Discussion

John L. Czajka

I must begin by noting that I am currently serving as a member of the National Academy of Sciences (NAS) Panel on Alternative Census Methodologies. While I have been asked to comment from my perspective as a member of that panel, none of my remarks should be construed as representing the opinion of the panel.

■ What Makes A Data Base Useful: An Overview

Figure 1 delineates several characteristics that affect the usefulness of a data base for secondary analysis--that is, analysis for purposes other than those for which the data were originally collected.

Figure 1. Characteristics Affecting the Usefulness of a Data Base for Secondary Analysis

- Universe
- Coverage
- Sample design
- Item content
  - Items common to other data bases
  - Unique items
  - Items repeated over time
- Data quality
  - Completeness
  - Accuracy
- Form in which data are maintained
  - Medium
  - Format
  - Consistency over time and geography
- Accessibility
  - To whom
  - For what purpose

Very briefly, the first three characteristics are the universe that the data base purports to represent, the extent to which the data base actually covers that universe, and the design and implementation of the sample by which observations were selected from the full universe. I note under item content that unique items, items that are common to other data bases, and items that are repeated over time are all important, potentially. I list two aspects of data quality: the completeness of the data (how few missing responses) and their accuracy. I also stress the importance of the form in which the data are maintained, and by this, I mean the medium and format in which they are recorded and the consistency with which these are utilized over time and, if applicable, geography. Finally, I include accessibility--both to whom and for what purpose the data can be released.

■ Assessing Administrative Records

How do administrative records data fare by these criteria? Typically, these data represent only a limited universe--the clientele of a program, quite often, or, in the case of vital statistics, the small part of the population experiencing a comparatively rare event during a narrow interval of time. The coverage of this narrow universe is often complete, however; that is, no part of the universe is excluded from representation. Indeed, the level of coverage provided by administrative records is a great strength of such data. Furthermore, administrative records often include the entire population, not a sample. This greatly increases their potential usefulness for analysis, although the sheer size of an administrative data base may make the data unwieldy.

The item content of administrative records is often rather limited. For example, data derived from tax returns are weak on "common" items--particularly demographic variables. Furthermore, the repetition of items over time is subject to legislative action. In Canada, a few years ago, the tax code was altered, eliminating personal exemptions. As a result, data on the size of the

filing unit were no longer collected on the tax form.

With regard to data quality, the completeness of administrative data, both at the unit and item levels, can be very high--better than survey data quite often. The accuracy of the data can be very high as well. The legal requirement to respond accurately is a big inducement to do so. At the same time, however, the possibility of personal gain from responding inaccurately may work against this inducement in some cases. If so, it may introduce error that, for some individuals, cannot be dismissed as mere "noise."

The form in which administrative data are maintained is often not conducive to their usefulness. On a project that I directed for the Department of Labor, we needed to collect extensive county-level data from two States. We surveyed a number of States regarding their holdings. We were particularly interested in whether their data were available in machine-readable form. After selecting two States on the basis of their responses, we received a large box of materials from one of the States and were dismayed to find that the data we had requested were provided on microfiche. I guess somebody felt that we could put the data in a machine and read them.

Finally, a significant factor inhibiting the wider use of administrative records for secondary analysis is that access is very limited, generally.

■ Making Administrative Records More Useful

There are approaches to making administrative records data more useful. I would be remiss if I did not highlight the work of the Internal Revenue Service (IRS) Statistics of Income (SOI) Division in producing data from individual and corporate income tax returns.

To create useful data from individual tax returns, the SOI Division first samples the administrative data base, reducing 114 million records to about 100,000. IRS staff then edit the electronic data by accessing the actual paper documents. Through this editing process, the SOI Division adds new items from the forms; corrects errors in taxpayer arithmetic, interpretation of tax law, and IRS data entry; and improves the consistency of data items across records. Taxpayers' original entries that have particular interest are retained along with the corrected items.

The SOI Division then assembles the data in a form that is more usable than the administrative files. They publish extensive tabulations, provide microdata for tax policy analysis within the Treasury Department and Congress, and even produce public-use microdata.

