## DISCUSSION

Ralph B. Bristol, Jr., Department of the Treasury

It's not easy for an audience, faced with a pot pourri of such heterogeneous papers as these, to perceive any unifying theme. One of the goals of a discussant is to attempt to provide such a theme. Even the title of the session--"Topics in Administrative Records Research"--is rather a cop-out in itself. At least the papers have in common the fact that they are all based on "administrative records." Meaning what? Meaning that the data were generated for, defined for, optimized for, some purpose other than what the present authors wished to use them for. It's rather like organizing a golf tournament in which the entrance requirement is that participants are permitted to use any equipment they wish except golf clubs and golf balls!

Why then, apart from masochistic gratification, did these people write these papers? I recognize two syndromes, which I label the "Mount Everest Syndrome," and the "Monopoly Casino Syndrome."

By the first, the Mount Everest Syndrome, I refer to the situation in which an author, asked the question "Why do you want to analyze this huge pile of data?" replies, "Because it's there!"

The U.S. Internal Revenue Service (IRS) receives several hundred million pages of tax forms each year; Revenue Canada, considerably less. Some people, faced with any vast quantity of undigested data, will say "There must be something in there worth analyzing!" Unfortunately the truth is: not necessarily!

Sometimes we see aggregations of information which appear so rich, so encyclopedic, that we say "Whatever the question is, the answer is sure to be in there, somewhere." Perhaps, but that's still a far cry from being able to extract answers from the data without knowing, or being willing to state, what the questions are! Such analysis produces frequency distributions and cross-tabs as its final product. It may be important work, very necessary for the advancement of our knowledge, but its value tends to be limited by its very irrefutability--its lack of testable hypotheses. There are few subjects which are so interesting, inherently, that an author is not forced to begin an analysis by explaining why the reader should tarry over his words. The IRS papers, I'm afraid, fail to provide this motivation to the reader.

The Hoskins-Barclay paper on the evolution of the sample design for Canadian personal income tax statistics is a classic situation of a statistician trying to provide the answers when the questions are unknown. Such a statistician really spends his time trying to anticipate what questions will be asked by others in "states unborn and accents yet unknown." Since we in the U.S. do the same thing, I was interested in the similarities and differences between the two evolutionary paths of Canadian and U.S. personal tax statistics.

Many of their remarks apply equally to the U.S. experience. For example: "Users were growing and requiring different analyses"--the U.S. Statistics of Income Division faces exactly the same sort of changing set of customers with changing needs.

"A statistical subsample was built as the basis of a simulation model"--the U.S. Office of Tax Analysis no longer uses the Internal Revenue Service's Statistics of Income (SOI) sample directly in its tax analyses, but rather, a subsample (of about 75,000 returns), which it achieves by a data reduction process based on cluster analysis. This subsample, or "Tax Model" as we call it, is shared with our legislative counter-part, the staff of the Joint Tax Committee of the Congress.

"A customized tabulation service was initiated"--this is just what the Statistics of Income Division of IRS is now doing, providing custom-tailored output on a reimbursable basis.

"To satisfy our own Department's statistical requirements, another subsample was created...supplemented by data related to taxpayer errors, time requirements for assessing, etc."--in the U.S. these are the TCMP or Taxpayer Compliance and Measurement Program sample and TPUS or Taxpayer Usage Study.

At the same time, there are certain differences in the evolution of income tax statistics in the U.S. and Canada. For one thing, our sample is much smaller. Even though we receive almost 100 million returns, rather than 16 million, our SOI (Statistics of Income) sample is less than one-third that of the Canadian, viz 100,000 as opposed to 450,000. For this, of course, we have paid a price. For example, they stratify on province and urban and rural geographical area; the U.S. no longer

provides state data unless specifically requested and paid for by the state(s) involved.

