

THE FORWARD TRACE STUDY: ITS PURPOSE AND DESIGN

Howard Hogan, Bureau of the Census

1. Why is the Census Bureau Doing the Forward Trace Study?

To evaluate the 1980 census the U.S. Bureau of the Census relied upon two techniques. One technique is aggregate analysis. Basically, the total population is estimated from independent records such as previous censuses, birth records, death records, and immigration records and this total is compared with the census count.

Aggregate analysis has two serious limitations. It only works for those groups for which there are historical statistical series, which for practical purposes means only for whites and blacks. Aggregate analysis cannot be used to make estimates for Hispanics. Further, aggregate analysis has been successfully used only at the national level. The effect of internal migration is too strong to make state or local estimates, although attempts have been made.

The second technique was the post enumeration survey. This approach is based on a case-by-case matching. It is adaptable to any demographic group and to any level of geography.

In a post enumeration survey or PES, a household survey is done soon after the census. One then searches the census records for each person counted in the PES. The proportion of the PES cases not found in the census is taken as the estimate of the proportion of the total population missed by the census.

The post enumeration survey of course, will also miss some people. It will miss some of the same people who were missed in the census. But matching studies will work as long as, on the average, the people counted by the second source are representative of both those counted by the census and those missed by the census. That is, one assumes that the PES is statistically independent of the census. If the people counted by the PES are more likely than average to have been counted by the Census, then the two systems are statistically correlated and the PES results will be biased.

A major criticism of the post enumeration survey as a method for census evaluation is the correlation inherent in doing a household survey near the time of the census and employing similar methods. No matter how carefully one monitors operational dependence between the PES and the census, there is a presumption that those people who are hard to locate in the census will still be hard to locate in the PES.

One way to lessen this problem is to conduct the PES several months after the census. Many hard to locate people will have moved and perhaps be easier to find. The longer after the census, the more independent the PES can be. In 1980, we conducted a two-part PES, doing one survey just after the census in April and a second survey in August, four months later. However, the longer one waits after the census to do a PES, the more delay in producing the final estimates, and, the harder the matching becomes due to increased response error, especially for people who have moved.

One alternative, then, is a reverse record check, or RRC, a technique tested by the Bureau in 1960, but not used by the U.S. since. In order to do a matching study, one needs to create an independent list to

match against the census; a list which represents the total population. A PES creates that list from a survey done after the census. A reverse record check draws that list from a sample drawn some time before the census. A sample drawn from the previous census is usually used. The sample is supplemented by a sample of birth records, immigration records and a sample of people missed in the last census (as determined by a PES or previous RRC).

The basic logic of the RRC can be expressed simply. With time, a person's chances of being interviewed change. Some people who were 'countable' five or ten years ago become 'uncountable' at census time. Some 'uncountables' become 'countable'. The 12 year old five years ago is now the 17 year old hanging around on the street. The 17 year old of five years ago is now 22 and perhaps now has a job, a wife, and high degree of social visibility. With respect to response correlation bias, a 10 year separation is better; a 20 year separation might be better still.

A reverse record check, when based on a previous census, has other advantages worth mentioning. It has been the U.S. experience, that a census achieves better coverage of the population than can be achieved by any survey, including a PES. This improvement in coverage is especially great among the hard to enumerate, low visibility groups. The publicity in a census, and special procedures for hard to count groups have a pay-off that no survey can duplicate.

This advantage of going back to an earlier census is increased when a sample of missed people is taken from an earlier PES or RRC. Thus the sampling frame for the RRC tends to be both more independent and more complete than can be achieved by a PES.

Another advantage is that the data are collected and can be ready for initial matching as soon as the census is ready. A PES must normally wait several months after the census in order to avoid overlapping with and interfering with the Census itself.

