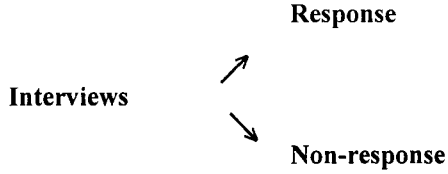
# DISCUSSION

Howard Hogan, Census Bureau
Howard Hogan, Bureau of the Census, Services Division, Washington, DC

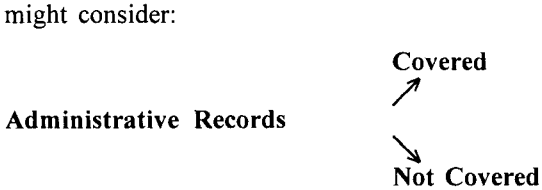
Before talking about the individual papers, let me say a few words in general. All three papers address designs for new longitudinal surveys. Each designer examined the survey objectives, the cost structure, the time frame, and made reasonable choices, sometime clever choices. However, they did not stop there. Each paper describes a process of testing and gathering information to see whether that clever choice was quite so clever. As discussant, I was privy to both early drafts and the later paper. It was clear that the authors' thinking was evolving as the research progressed. We all have ideas which to us seem clever. The lesson of these papers is that clever ideas, tested against real data or in real situations, can be turned into useful idea.

The first paper, by Michaud, Dolson, and Renaud, concerns combining administrative and survey data. I showed this paper to a number of people at the US Census Bureau to get their ideas, and all were quite impressed. The paper contains, I think, quite a clever idea. We have mostly tended to think of using administrative record information as a substitute for collecting the data in the field. Other times, we have thought of using administrative records to fill in non-response records. To my knowledge, we have never given the respondents a choice in the matter. We have never asked them what they wanted!

We usually think of data collection designs as:

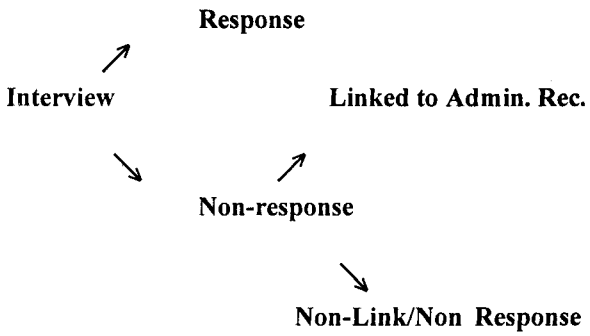



This would be a usual survey. Other times we might consider:



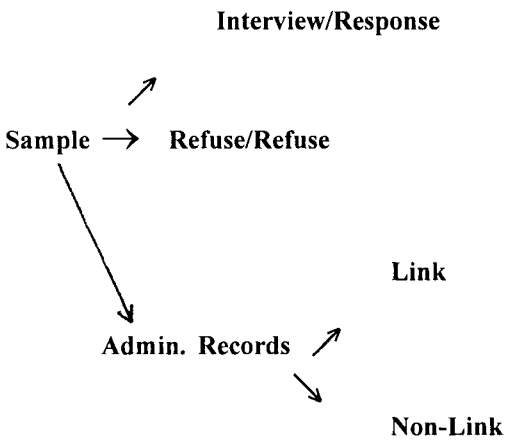

For example, we might just tabulate the administrative records to measure migration or revenue of non-employer businesses.

Sometimes we try




For example, we use the tax returns to impute for non response in our business surveys.

In the Michaud, *et al.* approach, we have




Let the respondent choose.

Of course, there are important differences between the legal systems of Canada and the US. In the US the respondent need not give explicit permission. If we tell the respondent that we plan to use his administrative records and the respondent does not explicitly object, then we have authority to proceed with the matching. In Canada, they must ask, at least sometimes. However, for a number of reasons, matching to administrative records in the US is quite difficult unless the respondent gives us their Social Security Number (SSN). So, in essence, they must give us permission. The Decennial Census has avoided the use of administrative records since we have seen a negative impact that it has on the

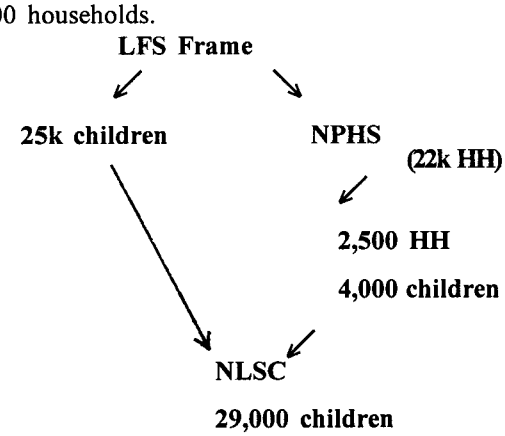
response rates.

Of course, we have never given the respondents anything back in return for letting us use their records, even in the form of a shorter interview. I do not know where these ideas will lead us, but I do know that those people at the US Census Bureau working on administrative records are already reading the Michaud, *et al.* paper, and trying to decide how we can use the ideas.