Among other noticeable differences, their sample does not include either late-filers or prior-year returns. We have found these people to be a rather special breed of cat (more tax-shelter and other negative-income returns, for one thing), so we, make it a point to include last year's late-filers in our sample for this year, even though sometimes this causes severe problems, when there is a change in law, for instance. Finally, I notice with a twinge of jealousy that Revenue Canada includes occupational distributions in its output. We have to be struggling to do that, so far unsuccessfully.

In brief, as I said earlier, it's hard to provide the answers when you don't know the questions but are attempting to anticipate them. There is no difference there, between the U.S. and Canadian experiences!

The other two papers do not represent the Mount Everest or "Because the data are there" syndrome, but rather the Monopoly Casino, or "Because it's the only wheel in town" syndrome. They at least know what the questions are, but for one reason or other they are denied the data they need to answer the questions.

Greenwood continues her analysis of the distribution of wealth in the U.S., using income flows from the 1973 Statistics of Income, as originally reported in her paper presented to the 1981 ASA meetings. I am quite disturbed by stock estimates based on the capitalization of single-year cross-section flows, but such concerns were presumably expressed and discussed at the original presentation. This version examines the impact of age and household size. Cross-section regression analyses using these variables seem unable to explain even as much as 10% of the variation of wealth across households. The impact of these two variables on Gini coefficients appears somewhat greater--about 25%. This procedure of comparing an actual distribution not with an ideal, perfectly equal distribution, but rather with one corrected for certain variables is a welcome one. I hope she will continue to refine her research in this direction.

The Norris-Haché paper is another classic administrative records situation: one agency defines the variables, immigration, but doesn't really measure them, so the statistician must work with records generated for entirely different purposes: payment of taxes or family benefits. One of the problems encountered here is the philosophical one of providing an operational definition of "truth"--if alternative estimates disagree, how do you decide which one to accept? Their results demonstrate that the use of alternative data sources can be very valuable in shedding light on otherwise obscured behavior. In this case, the importance of non "landed immigrants" in total immigration had been overlooked by conventional statistics. Similarly, tax and benefit data pointed up recent declines in emigration.

The authors mention that "tax data can be used to estimate the ratio between the migration rate of children and the migration rate of adults." It's not clear to me just how they can be so used, since the authors refer earlier to imputing dependents to tax-filers.

They conclude optimistically that "although each data source has a number of limitations, the strengths of both can be combined..." I'd like to ask how they can be so sure that the strengths rather than the limitations will emerge triumphant. An actress once attempted to seduce George Bernard Shaw, saying "Wouldn't it be marvellous to have a child with your brains and my looks!" Shaw demurred, reminding her, "It might have my looks, and your brains."

To return to the question I posed at the outset: Why did these people write these papers? Why wrestle with data sources which, if not hostile, are at best indifferent? Why put up with variables which are ill-defined and ill-measured for your purposes, only to present them to ill-mannered critics? The answer is: Because they didn't have a choice! For many, if not most, questions in this world, the correct data--correctly defined and correctly measured and correctly collected--are just not available, so we must make do with what we have. It's frustrating work, but I admire people for undertaking it. Only by tackling the job with the tools at hand, be they ever so unsatisfactory, can we make progress toward solutions.

The discussant's comments can be said to deal with the papers at this session from two distinct viewpoints: data user and data producer.

As a data user, the discussant is rightly impatient with the extent to which the papers offer only small advances of knowledge on subjects about which there is no apparent immediate need to know more. The Proceedings versions of these papers attempt to take some account of this criticism. Even so, readers may still feel this viewpoint has merit.

On the other hand, as a data producer, the discussant does find aspects of these papers that are of some interest. After all, the process of converting data into information is always fraught with difficulties, especially where the data came from administrative records — typically collected for another purpose entirely. Even small gains in such an environment may seem well worth it.

Data gathering activities can, as the discussant points out, be likened to mountain climbing and other risky ventures. However, who is to say that just because one cannot scale the highest mountain, the view is not worthwhile from a lower slope?