These advantages are balanced and may be overwhelmed by the problems of tracing. Census records are arranged geographically by the place where a person was enumerated. To locate a person in the census, one must first learn his census day residence. To find out the census day residence, one must normally find the person and ask. Although the starting frame may be more complete and more independent than a PES, the final sample that can be traced successfully may be neither complete nor independent. Failure to trace is a major problem. In 1960, the last time the U.S. tried an RRC the results were:

<u>Source</u>	<u>Percent Not Located</u>
Total	12.2
Census	9.0
Missed	16.8
Births	14.4
1950 Registered Aliens	0

Canada has been using the RRC technique since 1961 and has had considerable experience. They also have the advantages of a quinquennial census, thus they only have to trace for 5 years. Even so, their failure-to-

locate runs high. In 1976, the results were:

<u>Source</u>	<u>Percent Not located</u>
Total	4.8
Census	3.1
Missed	9.6
Births	7.6
Immigrants	10.6

A non-response rate of 4.8 may not seem intrinsically high, but it is worrisome when one considers that it may be bigger than the miss rate in the census. It is also troubling to note that hard to enumerate groups are thought to be hard-to-trace groups.

Tracing in both the United States and Canada was done after the current census was over, that is it did not begin until the end of the tracing period. For example, we started tracing in 1960 knowing only the 1950 address. We call this approach Retrospective Tracing.

The problem with retrospective tracing is that the trail is cold before tracing is begun. If we start tracing at the beginning of the period, and follow these people along, we might reduce dramatically the untraceable rate. We call this Forward Tracing.

There are basically three approaches to Forward Tracing:

1. Tracing without any personal contact.
2. Tracing with contact at the beginning, but no further personal contact.
3. Tracing with periodic contact, year-by-year if necessary.

Personal contact is sure to improve the tracing, but it might introduce a conditioning bias. It may also dramatically raise cost. These considerations lead to the Forward Trace Project.

2. What is the Forward Trace Study?

The Forward Trace Study is designed like a reverse record check. We have a sample from the 1980 Census, and a sample of missed people from the 1980 PES. We are drawing a sample of immigrants from Immigration and Naturalization Service (INS) records. Finally, we are drawing a sample of births from state records.

The Forward Trace Study differs from a reverse record check in that it is purely experimental. We are only going to continue until late 1984, at which point we will close out the study. We will not have an actual census to match against. Instead, we will have to simulate a census by sending census-like questionnaires to addresses where we believe sample people are living.

As in a census, these questionnaires will be sent to the address, not the people, and will ask those living there to enumerate themselves. We will then check the self-enumerated household roster to see if the sample people were listed. Any sample person not listed on the census questionnaire will be followed up to determine whether the problem was incorrect tracing or undercoverage in the simulated census. The Forward Trace Study also differs from a Reverse Record Check in that we are testing different tracing strategies. We have defined the following experimental groups:

- A = periodic tracing with annual personal contact
- B = periodic tracing with one initial contact

- C = periodic tracing without personal contact
- D = retrospective tracing at the end of the period

One-fourth of the sample is assigned to each of the four treatments. Sample persons assigned to treatment group D will have no contact or tracing of any kind until the end of the study.

FORWARD TRACE STUDY DESIGN Sample Size People

<u>Frame</u>	<u>Treatment</u>			
	A	B	C	D
C-Sample	3967	3967	3967	3967
M-Sample	1333	1333	1333	1333
B-Sample	900	900	900	900
I-Sample	900	900	900	900
Total	7100	7100	7100	7100

Sample persons assigned to treatment group A will be traced with annual personal contact. Tracing will begin using the name of the sample person. If this is unsuccessful, the other household members will be used in order to trace the sample person. For example, if the sample person is a child missed in the census or born since the 1980 Decennial Census, the parents and other adult household members not in the sample will be used to trace the child.

