DISCUSSION

Greg J. Duncan, Survey Research Center

The papers of Jean and McArthur and Doyle and Citro focus on several aspects of data collection and management associated with the Survey of Income and Program Participation (SIPP) and its elaborate pretest, the 1979 Income Survey Development Program (ISDP). Both must address issues arising from the basic design of longitudinal surveys of individuals and households and it is worth beginning with a brief review of the sampling theory behind the SIPP design. Since this design is so similar to the one used in the longitudinal study with which I am most familiar--the Panel Study of Income Dynamics (PSID)--I will draw heavily from the 17-year history of that study.

Many cross-sectional surveys obtain their samples of individuals and households by sampling dwellings. Longitudinal surveys can do this as well, as evidenced by the procedures of both the SIPP and PSID. Representative samples of dwellings provide representative samples of subunits within those dwellings--households, families, Food Stamp recipiency units, AFDC recipiency units and individuals. The selection probabilities of each of these subunits are identical to the selection probabilities of the dwelling itself. With a properly specified set of rules regarding the definition of units and the tracking of those units over time, a longitudinal study such as the SIPP or the PSID can maintain a representative sample of each of the various subunits over time. This requires that newly formed subunits of interest (families, AFDC recipiency units, etc.) enter into the sample with known selection probabilities in order to reflect corresponding changes that are taking place in the population at large. It requires that individuals be classified as either "sample" or "nonsample" and that explicit rules be followed consistently in the event of dramatic changes in the composition of units. In the SIPP, as in the PSID, for example, nonsample individuals who join the sample through marriage are followed only as long as they remain attached to a household containing a sample member. Once they regain their independence from all sample members, they are no longer followed.

In general, these sampling considerations require that the study have good systems for (1) tracking all sample individuals, regardless of where they go, (2) allowing individuals to join the sample to provide accurate information about the household in which sample individuals reside, (3) having a fool-proof system of identification numbers for all individuals, sample and nonsample, and (4) storing the data for the individuals and aggregations of individuals so that an analyst can perform a variety of analyses on these individuals in an efficient way.

Most of the field control procedures outlined by Jean and McArthur are quite similar to the ones that have been used successfully by the PSID for 17 years. I do have a few comments about some of them, however.

1. Not all individuals who are institutionalized appear to be carried along as a part of SIPP households. In the PSID, individuals who are institutionalized and cannot be interviewed are associated with a sample family for as long as they remain institutionalized. Of course they are not considered part of the family for most purposes, but the family provides us with the means of tracking them and then reestablishing contact with them when and if they leave the institution. It may be tempting to drop institutionalized individuals from the sample, but there are a substantial number of them, especially at younger adult ages. A strategy of dropping institutionalized individuals in a country with a compulsory, universal military service, would result in all young people being dropped from the sample! Not keeping track of young children who move into institutionalized housing of various types or with relatives who are not sample members means that the SIPP will be unable to inform analysts about such children. (The PSID does not follow these young children either.) They may be too expensive to follow, but the decision of not following them should be based on an appreciation of the consequences.

2. Model-based statisticians may not appreciate the distinction between sample and nonsample individuals and will lament the fact that nonsample individuals are dropped by SIPP once they leave sample households. The PSID does not follow nonsample individuals either, but perhaps this is a mistake. Some methodological work conducted by Finis Welch and his colleagues on the PSID has detected no significant differences between behavioral models estimated for sample and nonsample individuals. (Becketti, et al. 1983.)

3. The nonresponse rules for the SIPP are not entirely clear from the Jean and McArthur paper, especially the rules regarding attempts to contact nonrespondents to waves subsequent to the first one. The PSID does not attempt to recontact these nonrespondents and I think that that is the biggest flaw in the PSID design. Evidence from the new

youth cohorts of the NLS indicates that nonrespondents in one wave are often quite willing to respond to subsequent waves. One gets the impression that refusals or contact difficulties are often quite transitory in nature.

4. The Jean and McArthur paper mentions but does not emphasize the importance of obtaining the name, address and phone number of a contact person who might know the whereabouts of sample households if they move. More conventional means of following individuals such as through forwarding addresses sometimes do not work precisely because the individuals do not wish to be followed easily. In our experience, the contact information is invaluable.

