of 1983. The remaining untraced cases were followed up with a personal visit in August of 1983. A person who is difficult to trace to a 1980 address in 1980 is

more difficult to trace in 1983. In addition a recall bias is introduced. A respondent may not recall in 1983 where a sample person, such as a child, lived in 1980. The sample person may also have difficulty remembering accurately the exact address where he or she lived. Whenever an address is given where the sample person may have lived on census day, the address is converted to census geography. The census geography allows us to search the address and if the address is located, the census questionnaire for that address is checked to determine whether the sample person is listed.

Comparison of Preliminary Tracing and Matching Results

All phases of the processing of the IRS and CPS samples were performed at the same time and in the same manner. As a result any differences in the outcome are due to the nature of the sample and not to any processing or time differences. Also, there cannot be any learning effect by the clerks in the processing center or by the interviewers when comparing two studies that are implemented concurrently. The numbers and percentages in this paper represent unweighted CPS households in the CPS/Census Retrospective Match Study and single or joint filers in the IRS/Census Direct Match Study. A case is coded if all persons in the CPS household or if the single filer or both joint filers have been assigned a final enumeration status with respect to the 1980 Decennial Census. The enumeration status is one of the following: matched (M), not enumerated (N), linked to a close-out case on the census questionnaire (L), census questionnaire not on microfilm (Q), possible match (P), refused followup (R), unable to geocode (G), unresolved (U), deceased before April 1, 1980 (D), APO/FPO addresses (S), or emigrated before April 1, 1980 (E). All of the above are classified as "traced." "Traced" implies that we were able to contact either the sample person or someone who could provide some information about the sample person's census day residence, even if the information was minimal.

The M code indicates that the sample person was counted in the 1980 Decennial Census. The codes N, L, and Q indicate that the sample person was missed in the census. The codes R, G, U, and P will require a noninterview adjustment or imputation. The codes D, S, and E will be out-of-scope since they were not eligible to be counted in the census.

An enumeration status of not traced or tracing failed (T) was assigned only after mail, telephone, and field followup. The mail followup questionnaire was a postmaster return or nonresponse, no current telephone number could be located, and no one at a previous number had ever known the sample person. Attempts by a field interviewer to contact the sample person at a current or previous address or attempts to contact anyone who had any knowledge of the sample person were at times unsuccessful. If these attempts were unsuccessful, the sample person was classified as a tracing failure.

The first step in matching the CPS sample to the 1980 Decennial Census was to search for the sample persons at the March 1977 CPS address. As shown in Table 1, 42.5 percent of the households were matched for all sample persons in the household. The remaining sample persons with social security numbers were compared to the IRS/IMF file of 1979 tax returns to get the address they used for filing in April 1980. This address was searched in the 1980 census. For the CPS sample an additional 14.0 percent of the households were completely matched or coded. The cumulative total for the CPS sample is 56.5 percent coded with a final enumeration status. For 60.3 percent of the IRS cases the single filer or both of the joint filers were coded with a final enumeration status. At this time "coded" means that the sample person was matched or determined not eligible to be included in the 1980 census.

Table 1: Tracing Comparison: Percent

	CPS	Cumulative	IRS	Cumulative
Coded or Traced				
At 1977 addresses	42.5	42.5	—	—
At 1979 IRS/IMF addresses	14.0	56.5	60.3	60.3
After mail followup	7.4	63.9	9.5	69.8
After phone followup	16.6	80.5	8.5	78.3
Not Traced				
After phone followup	19.5		21.7	

There were no preconceived ideas as to which sample would have the highest percentage coded at this point in the matching process. The CPS sample has more persons per case and should be more difficult to get all persons in the household coded before followup. The IRS sample is composed of taxpaying adults and as a result has one or two persons for each IRS case. Also, there is the problem of CPS classifying college students as a part of their parents' household, while the census counts them at their school residence. But the IRS also contains college students and other young adults who consider their parents' address as their permanent residence even though they do not live there. As a result they filed their income tax return at their parents' address. The CPS sample also contained unrelated household members in 1977 who were not together in April 1980. The IRS sample contained some couples who filed jointly in April 1980, but were divorced or living separately. Followup was necessary for at least one sample person for 43.5 percent of the CPS households and for 39.7 percent of the IRS cases.

Mail followup was attempted first, because it was the least expensive. The post office will forward mail, but only for a short period, generally six months to a year. As seen in Table 2, the PMR rate for CPS was 33.8 percent and for IRS was 18.5 percent. This was expected since the IRS address was more recent in many cases than the CPS address. Many people in the CPS sample had moved between 1977 and 1980. On the other hand, the nonresponse rate for the IRS sample was much higher. Further investigation of this difference is planned after the study is completed.

Table 2: Results of Mail Followup

	% of CPS Mail Followup	% of IRS Mail Followup
Mail reply	28.4	21.3
Postmaster return (PMR)	33.8	18.5
Nonresponse	37.8	60.2

After processing the results of the mail followup, 63.9 percent of the CPS cases and 69.8 percent of the IRS cases were completely coded (see Table 1). The untraced cases (36.1 percent for the CPS sample and 30.2 percent for the IRS sample) were subsampled for telephone followup. One half of the PMR and nonresponse mail followup cases for both studies were sent to telephone followup. The IRS sample contained a small group of cases that had no characteristics for the primary filers. One fourth of these cases with no characteristics that were PMR or nonresponse after mail followup were subsampled for telephone followup. The results of telephone followup are displayed in Table 3. An interview was classified complete if the sample persons could be traced. The CPS sample had a higher success rate than the IRS sample for telephone followup mainly because in most cases the CPS sample had a telephone number for the March 1977 address on the CPS control card. Telephone interviewers had to rely on directory assistance to obtain a telephone number for all of the sample persons in the IRS sample.