In discussing ways to enhance the value of administrative records, it is important to note that with appropriate justification and advance work, items can be added to any administrative data system. The Census Bureau has at times added a geography item to the tax return in order to obtain data needed for revenue sharing.

■ External Influences

Changes in the policy arena can affect the usefulness of an administrative data base, as well. Witness the Food Stamp Quality Control (QC) data base, which is assembled from quality control reviews of monthly samples of food stamp case files. These data support one of three microsimulation models used by the Food and Consumer Service (FCS) of the U.S. Department of Agriculture. The other two models use survey data--namely, the Current Population Survey (CPS) and the Survey of Income and Program Participation (SIPP). When Congress was in an expansive mood with respect to the Food Stamp Program and other entitlement programs, a data base that was limited to current participants could not address the most important policy questions, such as how many people would be induced to participate if they were made eligible for any or more generous benefits. More recently, as pressure to contract the Food Stamp Program has increased, the QC data base has become a critical element in policy analysis. Its more complete coverage of the participant population and much larger sample size relative to SIPP and the CPS matter much more now than they did a number of years ago.

Pure necessity, too, can make administrative data more useful--or, at least, more used. The subnational poverty estimates that the Census Bureau is producing

for the Department of Education provide a good example. IRS data, aggregated to subnational areas, will play a major role in the development of these estimates. Here, the deficiencies of the administrative data are addressed in the way that the data are used--specifically, as one component in the estimation methodology. I would note that, for at least four decades, the Bureau of Labor Statistics (BLS) has used data on Unemployment Insurance (UI) claims to estimate for substate areas the total number of unemployed persons in each month. The UI claimants typically represent about half of the total unemployed, with the actual fraction varying widely across geographic areas, as well as over time. Several other sources of data are used to estimate the nonclaimants among the unemployed, component by component.

■ Administrative Records and Privacy

A key ethical and often legal requirement in the collection and use of any data base of personal data is that the respondents or subjects provide "informed consent." Critical to a respondent's ability to provide consent is the right to deny such consent. If a person is required by law to respond to a set of questions, then consent--in any meaningful way--cannot be given. Thus, we have the Privacy Act and other laws relating to data sharing among parts of the government. Lacking the ability, as individuals, to deny consent, the people, acting through Congress, regulate the uses of the data that they are required to provide. (Keep in mind that the same "people," also through Congress or even the Constitution, mandated the legal requirement to provide the data.)

This is not a very satisfying resolution, however. The issues of privacy raised by secondary uses of administrative records must be addressed more adequately. A suggestion to consider is that any planned usage under the census be addressed specifically in legislation or at least a ruling preceded by public hearings.

On this matter, we should not understate the importance of the word "informed." My gut feeling is that there are a lot of people prepared to give uninformed dissent. Don't laugh, but legislators and others who act as representatives of the people can be made more informed.

In the legal mandate that it carries, the decennial census is more like an administrative records data collection than it is like the Bureau's other surveys. Of course, nobody goes to jail for census evasion. But to achieve its mission, the Census Bureau can ill afford to have people refuse to respond to the census because they do not agree with the uses or even perceived uses of their data.

The Bureau is under the legal obligation not to divulge--whether intentionally or not--data that can be associated with the person or other entities to which they apply. Staff of other agencies who have tried to obtain data from the Bureau know how seriously the Bureau takes this obligation. Adding to the Bureau's resolve is the fact that the threat to response rates comes from people's mere belief that such disclosure may occur.

■ Administrative Records and the Census

Let us move now to prospective uses of administrative records in conducting the census.

There are countries--not many--that conduct their censuses entirely with administrative records. A key element in their ability to do this is the existence of a particular type of administrative records data base called a population register. In all of these countries, the availability of a national identification number to identify persons in administrative records systems is helpful in updating the population register. It also facilitates linkages to other data bases that utilize the same personal identifier.

An enormous appeal of an administrative records census is its cost savings. Another appeal is the potential that it provides to generate intercensal data by replication of the (low cost) census procedures. This is possible because the data are always there (and as up-to-date as they are for the census). Tom Jabine, in his comments in the preceding session, identified another appeal: the progressive failure of traditional census methods to get the job done. Certainly, this has been relevant in a number of European countries, and there are signs of such failure here, as well.