The interviewer will begin with the original source address and attempt to verify the address or update with the most current address. If the interviewer is unable to contact the sample person or other household member, we will trace the sample person by contacting the relatives, neighbors, or any other contact persons that are known to us. Sample persons assigned to treatment group B will be traced similarly to those in group A except personal contact will be made only once, at the beginning of the tracing period. Sample persons assigned to treatment group C will be traced without making personal contact. With treatment groups A, B, and C attempts will be made to obtain the proper address location for each sample person.

Sources of tracing for the non-contact groups (B and C) will be the post office, telephone books, directory assistance, city or suburban directories, utility companies, or any other lists that may be available for the area. We can also trace the sample persons by matching to other lists or administrative records, such as the Social Security Administration, Internal Revenue Service, Immigration and Naturalization Service, Department of Motor Vehicles, and Department of Labor, to get more current address information.

At the end of the Forward Trace project, the Bureau should be able to assess its ability to implement and carry out a reverse record check. Realistic estimates of cost and staffing requirements can be made. Important design issues will be clarified. Should the Bureau of the Census choose to utilize a reverse record check as part of its strategy to analyze 1990 census coverage, this project will constitute a dress rehearsal.

3. Specific Questions the Study will Answer

Now that we have had an overview of the project, let us turn to some specific questions the study will answer.

3.1 What is the percent of subjects who are successfully traced using each of the four tracing methods?

We will not really know until we are done. For the non-contact groups, it is too easy to think you have traced someone when you really have not. Only when one actually contacts a sample person is the trace confirmed.

3.2 How does success in tracing vary by tracing approach for the different frames and different demographic groups?

Some tracing approaches may work for suburban areas, but not in central cities or rural areas. Some techniques may work for the poor but not the rich. Some techniques may work for blacks and whites but not for Hispanics. The birth sample and the immigrant sample clearly will take different approaches. We must learn these things in order to design an RRC which is both effective and affordable.

3.3 What are the costs of each of the four tracing methods?

In order to achieve the desired precision, the 1980 PEP sampled around a half of a million people. With samples this size, the question of per unit cost becomes an important issue in choosing an approach. Average costs of \$100 per case would have put the total cost at \$50 million. The forward trace will show if the U.S. can afford an RRC.

3.4 Is the RRC approach more cost-effective for some population segments than for others?

We may decide that the RRC yields acceptable results at acceptable costs only for some groups. For other groups another approach, such as a PES, may work just as well and at lower costs.

Gathering cost data directly on different population groups is not possible. Interviewers, clerks, as well as the computer, process all cases together and don't give separate changes for each age, race, sex, etc. group. But, by examining the steps each case went through, one can determine approximate costs.

3.5 Does repeated contact in annual tracing cause a bias due to a conditioning effect?

Periodic personal contact is the only way to be sure that a person is traced. It may actually be the cheapest way by avoiding an expensive trace at the end, when all the good leads are lost.

One must worry about the effect of calling up a person every year and saying, in effect, "We are the Census people. Who is living with you?" Will these people react to the census the same as other, non-sample people.

We hope that we can get a handle on this problem with the pseudo-census. If we tend to get a perfect match-up between the traced household and the reported one, we have evidence of a conditioning bias. If the people left off tend to be the same as those omitted in a real census, conditioning may not be important.

3.6 Can methods of sampling immigrants be

developed?

The 1960 U.S. Reverse Record Check drew its sample of immigrants from the lists of resident aliens who registered through the Post Office each year. This list included some aliens who had entered before 1950 and were counted in the 1950 census. The list excluded aliens who entered after 1950 but became citizens before 1960. It, however, had the advantage of having a current address.

In any case, alien registration stopped in 1981. We had to develop a different method. Essentially, we sample immigrants from a INS computer file based on date of entry. INS clerks then pull the immigrants file and transcribe the address and other tracing information. These forms are sent to us to begin tracing.

