DO WE LEARN FROM PAST EXPERIENCE WHEN CONSTRUCTING COMPLEX DATA?

Pat Doyle, Mathematica Policy Research, Inc.

The Survey of Income and Program Participation (SIPP) is a nationally representative longitudinal survey of households in the United States, designed to alleviate a number of problems with currently available socioeconomic data. The survey content and collection methods provide the ability to measure intrayear fluctuations in transfer program eligibility and participation, net worth, income, expenses, employment, and household composition.[1] SIPP grew out of the Income Survey Development Program initiated by the Department of Health and Human Services in the mid-1970s in response to the need for an improved measure of the economic situation of households in this country. The Income Survey Development Program sponsored extensive research into ways in which the measurement, collection, and processing of income, transfer program, and wealth data could be improved. In the late 1970s several site tests and two nationally representative research panels were fielded in order to test alternative collection and processing methods. The last of these tests, known as the 1979 Income Survey Development Program Research Test Panel (ISDP), was sufficiently large to provide reliable national estimates of many household characteristics.[2] However, the survey was so complex and so new that there were major hurdles to overcome before the data could be used for analytical purposes.

The objective of this paper is to compare the public use microdata products from SIPP to those available from the ISDP. Areas where improvements have been made will be discussed as will areas where more work should be done to facilitate the use of SIPP for the analysis of public policy, particularly that policy which affects the low income population. The paper first provides a brief history of the ISDP and lists a series of difficulties experienced with the ISDP from the perspective of a data coordinator whose job was to generate analysis files for specific research tasks and to assist researchers in understanding the data and its limitations. Following that is a section describing the areas in which improvements have been made to SIPP to minimize these difficulties. The paper concludes with the author's "wish list" for further enhancements to the SIPP data products. I might note before proceding that the answer to the question posed in the title of the paper is yes, much has been learned from the ISDP experience.

History of SIPP and ISDP

During the early 1980's the ISDP was being used principally by the government and its contractors as a tool for research in the development of SIPP and little attention was paid to the development of public use files. Unfortunately the program was abruptly halted in early 1982 with no prospects for the ultimate adminstration of the SIPP. With the likelihood that SIPP would never be fielded, a need for useable data products from the ISDP became apparent. Recognizing the need for finishing the work necessary to use the ISDP data, several government agencies jointly provided funding to complete the task of making the data available to the public. At the same time, the Food and Nutrition Service of the U.S. Department of Agriculture sponsored an effort to resolve the data access problems posed by the ISDP through data base management technology (Doyle and Citro, 1984) in order to provide a means to construct analysis files needed for their research.

The work that had to be done to construct usable analytical files from the ISDP consisted of the detection and correction of numerous errors in the data and construction of variables important for determining the composition of households and other units, as well as restructuring the data into a convenient form for access. A substantial effort was required to account for complexities in the file structure due to experimental reporting schemes, interview group rotation and the exclusion of part of the sample in one wave. Many of the enhancements to the publicly available cross-sectional ISDP data were prepared by MPR based on its work in developing analysis files for FNS and in conjunction with the consortium of agencies that collaborated to establish public use files. MPR constructed monthly household, family, and food stamp unit composition indicators to facilitate longitudinal analysis; identified errors and made corrections to the unique person identifiers; performed AFDC, Medicaid, and Food Stamp unit edits on the cross-section files and created variables to denote the program filing units for each of these programs; provided the means to disentangle the results of an experiment in which the reference period for reporting asset income varied; and, finally, was the first organization to successfully link all waves of information collected in the ISDP.

In the midst of the efforts to render the ISDP useable for methodological and public policy research, the Census Bureau obtained the funding needed to conduct the full SIPP. The Bureau, faced with stringent deadlines for fielding the initial wave, proceeded to develop questionnaires, data collection strategies, and processing systems immediately, relying heavily on the previous ISDP experience -- both in the correction of previous errors and in the development of questionnaires and processing systems. As a result, the current SIPP is very similar to the 1979 ISDP. The questionnaires have similar content and organization, many fielding techniques such as the use of staggered interviewing are the same, the procedures for editing and imputing income and recipiency are similar and, finally, the initial microdata product for the first cross-sectional file has essentially the same structure as the ISDP working files.