5. Telephone interviewing is mentioned as a possible way of preserving high response rates. The PSID experience suggests that this is indeed true and that data quality does not suffer unduly from switching interviewing modes. Indeed, a substantial number of recontact calls are made to PSID respondents to clean up unclear interview information. Telephones also provide a means of not only reaching geographically remote respondents but also respondents whose time schedules make telephone interviews much more likely to succeed.

Before turning to the subject of the Doyle and Citro paper I would like to make a comment on the interaction between the data collection and data management. Too often we compartmentalize the two without realizing how intimately they are related. As illustrated in the Doyle and Citro paper, data analysts often discover apparent inconsistencies or outright errors and are in the worst position to make an informed judgement about the problems. Data collectors ought to anticipate problems of this sort and have significant resources allocated to solving them. Most of the problems must be resolved by returning to the original protocols, at least briefly, to understand the nature of the problem.

What now of the data structure and methods proposed in the Doyle and Citro paper? Several basic questions come to mind.

1. The most basic question to be asked of any proposed data structure is "Is it feasible?" That the proposed structure has been used with success for several ISDP projects suggests an affirmative answer to this question.

2. The second question, more difficult to answer from the information contained in the paper, is "Is it efficient?" or, more properly stated, "Under what circumstances is it efficient?" Does one need a dedicated machine capable of grinding away throughout the night to select an abstract from the data set with this system, or is it feasible to use the proposed system in a computer environment in which CPU is priced at its marginal cost? I suspect that the proposed system is not very efficient in the latter type of computing environment but I could not tell from the information contained in the paper.

3. Since most "computing" costs are the labor costs of the programmers and other analysts rather than the machine charges, the third question is "Is it easy to use?" Apparently once one has acquired a great deal of specific training about the proposed system, it is fairly straightforward. But outside analysts are encouraged to consider avoiding the data abstracting complications by delegating that work to those who are more familiar with the system.

The data structure that is proposed is modelled after the exceedingly complex file structure used by the Census Bureau. Surely there is a simpler method than an eight-level hierarchy for each wave and four files each with a fifteen-level hierarchy and a completely separate six-level hierarchy that can be used to sort out different aggregations of individuals. The PSID files are more complicated in that they have more waves of data but are simpler in that they are in only one aggregation--the family. It has but two levels to its hierarchy--the family history and the individual. The term "family history" is chosen with care because a major insight, obvious now but not during the first twelve years of the study, are the data structure implications that stem from the fact that not all individuals in a given family in the most recent wave share the same "family history". In fact, we have about seven thousand current families but over nine thousand family histories. The first level of the hierarchy, then, is the family history; the second level consists of the individual histories of all of the individuals who share that same family history. One could also construct "household histories", "Food Stamp recipiency histories", etc. as additional hierarchical levels or as separate records in a networking data structure. These simpler hierarchies require that some of the information from the individual data record be aggregated into the family or household record and this work is probably best done at the Census Bureau rather than having outside analysts attempting to do this with the information they have at their disposal.

A final comment concerns a limitation again attributable to the way in which the Census Bureau processes its data rather than to the organizations such as Mathematica Policy Research that attempt to make sense out of it. Implicit in the file structure is the assumed need to aggregate individuals into households or other sensible units, but not the

possible need to relate individuals to one another. One could think of a file in which all sample individuals had data records that contained information on all individuals who had been or were about to be related to them in some way (by blood, marriage, adoption, or sharing the same dwelling). Information on the related individuals would include wave by wave (or, in the case of SIPP and ISDP, month by month) information on how the individuals were related and whether they shared the same dwelling, family, household, Food Stamp recipiency unit, etc. For most purposes this would be the most general file structure for SIPP, enabling the analysts to distinguish step children from natural children, ex-spouses, and other relatives so that one could analyze the economic consequences of divorce, etc. This would, of course, require a great deal more information than is now currently provided in the Census Bureau's current "relation to head" coding. But the added detail would enable the construction of a file structure that would be of greatest use.

References

Becketti, Sean; Gould, William; Lillard, Lee and Welch, Finis Attrition from the PSID, Santa Monica, CA: Unicon Research Corporation, November, 1983.