In the United States, we do not have either a population register or a national (or universal) ID number. When broad health care reform legislation appeared a possibility, an earlier NAS panel seized on the potential provided by a national health care data base--a kind of population registry, in effect. But health care reform--and the national data base that it would have required--did not come to be.

The Census Bureau, in developing strategies to use administrative records, is responding to recommendations that it make greater use of administrative records. I was going to say that the Bureau has not embraced the concept of an administrative records census, but Tom Jabine's observation in the preceding session that the Bureau is planning a large-scale experiment with administrative records during the 2000 census suggests that this may be something of an overstatement.

■ Population Coverage of Administrative Records

From tax returns filed in 1990, Pete Sailer and colleagues at the SOI Division of IRS developed an estimate of the population represented by annually collected IRS data. Their estimate was unique in that, in addition to persons filing or claimed on tax returns, it included persons captured only on information documents--the documents with which wages, retirement income, capital transactions, and other sources of income are reported to the IRS. Their total estimate amounted to 97.5 percent of the 1990 census count and 95.7 to 95.9 percent of the estimated total resident population, based on demographic analysis or the post-enumeration survey.

The estimate was prepared in such a way that we cannot determine the net contribution of the information documents (or, alternatively, the number captured on tax returns alone). There were an estimated 43.7 million persons who were nonfilers with information documents. Some of these were claimed as dependents, but we cannot tell how many. Some may have also filed only prior-year returns and not been counted as (current-year) filers. I would guess that the information documents netted about 5 to 8 percent of the total population, or more than half of the population that is not covered by tax returns.

Information documents do not include income from food stamps or AFDC. Persons with income only from these sources would not be counted in the IRS data.

I developed a crude estimate, using 1990 March CPS data, of persons in families with no reported income other than food stamps or AFDC. My estimate of 3.4 million (mostly children and young women) would account for about one-third of the IRS coverage gap relative to the Census Bureau's alternative 1990 population estimates. That is, the combination of IRS records with food stamp and AFDC caseload data would account for about 97 percent of the total 1990 population and almost 99 percent of the 1990 census count.

Note that if income from AFDC and food stamps had been required to be reported on information documents, the IRS estimate of the total population would have included the direct recipients but not the children in these families. At the same time, some children who would be counted in food stamp households are not included in my crude estimate because a parent had other income and, therefore, would have been counted in the IRS estimate on the basis of an information document.

■ Census Bureau Uses of Administrative Records

The Census Bureau has announced plans to use administative records in the following ways to assist with the taking of the 2000 census:

- ❒ To impute nonreporting units (about 5 percent of them)
- ❒ To impute some missing items for responding units
- ❒ To prompt respondents in the coverage measurement interviews

(Robert Marx reminded me that administrative records will also be used to help build address lists for special populations, such as residents of Indian reservations.)

In the first two listed uses of administrative records,

their application will be entirely transparent to respondents (and nonrespondents). This is clearly not true of their proposed usage to prompt respondents during the coverage measurement interviews. This is troubling, and I will return to this point below.

Despite these and numerous other innovations, the 2000 census is being prepared with fewer large-scale tests than the much less innovative 1990 census.

■ Recommendations Regarding the Census

Let me close with some brief recommendations with regard to the use of administrative records to help conduct the 2000 census.

- ❒ For both operational reasons and ethical considerations, limit the number of administrative files to a few.
- ❒ Learn these files well. In researching their strengths and weaknesses, develop strong priors regarding expected outcomes when these files are matched to census data and other files. With something as complicated and susceptible to error as record linkage, it helps enormously to have independent estimates of expected results.
- ❒ Explore ways to improve the selected files--for example, by getting new items added or introducing new edits.
- ❒ Reconsider plans to use names from administrative records to prompt respondents during the coverage measurement interviews.

On this last point, concerns about privacy and confidentiality alone would question such usage, but there is reason to be apprehensive about adverse effects on measurement, as well. Prompting residents with names identified as coming from administrative records may influence their responses in ways other than stimulating more accurate recall.