Difficulties in Working With the ISDP

In the course of producing analytic files from the ISDP, MPR experienced a number of

frustrations which need to be alleviated with SIPP if that new survey is to become a popular data source for the analysis of public policy. Some of the frustrations we faced resulted both from our naivete as consumers in the use of true longitudinal data describing intrayear patterns of individual behavior and the naivete on the part of the producers of the data as to how the data would be used for this purpose. Other difficulties arose from the fact that the ISDP was at least originally intended as an experimental project from which both the producers and the consumers could gain the experience necessary to efficiently comprehend and analyze SIPP. The principal problem faced with the ISDP was (and still is) the total lack of any longitudinal data products. With one exception, all data available to the public now and to the government over the last four years have been issued as a series of cross-section files. The data files from each of the six waves administered for the ISDP were produced independently of one another using procedures developed for a cross-section survey of households. They can not be directly linked to form longitudinal files. Furthermore, when they are linked, observation of changes in individual behavior or income receipts is obscured by the fact that imputation for item nonresponse on any given wave was performed without regard to responses to the same questions on any other wave. Finally, the appropriate sample weights necessary to analyze longitudinal data from the ISDP were simply never developed.

This lack of longitudinal data products was the source of a number of difficulties MPR faced in using the ISDP for policy research. One difficulty was that the accounting period for income data on the cross-section files was monthly while unit composition on the available data files was structured around the relationships that existed at the time of each interview. The interviews were conducted every three months. Researchers, of course, preferred to have the same time frame for both income and composition and the logical approach to meeting this demand was to construct monthly unit composition indicators. MPR proceded to develop these and in the course of so doing several obstacles arose which had to be overcome. In particular, the method of uniquely identifying individuals over time did not work, the use of a cross-sectional approach to determine family composition in each wave obscured some of the true relationships existing within a household, and records of why some individuals left the sample were lost entirely.

Another difficulty faced in using the ISDP was that although one of the survey goals was to study participation in transfer programs such as social security or welfare, the information collected on program units was not restructured into useable form to facilitate the analysis of program participation. Furthermore, the program unit data pertained only to composition at the time of each interview and were not edited to be consistant with changes in household or family composition occurring between waves.

A third area in which the lack of planning longitudinal products compounded the difficulty in using the ISDP was the design of the contents of each of the cross-sectional files. Most of the the data collected was recoded into variables of interest for cross-sectional applications in lieu of (rather than in addition to) recording it as responses to the original questionnaire. As part of this restructuring of the data, income recipiency fields were edited and imputed cross-sectionally. Unfortunately the publicly available products lack flags necessary for identification of when cross-sectional imputation had been made. Imputation flags were recorded when income amounts were imputed but not when income recipiency was imputed.

A fourth area in which the lack of planning for longitudinal use complicated access to the ISDP was in the design of the public use files and the associated documentation. The public use data products consist of one file for each wave containing data arrayed in a complex modified hierarchical fashion plus a few supplemental files necessary to link the data across waves for longitudinal studies. The cross-section files are cumbersome to use both because of the structure and because the documentation is incomplete. There is a record layout but the details of how many of the variables were constructed are simply not documented. Furthermore, there is no system established to provide assistance to users except the Wisconsin data center (Institute for Research on Poverty, 1985) funded by the National Science Foundation in 1984. Another related problem faced with the ISDP was the repeated reissuance of files every time a enhancement was made or a bug fixed. For example, over the years MPR has received a half dozen different versions of the first wave.

There are a few other characteristics which make the ISDP difficult to use for longitudinal analysis which result from the design of the sample and the questionnaires rather than from the lack of preparation of longitudinal data products. These include inadquate identification of truncated spells of program participation, inadequate identification of school enrollment, lack of coverage for persons entering the universe after the initial wave, the reduction of the sample size for one wave, the use of staggered interviewing techniques, and a reference period that is too short for many studies of duration and turnover in program participation.

How SIPP has been Improved

The producers of SIPP have learned a great deal from the collective ISDP experience. The initial data products available for the early waves contain numerous improvements over the ISDP. The data products now available consist of cross-sectional files from the first three waves, each in two alternative formats, and associated documentation in both machine readable and hard copy form. Even though these initial products are cross-sectional and were developed using similar procedures as the ISDP, there are a number of enhancements which will facilitate the use of the data in both cross-sectional and longitudinal studies.

One major enhancement is the existance of more extensive documentation. This is coupled

with more user support which is essential for a survey this complex. The documentation includes a section on the source and reliability of the estimates, an index of the variables, an overview of the survey design, and a copy of the questionnaire. All of these are in addition to the file layout. At this time the document does not contain a description of how the added recodes are constructed but I am confident that it will follow if users indicate a need for it.

