DISCUSSION

Jeffrey S. Passel, Population Division, U.S. Bureau of the Census

The authors' technical abilities and knowledge of their data sets as displayed in these papers do not require comment or criticism. Consequently, I will not discuss specific issues addressed in the papers but rather I will comment on the coverage problem in general, with illustrations from these papers.

The problem of measuring coverage of a census or survey, such as the CPS, is very difficult indeed. It is difficult enough to try to measure the coverage of population in age-race-sex groups for a census but when we attempt to measure coverage for geographic areas or according to socioeconomic characteristics, the solution of the problems involved can become practically impossible. As the three papers presented here have shown, the problems are not insoluble, but may require numerous assumptions and may still be very difficult.

Estimating coverage involves comparing the count or estimate obtained from the census or survey with an estimate obtained from independent sources. Because the undercount is a residual, the estimation procedure requires highly accurate and precise data for both the count and the independent estimate. When the survey estimate is not precise, it can be almost impossible to measure coverage. For example, if the standard error of the survey estimate of group size is larger than the probable undercount of this group, it may be impossible to derive a sensible coverage estimate. In such cases, we must resort to making plausible assumptions about the data and the resulting undercount estimates (if any) can be highly variable and unreliable.

The quality of data used for the independent estimate is also important. Let me say first that what we are doing is attempting to measure the survey data against a standard which is almost always unknown and, in most cases, unknowable. Sometimes we may feel we know what the true value of an aggregate total is, but seldom do we know with any confidence what the true distribution of a characteristic is. If we did know the true characteristic distribution then the problem of measuring coverage would be trivial. Furthermore, we really wouldn't have to take the survey in many cases, since the independent estimate would suffice. However, when the true distribution of characteristics is unknown, we must resort to inference or general indications from partial estimates of coverage. A good example of this sort of detective work is the paper by Alex Korns.

In attempting to explain anomalies in the CPS employment series, Korns hypothesizes that coverage problems in the CPS could account for the observed irregularities. To bolster his case, he examines alternative explanations which turn out to be lacking for one reason or another. Then, he turns to some information about missed persons in the Census and CPS. From these bits of information, he builds a strong case for coverage problems in CPS being the cause of the anomalies. His conclusions seem correct but his case is basically inferential because the true distribution remains unknown and the information about missed persons is generally sketchy.

The methods for measuring coverage are limited only by the availability of independent data and the ingenuity of the researcher but they fall generally into four categories:

1. Component or demographic analysis involves building an estimate from components (births, deaths, and migration for population estimates), as well as using information regarding known internal regularities in the data, such as sex ratios. Population estimates made this way by Siegel and others in connection with the 1970 Census are employed by all the authors in one way or another. Bateman also refers to the difficulties involved in generating such estimates for housing.

2. Reinterview studies consist of the reenumeration of a sample of households to check their coverage in the census or survey. Reinterview studies generally do not provide good estimates of overall coverage because of problems in obtaining "true" matches and nonmatches and because of the so-called "correlation bias"; that is, the resurvey misses people, too, and these tend to be the same people who were missed originally. However, reinterviews can provide a great deal of information on the components of error. We can generally distinguish underenumeration from overenumeration, misses within covered units from omitted units, errors of omission from reporting errors, and types of persons missed.

   The primary value of reinterview studies is that they can provide a great deal of information about the characteristics of missed persons as well as components of error. Much of the inference in the Korns and Yuskavage-Hirschberg-Scheuren papers is based on such information from reinterviews. Note that a reinterview study may not provide a quantitative estimate of the error, but can obviously still be quite useful.

3. Record checks involve comparison of census or survey records with a list of persons who should be in the census or survey. This list (or lists) is usually a set or sets of administrative records, such as driver's licenses, social security files, etc., or it could be another survey. By using a set of records which are independent of the census or survey, the correlation bias can be greatly reduced. However, the problems of

obtaining true matches and true nonmatches are increased because of differences in format and scope of data. This method also can provide information on components of error and limited information about characteristics of missed persons. The paper by Yuskavage et al is based in part on coverage information obtained from record checks using Social Security and Internal Revenue records.

4. Comparison with administrative aggregates (used by all of the authors) is another general type of method for estimating coverage. Birth records, social security data, Medicare files are examples of the type of data used. The data may refer to the entire population, or more often, specific age-sex segments. In most cases, the administrative data must be adjusted for known classes or omitted persons or for differences in definition of characteristics.

Results from all four basic techniques for estimating coverage can be manipulated with various statistical methods such as regression or contingency table techniques.

Given that we can estimate coverage of censuses and surveys within some range of error, there are some other issues which must be faced in using such estimates. First, let us be sure to note that even though the undercoverage of a survey may be substantial, much of the information obtained may be virtually unaffected by coverage error. One example is the percentage distribution of population into income classes shown in the Yuskavage-Hirschberg-Scheuren paper. Although the income distribution of missed persons is substantially different from that of covered persons, the corrected distribution is quite similar to the uncorrected and some parameters (e.g., the Gini index) are almost identical. Another such example comes from our work at the Census Bureau on estimating the coverage of the population of States. Although the undercount in some States is moderately large, the percentage distribution of the State populations change very little when corrected for undercount. In cases such as these, it is only the differential undercount of income classes or States that change the percentage distribution. Furthermore, it takes a substantial difference to alter the basic distribution more than a very small amount.

Another issue that must be faced in using any coverage estimates is what level of error can be tolerated; in other words, when is it preferable to use the corrected numbers over the uncorrected ones. The most stringent error limitations should be placed on corrections for numbers which are used to disburse funds competitively. If the coverage estimates for such numbers can be in error, then the allocations for some areas based on the "corrected" numbers may be further from the "true" allocation than those based on uncorrected numbers. In such cases, the cause of equity will not be served by using "corrected" numbers. This type of competitive allocation requires coverage estimates of a high degree of accuracy and precision, as well as uniform quality for all areas considered.

For noncompetitive allocations, such as capitation grants, the requirement of uniformity in quality may be relaxed, but the accuracy requirements remain. A lower level of accuracy and precision can be tolerated for coverage estimates used in research. Such estimates can be used to indicate whether or not research results are caused by or altered by coverage errors. A still lower level of accuracy and precision can be tolerated in coverage estimates which are used illustratively. Such estimates can still provide qualitative indications of errors and can be useful as rough guides in the broad interpretation of census or survey data even though they may be somewhat inaccurate or imprecise.

The estimates presented in this session generally fall in the middle category. The results obviously have found research applications, but must be refined for the more demanding uses. In conclusion, I would like to commend the authors for their work. Furthermore, I would like to recommend strongly to users of CPS and census data that they take heed of the findings presented here in the course of their own research.