Table 3: Results of Telephone Followup

Percent of Telephone Followup

	CPS	IRS
Total		
Complete	50.6	37.6
Incomplete	49.4	62.4
PMR		
Complete	29.5	19.6
Incomplete	70.5	80.4
No response		
Complete	46.4	40.7
Incomplete	53.6	59.3
No characteristics		
Complete	—	24.1
Incomplete	—	75.9
Other		
Complete	85.8	73.8
Incomplete	14.2	26.2

In both studies there was no telephone interview for persons with an unlisted number or for persons who did not currently live in the same general area of the 1980 address. Telephone interviewing works well for persons with telephones and obtainable telephone numbers. The interviewer spent an average of one hour on each CPS and IRS case attempting to locate a telephone number and conducting the interview.

As expected, the success rate for tracing the sample persons in the nonresponse category was higher than for persons in the postmaster return category. The persons who did not respond to the mail followup questionnaire were in many cases still living at that address. The mail followup questionnaire was returned as a PMR because the sample person or persons had moved and the forwarding order had expired. A current telephone number was more difficult to locate for these PMR cases. The cases in the "other" category were ones that returned the mail followup questionnaire, but additional information was needed to geocode the 1980 census address or because the questionnaire was not completed properly. The success rate for these cases was much higher because many of the mail followup questionnaires were returned with a current telephone number for at least one household member.

After processing the results of telephone followup, 19.5 percent of the CPS sample cases and 21.7 percent of the IRS sample cases were untraced (See Table 1). These cases were subsampled further to be sent to field followup.

One fourth of the untraced IRS cases were sent to field followup. Two thirds of the households identified by race as black and twenty percent of all other races in the untraced CPS cases were followed up in the field. As seen in Table 4 190 CPS cases and 176 IRS cases will be followed up with a personal interview in the field. The smaller number of field cases will allow more time and effort for locating or tracing the remaining untraced sample persons. Since these cases in both samples were not traced through mail and telephone followups, these sample persons will be difficult to trace, but the field interviewer will have more resources available for finding people. The interviewer can talk to neighbors or apartment managers and use the most current telephone, city, and suburban directories available. There will be other lists or administrative records that may be unique to the area, but were unavailable for use, except by personal field interviewing.

Table 4: Field Followup

	CPS cases	IRS cases
Total	190	176
Black	71	53
Other races	119	123
Hispanic	—	51
Remainder	—	72

Nonresponse and Nonresponse Adjustment

Treatment of the nonresponse cases is especially critical in coverage evaluation studies. One is trying to measure the proportion not enumerated by equating it with the proportion not matched. Having found the person in the census is usually considered sufficient information to consider him enumerated. The converse does not necessarily hold. More evidence is required of the "not enumerated" cases than of the enumerated ones. The nonresponse group becomes disproportionately not enumerated.

In these studies, we deal with nonresponse in two ways. First, except for the initial match phase, we defined nonresponse status independently of enumeration status. This meant that we coded the

cases as nonresponse before any searching of census records was done. We asked ourselves: "If we do not find this person in the census at this address, will we be confident in declaring the person not enumerated." If the answer was "no," the case was never searched. Because of a strict adherence to this rule, our nonresponse rates will be relatively high, but the proportion not matched will be a truer reflection of the proportion not enumerated.

Secondly, we have designed the processing with the imputations in mind. Normally, inputations are based on such variables as age, race, sex, and rural vs. urban. These are important, however, we have created new variables which we hope to be more closely linked with enumeration status. For example, after initial matching, all cases were classified based on their status at that point. There were three general groups:

A: Cases that could be matched or declared out-of-scope without followup.

B: Cases where we were unable to locate the address in the census.

C: Cases where we could locate the address, but could not find the sample people.

These cases were then further classified. Three examples of the group classifications follow. Group B1 included rural and vague addresses which had never been assigned a geographic code and searched. Group B5 included cases where the correct address region had been searched, but the address not found. Group C3 indicated that a possible spouse was found for the sample person. Clearly, a followup nonresponse case from Group B1 is less likely to be a census miss than a followup nonresponse from Group C3. This information, together with age, race, and sex should allow us relatively homogeneous imputation groups.

Comparison with Canadian RRC

The design of tracing in these surveys was inspired by the Canadian RRC. Since the tracing and matching is not yet complete, a full comparison of results between these studies and the 1981 Canadian RRC is not possible, but some preliminary comparisons are possible.

Table 5: Comparison with Canadian RRC: Percent

	Canadian	CPS	IRS
Matched at Sample address	46	43	60
Additional traced after phone contact	40	37	18
Total traced after phone contact	86	80	78
Sent to field	14	20	22

Although we started with addresses that were more up-to-date, we still had to send a higher percentage to field followup. A prime reason for this difference is the difficulty in accurately assigning geographic codes and searching the census records. A higher percentage should have been matched at the sample address, but was not because we accidently overlooked them. The difference was not a function of the RRC, but of the way the census was conducted.

Experience, or in our case, lack of experience, must also have contributed. Our telephone interviewers and their supervisors were new to this type of survey and they simply did not know the tricks of the trade. The Canadian RRC is the result of 20 years of refinement.

The fact that the U.S. population is some 10 times larger than Canada's must also make tracing more difficult in the U.S. Local knowledge and just plain luck play a large role in tracing and they do not work as well as in a larger country. Accessing administrative record is more difficult. Our IRS records contain 90 million records. Just mounting the data tapes and passing the file requires careful planning. We were not able to use administration records as efficiently as does Statistics Canada. That there were differences did not come as a surprise. In a sense the purpose of the study was to identify those differences. We were pleased that the results came out as close as they did.