Another major enhancement is the issuance of the data in alternative formats. One format is a modified hierarchical file similar to the ISDP files. The other is a rectangular person file where all the household and family information is replicated on each individual's record and all of the income data is summarized at the person level. The first file is more efficient in terms of storage (except for the padding which is discussed below) but more difficult to process. The second file is less efficient in terms of storage but much easier to use. The second file will also facilitate use of the data longitudinally since the linkages must occur at the person level.

The contents of these SIPP cross-sectional products are greatly improved over the ISDP. The questionnaire and control card information is almost entirely replicated on the public use files. In addition to the questionnaire image portion, the files contain very useful variable constructs. In particular, there are program unit variables for all the transfer programs. There are also household and family summaries of income by type and program participation. Of vital importance is the addition of imputation fields for recipiency as well as income amounts.

One very important set of enhancements to the cross-sectional files is the addition of monthly household and family composition indicators. These will greatly improve the researcher's ability to study individual behavior patterns over time. On the first wave these composition indicators do not vary across the four reference months because there was no baseline information with which to describe variation in household and family relationships during that period. Wave II will be the first opportunity to observe changes such as these.

The final improvement to note is that longitudinal products are being planned. The importance of longitudinal weights, longitudinal imputations, and longitudinal unit construction is recognized as is the difficulty of the task of creating them. Extensive research is being carried out in these areas (McMillan and Herriot, 1984. Judkins, et.al, 1984. Ernst, et.al, 1984. Samuel and Huggins, 1984; and Kalton, 1985) and we are informed that perhaps in early 1986 we will have a true longitudinal data product.

What More Can Be Done?

Anyone who has ever produced a data product for general consumption will know that users are never completely satisfied. Of course, I am no exception. Hence I would like to take this opportunity to list the areas in which SIPP could be improved to enhance its popularity for public policy analysis.

There are a number of improvements which could be made at relatively low cost and without redesigning the survey. First, the hard copy documentation could be made easier to use with a simple format change. Variable names could be more informative in the case of variable constructs and imputation flags. (I do not recommend changing the names of the questionnaire image fields which now are a function of the source code numbers appearing on the questionnaire). Finally, more information in the variable definitions would greatly facilitate use of the file layout. For example PP-IMP01 says imputation flag for 'SC1002'. It would be helpful if more text was included in these definitions or if these variables were grouped by topic with the topics labeled. Second, the public use tape for the complex file is now padded out to the length of the longest record. That means over 1,000 characters of storage are zero filled on all record types except the person record (4 record types have over 1,400 fill characters each). This results in the file requiring 3 reels of 6,250 bpi tape to store. When the filler characters are removed the file is reduced to 1 reel. It seems there must be a better way to supply data to users requiring fixed length files. Thirdly, for confidentiality reasons miscellaneous rarely received income amounts have been lumped together without regard to the nature of the income type. It seems the desired level of confidentiality could be achieved with a more meaningful grouping of rarely received income amounts (such as state administered SSI) with other similar sources already individually identified (such as Federally administered SSI). Finally, faster notification of apparent problems with previously released files is needed.

Another area in which I would like to see a change that does not require redesign of the survey (but is more expensive than those listed above) is early release of the results of longitudinal weights and imputations. I realize this cannot happen in the next few years simply because there is considerable methodological research to be conducted. However, once the systems are in place, it would be nice not to have to wait two years after the completion of data collection to begin longitudinal analysis using the information collected in existing multiple waves of data.

Another area in which SIPP could be improved without redesign of the survey is to change the method of defining family groupings within households. The procedures now rely principally on the relationship of individuals to the head of the household. Furthermore, determination of a subfamily group where the head is not married relies somewhat arbitrarily on the age of the youngest subfamily member. Although it is desirable to greatly improve the measurement of relationships among household members through a redesign of the questionnaire (David, et.al, forthcoming) the Census Bureau could use additional information already being collected to improve the family unit construction.

Related to the issue of determining family groupings within households is the manner in which the data are organized on the complex