Several problems remain. Immigration law is extremely complex. It is difficult to make the legal categories agree with what a demographer would consider an immigrant. For sampling purposes, we want anyone likely to stay until the next census. Problem groups are:

1. Foreign Students: Although they are counted in the census, they are not considered immigrants. We still need to work out special procedures to sample this group.
2. Refugees: This group is large and complex. Until they actually apply for permanent residence, INS does not keep a file. Most apply a year or two after arrival in the U.S., so we can sample them then.

Administrative files are constructed and maintained for administrative purposes and not as clean sampling frames. The INS record system is currently undergoing major revisions which are creating difficulties now, but which should help in the long run.

3.7 Can methods of sampling births be developed?

When we started two years ago, we thought that this would be easy. Nearly all births are registered, and the Federal government has the computer file listing all the births. We quickly learned that this sample is one of the most difficult to construct.

Registration of births is the prerogative and domain of the states. Actually there are 52 independent organizations, counting the District of Columbia and New York City. Each has its own rules, its own procedures and its own concerns. Most state vital statistics offices have been as cooperative as they can, given the constraints of their laws. Only a few states have been unable or unwilling to cooperate.

Many other states have imposed constraints on the handling of illegitimate births and adoptions. These restrictions are not unreasonable, but they greatly complicate doing a national survey.

The National Center for Health Statistics collects the statistical information from the states and prepares the national vital statistics publications. NCHS does not get names and addresses from the certificate. It does get birth certificate number which permits sampling. NCHS is unable to share this data with the Census Bureau without authorization of the states. The result is that, we are just now able to select a sample and hope to begin interviewing in a few months. This has been a real lesson in the fragmented nature of the U.S. statistical community.

Another problem we have not solved is that of timing. NCHS does not get its file ready until some 18

months after the end of the calendar year. Thus, one would have to wait until June 1992 to get the birth for January-April 1990. This delay would be intolerable in a real RRC. In the 1960 RRC, the Bureau drew the sample directly from the states. We may be forced to again return to this system.

3.8 What are the procedures, forms, questionnaires, computer programs, etc. necessary to implement an RRC?

Families move, split apart, come back together. People immigrate, die, get married. We are tracing through post office checks, telephone interviews, personal visits, administrative records. A major issue is learning how to control all the information, and forms from all the households, and people. Much of the success of the Canadian RRC derives from the fact that they have done it five times and have had time to develop the necessary forms and procedures.

4. Questions the Study Will Not Answer.

It is as important to understand the limitations of the study as it is to understand its promises. Let us turn to some questions the study will not answer.

4.1 What information do you need to decide a person was not enumerated?

This question may be the most important in census coverage research. It is easy to show someone was counted; impossible to prove a person was not enumerated. We can trace a person, interview him, then search the addresses. The problems of missing data, insufficient information, response and recall bias remain. The matching in the Forward Trace will not be a true test of matching rules and procedures. These rules may be a more important cause of bias than the completeness of the frame. The CPS/Census Retrospective Match was an attempt to address this question in the context of tracing.

4.2 Can undocumented immigrants be sampled?

Some illegal aliens are no doubt included in the census frame, others in the missed frame. People entering illegally since 1980 have virtually no chance of being sampled, unless like the Cubans and Haitians they are able to convert to a legal status. If the problem of undocumented immigration continues through the 1980's, a serious bias will be built into any reverse record check. After all, to do a record check, you need to have a record. It can be argued that a PES at least has some chance at sampling undocumented immigrants.

4.3 How can people moving from Puerto Rico be sampled?

The Canadian RRC ignores people moving into the provinces from the territories. Presumably, this group is small and not too different from the general population. Movement from Puerto Rico and other U.S. Territories such as Guam and Samoa is not small and the people are, in important ways, different from the general populations. In some local areas, the problem may be large. We need to find a way of sampling this group.