There are a few things I would like the authors to clarify. For example, it is unclear when permission for linking is required and when it is not required. They seemed to mainly use probabilistic (statistical) matching rather than just linking on SIN. Why?

For those who choose interview rather than administration records, the survey will use CAPI with premailed notebooks. However, their research seems to show that Pencil and Paper work better for respondents who have not prefilled the booklet. Do people who choose interview also tend to cooperate in filling in the notebook? Or, on the contrary, are these simply non-cooperative people, where the interview must be done from memory and thus be better with pencil and paper?

The next two papers are interwoven. The National Longitudinal Survey of Children (NLSC) draws a sample of 25 thousand households from the frame from the Canadian Labor Force Survey. The National Population Health Survey (NPHS) also draws a sample from the LFS frame. In this case a sample of about 22,000 persons. Now, the NPHS draws a third stage sample from these cases, one person per household. However, the NLSC also draws a subsample, in this case, all children in each of 2,500 households.




It's actually more confusing than this, because a special design is used for Quebec, but like the authors, I won't get into this. I take it that adding the 4,000 children from NPHS was not originally planned, but made necessary when the core sample fell short. It would be interesting to know what happened.

In any case, the NLSC is quite an intensive interview. The median length of the initial interview for a family with 2 children is two hours. Up to four children are allowed, so the interview could take, perhaps up to 3 or 4 hours in the worse case And the survey is longitudinal! What is going to be the effect in the long run? After all, if it took two or three hours last time, how eager will the respondents be to hear "Statistics Canada calling"? Can you get any of this information through administrative or school records?

There are teacher questionnaires, principal questionnaires, tests to take. You can see why they are concerned with burden. Are there incentives for the respondents? Do the teachers and principals get anything for cooperating? I found this part fascinating.

Apparently the U.S. Dept of Education is considering this type of survey. However, I do not think that at the Census Bureau we could possibly get away with contacting the schools and telling them that Johnny and Sally were in our surveys, and could you please tell them about them. Our rules of respondent disclosure normally prohibit identifying anyone as being in the sample.

Interestingly, the NLSC takes up to four children per household. The central focus of the Tambay and Mohl article is how to select only one person per household for another longitudinal survey, the NPHS. They investigated solving this problem by rejecting some households. They early on rejected the idea of selecting more than one individual per household. To make the sample representative, they investigated occasionally taking less than one person per household, but never more.

I find the concept of "Representativity" to be a tricky concept. I understand the concept of "efficient sample with known probabilities," and can see why you would want it. I understand the concept of "sample with equal probabilities" or at least not very unequal probabilities, and see why you would want that. If that is all that is meant, fine. However, the word "representative" carries

with it some posterior baggage. How well does the sample measure up against the universe. Clearly, there are an indefinitely large number of dimensions against which to test, so the question is, against which one.

Tambay and Mohl test first against the age distribution. A simple way to do this is to look at the cumulative totals, that the proportion under age 15, the proportion under age 20, etc. If we use the census as a standard, we can subtract the census cumulative percents, and clearly see the differences.

However, as the authors quickly found, things are not so simple when it comes to representativity. There are other domains to look at, proportion of parents as opposed to young childless adults, the proportion French speaking, the proportion of low income. One soon, I feel, begins to leave the concept of representativity and returns to the concept of efficient sample with known probabilities.

However, let me return to an earlier point. Why not more than one person per household? I do not find the reasons given to be persuasive.

One reason given was that taking only one person "Allows more in-depth questioning and shortens interview time." Well, we just saw that in the NLSC, they did not mind an interview of two hours or more. Why is the NPHS more squeamish? Specifically, why doesn't the NPHS take all 4,000 children drawn into sample for the NLSC when they turn twelve?

Taking only one member, we are also told, "Simplifies longitudinal follow-up operations." My experience is that tracing two people from the same family is much easier than tracing two people from two separate households. They may be still living together, which is great. One may have moved away, but even in this case the other respondents can likely tell you where they are. Because of inter-family correlations, you do not really get twice the information, but it costs you much less than twice the cost. So, the question becomes a careful trade off between costs and benefits.

Actually, one real strength of the Tambay/Mohl paper is the explicit discussion of costs, specifically the relative costs of the screener and the interview. This is tricky because the NPHS is a longitudinal survey. However, the effort is valiant.

Too often, survey statisticians treat costs only implicitly. By costs, I mean not just the total survey costs, but the detailed information about the marginal costs of follow-up, call back, travel, increased interview training, etc. There is a good reason for this. We have more difficulty collecting data on our own costs than on the respondents income, assets, and most inner thoughts. No one wants to pay what it will take to collect the costs data we really need. So we rely on hunch, intuition, common sense, good guesses. We probably make pretty good choices. However, as a profession, we need to make more of our assumptions explicit, so that they may be contradicted should the data be found. The Tambay/Mohl paper tries to do this.

Clearly, these are three good papers, about three new and exciting surveys. I hope and trust that in future meetings, the authors, or others, will come and report on the results, and that they will be equally interesting and exciting.