version of the cross-section files. Why is there one record per household per wave with monthly summaries but four family records for each family within the household, one for each reference month? Note this confusion is only compounded by the fact that there is one record per person per wave and one record for each income type per wave. Aside from the fact that this organization is not logical, it is simply cumbersome to use. For cross-sectional analysis of households and families there are two ways in which the data could be appropriately organized, neither of which is reflected in the current public use files. One way is to use a household month (or family month) concept. The other is to use household and family groupings as of the interview date with restrospective economic data for the previous four months. The former is the preferred choice because the economic and demographic data are (presumably) consistent. However, the latter is attractive because the composition detail is more precise. I think the Census Bureau should consider restructuring the SIPP cross-section files around one of these two concepts but provide enough information to allow the user to employ the other. One fairly simple way to achieve this goal is to insert a fifth family record in the current structure representing family groupings at the time of the interview. However, the preferred approach is to change the structure entirely to a household month file where there is a natural household-family-person hierachy within each time period. This file should contain five sets of monthly records, one for each of the four reference months and one for the interview month. Aside from the ease with which the latter file could be used for cross-sectional studies of the distribution of households and families, this latter approach does not impose any assumptions on what constitutes the same family or household unit over time. The current public use files do this at least at the household level. One argument often posed against this recommended structure is that it is not convenient for use in studying behavior patterns across time. This is certainly true but it is also the case that the cross-section files are not very appropriate for these studies anyway because the reference period is too short in a single wave.

My wish list for enhancements to SIPP extends to areas where redesign of the survey is necessary. The principal concern I have, which is shared by many others (Mathematica Policy Research, n.d.), is that SIPP does not appropriately deal with spell truncation. When an individual is first observed, it is determined whether or not he or she is participating in one or more programs. SIPP does not go one step further to determine the duration to date of this period of participation. For studies of duration of welfare it is essential to distinguish between spells in-progress and those just beginning.

Use of SIPP for analysis of participation in welfare programs is further limited by the absence of an integrated eligibility module such as the one used in Wave II of the ISDP. In order to determine program participation rates or to analyze the determinants of program participation, it is necessary to identify the pool of eligible units. Nationally representative household surveys such as SIPP are the only data sources which permit this identification. However, determination of eligibility requires information on assets and expenses not currently collected in the core module of that survey. Most of it is now being collected in various topical modules but these are administered at different times and some may be subject to elimination in future panels in order to reduce the average response time of the survey. Administering these questions in series of modules introduces a number of complexities because they must be combined in order to determine program eligibility. Aside from the increased cost to link the data, analytic problems arise because of sample attrition and changes in household and family circumstances between waves.

Another area where SIPP is weak is the identification of school enrollment. In the early panels, unless an individual is not working some weeks of the reference period or is age 17 to 49 and enrolled in post secondary school SIPP does not identify whether he or she is in school or the amount of time being spent in school. The later panels of SIPP have been modified to correct this weakness.

Although SIPP is an improvement over the ISDP in the collection and recording of transfer program units, the information gathering could be improved for the less well known assistance programs like the School Lunch and Breakfast Programs administered by the Department of Agriculture.

Finally, the author would like to reiterate one more time how difficult it is to cope with the cost saving measures of staggered interviewing and the reduction of sample size in one wave. I am fully aware that these techniques are necessary to control the cost of collecting the information. However, they immensely complicate the use of SIPP for public policy research.

FOOTNOTES:

[1]For an overview of SIPP see Kasprzyk and Herriot (1985).

[2]For an overview of ISDP see David (1983).

BIBLIOGRAPHY

David, M.; Rockwell R.; Monfort, F.; and Robbin, A. "The Scientific Potential of SIPP: Critiques of its Contents and Methods" in Journal of Ecnomic and Social Measurement. Forthcoming.

David, M. (ed). Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program. New York, Social Science Research Council. 1983.

Doyle, P. and Citro C. "The ISDP/RAMIS II System and Its Development". Washington, D.C.: Mathematica Policy Research, 1984.

Institute for Research on Poverty. "Center For Research and Retrieval of Data From SIPP and ISDP," in Focus vol. 8 no. 1, Spring, 1985.

Judkins, D., Hubble, D., Dorsch, J., McMillen, D., and Ernst, L. Weighting of Persons for SIPP Longitudinal Tabulations. Proceedings of the Survey Research Methods Section, American Statistical Association. 1984.

Kalton, G. "Handling Wave Nonresponse in Longitudinal Surveys." Paper prepared for U.S. Bureau of the Census Annual Research Conference. 1985.

Kasprzyk, D. and Herriot R. "The Survey of Income and Program Participation." Paper presented at the IASSIST Conference in Amsterdam, the Netherlands. May, 1985.

Mathematica Policy Research. "Food Stamp Research: Preliminary Results and Lessons for SIPP." Washington, D.C. Mathematica Policy Research. In preparation.

McMillen, D., and Herriot, R.A. "Toward a Longitudinal Definition of Households." SIPP Working Paper Series No. 8402. U.S. Bureau of the Census, Washington, D.C. 1984.

Samuhel, M. and Huggins, V. "Longitudinal Item Imputation in a Complex Survey." Proceedings of the Survey Research Methods Section, American Statistical Association. 1984.