DISCUSSION

Charles D. Cowan, U.S. Bureau of the Census

The role of a discussant is to bring in outside experience to a topic and to coalsce what has been presented by different authors. The thread that ties these papers together is that they represent different points in a continuum of research. In this case it seems that Rogot, Johnson, et al have concluded the second major step in a research program. Winkler is at the third stage, and Buckler is at the first stage. These stages might be labeled as 1) identification of the problem; 2) development of the basic methodology to attack the problem; and 3) using established methods to study the problem and improve methods.

Buckler has concluded the first stage: identification--he now knows what the problem is and has developed a strategy to do research. Rogot, Johnson et al are at the second stage--having developed a strategy, they tested it to see if it would work, what bugs there were in the system, and what they were missing. Winkler is in stage 3--in one sense he is using relatively tried and true methods--but he is also refining techniques and doing further research. Bercini, Sirken, and Mathiowetz are doing the same thing--using established surveys where they know how to do each separately and trying now to combine them to realize an efficiency not previously available--between stages 2 and 3.

Stage 2 of this process is the shortest and perhaps most rewarding; stage 1 is what makes or breaks the study--good planning is a must when given limited or no resources. Stage 3 lasts the longest and gives rise to new studies or means of studying techniques in the most realistic setting.

The first two papers are summaries of joint research conducted by Heart, Lung, and Blood Institute, Census, and National Center for Health Statistics. These papers present long awaited results of a difficult match study that makes use of information from the Current Population Survey and the National Death Index. The papers do not give any idea of how long it took nor the number of persons involved to actually get these studies off the ground. Discussions began in the late 1970's and since the inception of the idea, the staff at the Census Bureau has turned over twice--so it is good to see that the research has finally come to fruition. In order, the papers describe the methodology of the study and the results of the matching for the key characteristics when comparing mortality rates to other sources. The second paper is a glimpse into the future, showing how new and much more detailed data may be available on cause of death. The first paper shows how the match was done and the success of the effort.

It's hard to be a discussant and not be critical, but the terms discussant and critic are not synonymous. And yet as I read these two papers I find that most of my margin notes are unanswered questions. Before I ask these questions, let me applaud some of the answers I have found.

The first answer is to the question--how well does this technique work? Before launching into a major analysis of new data never before available, the authors have shown admirable restraint and fully tested how well the match works by making both internal and external comparisons. And they find that the match has worked fairly well. They have also questioned anomolies--searching for reasons why the match has been successful or not successful--in particular they have discovered that the cohorts with which they are working differ from one another in substantive ways and before analysis can continue, there must be research into methods for making the cohorts correspond to one another.

But I do then have to raise the question--having recognized that the cohorts do not correspond in composition--what has been done to make them fit more closely to one another? In particular, cohorts A and B seem to have different age distributions than the other cohorts. Does generation of age-standardization SMR's improve the results found in the paper? Has that been done yet? What about modeling in a more formal sense to get the cohorts to correspond more or use regression techniques to standardize for age/sex differences.

My second question is--with all these cohorts which come from correlated samples (because the first stage selections in all cohorts are the same)--how does one compute variances of estimates from combined cohorts? This problem was raised in the early '80's and is still not resolved.

Which may be the answer to another question--why are no formal tests conducted in this paper? Marginal distributions are compared between combined cohorts and known demographic results--but there is no test to determine whether these distributions differ. Why not--even nonparametric tests could have been used as approximations. I especially have to question comparisons of seasonality patterns when no test is performed.

In the third paper, Dr. Winkler is taking a well established matching strategy used for demographic files and applying it to business files. I've placed him in stage 3 of my construct because he's extending a well known technique and adding to the body of knowledge in this technique by refining it. Dr. Winkler might have preferred to have placed this in stage 2 since there is relatively little precedent for matching of business files in this fashion. Because of this, there are some new problems which must be resolved that haven't been addressed on the demographic side. One of these is the treatment of businesses that are related. The purpose of the match in this paper was to eliminate duplicates among multiple files and create one master file. But there's a question about how one treats subsidiary companies that's only partially addressed in

this paper. Given the wide range of ways businesses can be linked to each other, through partnerships, dealerships, etc.--how does one decide what is a duplicate; couldn't different uses of this file lead to different definitions of how it would be structured. To be complete in my discussion, I should point out that working with people, the researcher encounters the same problem with defining family units. In a paper given Tuesday, Constance Citro spoke about alternative household definitions and problems raised by extended families.

Dr. Winkler's work, viewed as an extension of the basic Fellegi-Sunter process, is a major contribution in two ways. Including the work cited in two previous papers, he has contributed substantially to understanding how matching can be conducted with large data files in a cost-efficient manner. The second contribution is one of completeness. Besides considering refinements to the methodologies, he also has tested alternatives to compare a variety of procedures and discussed the value of each procedure. The only omission was of a standard procedure in the Fellegi-Sunter process: that of using one set of weights for determining a match or non-match. By dividing the file into classes the discriminating power of differential weights might be great, but the variation on the results in each group because the classes are small might be increased. How does one decide when classes are too small or alternatively that the variance on the matching process is growing too rapidly.

Warren Buckler's paper is a description of a project at a very early stage: after it has been developed but before it has been tried. The description Warren gives of the problems of getting the different agencies involved to agree to the research hardly does justice to the process actually involved. In the earlier papers by Rogot and Johnson I pointed out some of the difficulties the three agencies had and the length of time it took to resolve procedural problems and problems with confidentiality restrictions. In each case however, three agencies have seen their way clear to collaborate on this type of research. The decision making process could have been greatly facilitated, however, with the presence of an arbitrator, something Warren and the earlier authors failed to mention. There was legislation drafted to establish such a function in the Chief Statistician's office and to facilitate the sharing of information. Warren speaks of legislation to mandate government-wide establishment reporting to ensure implementation of the SUAR committee recommendations. While I'm not sure legislation would resolve the problems addressed in the recommendations, it is clear that the current disjointed situation in leadership in statistics will not bring about a resolution. At these meetings we've heard about the scarcity of research monies and the increased push for a reduction in respondent burdens, and yet relatively little has been done to foster relationships between agencies to share information and cut costs of data collection. The NDI study shows a clear need for such collaboration; the work of the ERUMS Work Group also demonstrates this need.

The last paper is not a matching paper in a sense. If it must be categorized, it is more of a dual sampling paper. This paper is also concerned with a shrinking research budget and the need to reduce costs. This study examines some of the options available to reduce the costs of data collection by combining two surveys. It is hard to comment on this paper, however, without some basic information. What would be the result of this research if the data were weighted, and how do these rates compare to unweighted nonresponse rates from the ongoing NHIS? They seem low, but I don't know if this is an artifact of the lack of weights. I doubt it, and so this leads me to question one of the results given in the conclusions, that response rates are robust. It may be that they vary little between the treatments--but are they healthy? Which raises the broader question: How much nonresponse can be accepted in this program or any other when cutting costs? The reduction in costs is a fact we have to live with, but I don't expect respondents to be missing at random and so I question what the true loss is? The real control group for this study is NHIS, but the paper doesn't include results from the regular survey. I also question what differences there will be in item response or nonresponse. This isn't addressed in the paper either, and so I can only suppose that analysis will proceed on this survey and a subsequent paper will look at possible changes in results.