4.4 Does the success of an RRC depend on the "record" or on the time separation?

In its classic form, the RRC is a match from the previous census to the current census. This frame is

supplemented by a sample of births, immigrants, and persons missed in the previous census. It has several advantages. The previous census and supplementary records constitute a more complete frame than can be achieved through a sample survey, such as a Post Enumeration Survey. The second advantage comes from the randomization of the population over time. People move, households change and people who were easy to count now become hard to count and vice versa. This process is thought to increase the independence between the census and the evaluation frame.

These advantages are offset by the loss of both completeness and independence due to failure to trace sample people. This loss presumably increases with time.

A reverse record check which matches between censuses is only a specific case of the idea of a longitudinal evaluation study. One could also conduct an independent survey sometime before the census and match it to the census. By analogy with the Post Enumeration Survey, this approach could be called a pre-enumeration survey or perhaps a pre-census survey.

A pre-census survey loses the advantage of having a more complete sampling frame. However, the interval for tracing is flexible: one can choose an optimal time for the survey.

Among the choices open to the Bureau if it chooses to employ a reverse record check for 1990 are:

- A) A sample from the 1980 Census
- B) A survey drawn from a sample drawn late in the decade (for example 1987).

With its census every 5 years, Canada has never had to choose between the advantage of using a previous census and the advantage of a minimal separation. With our census every 10 years, the choice is more important.

4.5 How can an RRC be integrated with a check of erroneous enumerations to estimate net coverage error?

The RRC estimates only gross undercoverage. To get net undercoverage, one must estimate net overcoverage, i.e., duplicates, fictitious enumerations, out-of-scope enumerations and other census errors. Balancing the rules and procedures of the undercoverage survey to complement the rules and procedures of the overcoverage survey was one of the real challenges of the 1980 PEP. Designing an overcoverage survey to balance a RRC will require considerable thought.

5. Outcome of the Forward Trace Study.

By definition, when you are investigating the unknown you do not know what you will find. Sometime before the 1990 Decennial Census, the Census Bureau will have to decide what strategy should be used to evaluate the census coverage. The Forward Study is designed to provide empirical evidence to help answer that question, obviously with respect to the possible use of a reverse record check. It is unwise to prejudge the outcome of an experiment such as the Forward Trace Study. However, an understanding of the relation between the outcome of the project and the decision to be made, and the timing of that decision is warranted. The following

should bring the problem into sharper focus:

"The Reverse Record Check will not work". This outcome is clearly possible. As long as the study results are made available before a decision is made to implement a RRC, the results are useable.

"Restrospective Tracing is preferable/workable". In this case no decision is necessary until near 1990.

"Forward Tracing is necessary/preferable". In this case, the decision must be made several years before 1990. Just how far in advance will depend on the exact design of the reverse record check.

It should be clear from the earlier discussion that a mixed strategy is possible. A PES may be used for some groups, forward tracing may be used for others. Administrative record checks for still other groups. There are too many unanswered questions to put all our faith on the Reverse Record Check.

The Forward Trace Study is taking place within the context of other Bureau research into census coverage evaluation methodologies. How the Bureau chooses to evaluate the 1990 census and when the decision is made will also depend on these other research projects.

6. Forward Trace in its Wider Context.

I have been addressing the Forward Trace Study purely from the stand point of census coverage evaluation. I think it may have wider implications.

Longitudinal studies continue to grow in importance for the social and health sciences. Much is being written about tracing respondents. Most previous studies have focused on a narrowly defined group of respondents. None has attempted to trace complete spectrum of American society. None had tracing methods and tracing success as its primary focus. I think that the Forward Trace Study will have a unique contribution to this literature.

The Forward Trace Study may also answer some important questions on mobility. For example, what is the rate of return migration of immigrants? What is the emigration rate for other groups? These are not the primary focus of the study. We are not going to sacrifice our main purpose by being overly ambitious. But a potential exists.

Regardless of whether the Reverse Record Check approach is adopted to evaluate the 1990 Census, we will have learned much about American society and the American statistical system from the Forward Trace Study.