THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

ENHANCED DEMOGRAPHIC-ECONOMIC DATA SETS

No. 82

R. Herriot, C. Bowie, D. Kasprzyk
Bureau of the Census

S. Haber
The George Washington University

U.S Department of Commerce BUREAU OF THE CENSUS

TABLE OF CONTENTS

Enhanced Demographic-Economic Data Sets

By Roger Herriot, Chester Bowie, Daniel Kasprzyk, and Sheldon Haber

This paper explores the possible development and uses of data sets that combine demographic data-both survey and population census-with economic census administrative information. It describes the 1984 Survey of Income and Program Participation (SIPP) and various pilot projects to augment the SIPP data with information about the establishments and firms for labor force analysis and with tax return information for income studies. The ability to add industry or labor market variables is also discussed.

The idea of augmenting survey data with information from other sources is not new. Such microdata record matches have a number of uses.

- They can add information that cannot be collected from survey respondents-information, such as the amount of the employer's contribution for the respondent's contribution to social security over a wormlike.

- They can add "contextual" variables about the area in which a person lives or works-variables such as a city's unemployment rate or a neighborhood's racial composition.

- They can provide direct comparisons for evaluation of the accuracy of a respondent's answers to the survey questions-for example, the amounts of wages or social security receipts reported by respondents can be compared with the amounts on administrative records.

- They can be used as weighting controls to calibrate the survey and to reduce the variance for many items.

- Finally, they can be used to replace a respondent's answers in some situations to improve accuracy or to model estimates using both survey and administrative data.

The SIPP was the first Census Bureau survey designed from the beginning to facilitate such matching activities. The SIPP, which began in 1983, was preceded by an 8-year development program-the Income Survey Development Program (ISDP). With respect to matching survey data to administrative records, the philosophy, attitudes, and plans of the ISDP strongly reflected the experience gained in a 1973 exact match study (Scheuren et al. 1975). A review of the work of the ISDP with regard to the use of administrative records can be found in Kasprzyk (1983) and Griffith and Kasprzyk 1980).

SIPP design features

The primary goals in designing the SIPP were twofold: (1) To improve the reporting of income and program-related data in a way that would allow the analysis of changes over time at a microlevel, and (2)to accommodate the collection of a large quantity of information in a flexible manner that allowed some information to be collected more frequently than other information. These goals were met principally by using a survey design in which the same people are interviewed more than once.

Persons (15 years old or older) in households selected for a sample panel are interviewed about their income and other topics once every 4 months for approximately 2 ½ years. These sample persons are interviewed at new addresses if they move, and any other persons that they move in with, or vice versa, are also interviewed. In this way, a highly detailed record is built up over time for each person and household in a sample panel. This design minimizes the need for sample persons to recall most of the information for more than a few months, and it reduces the number of questions asked in one interview.

To enhance the estimates of change, particularly year-to-year change, a new sample panel is introduced every year rather than at the conclusion of a panel. Consequently, two, or sometimes three, panels are in the field concurrently. The overlapping panel design allows cross-sectional estimates to be produced from a larger, combined sample that is about double in size when two panels overlap and about triple with three overlapping panels.

The reference period for the primary survey items is the 4 months preceding the interview; for example, for the February interview, the reference period is the preceding October through January. When the household is interviewed again in June, the reference period is February through May. To create manageable interviewing and processing work loads, the sample households within a given panel are divided into four subsamples of nearly equal size. These subsamples are called rotation groups, and one rotation group, or one-fourth of the sample, is interviewed each month. Thus, it takes 4 consecutive months to interview the entire sample. This 4-month period of interviewing is termed a "wave."

SIPP content

Each interview is planned to take about 30 minutes, and it includes content that is divided into three main groups of questions-control card items, core items, and topical module items.

The *control card* is used to list every person residing at an address and to record basic social and demographic characteristics (e.g., age, race, sex, and educational attainment) for each person at the time of the initial interview. At subsequent interviews, changes in these characteristics are recorded on the card, as well as the dates when persons enter or leave the household. Some information relating to the housing unit or household also is collected (e.g., number of units in the structure and tenure).

The *core* is a set of questions that are asked at the first interview and then updated in each subsequent interview. The core collects the basic data on labor force, income, and program participation for each of the 4 reference months. Among the items included in the core are the following.

- Information associated with wage and salary earnings - e.g., industry and occupation codes, hours and weeks of work, and hourly earnings for up to two jobs;

- Data associated with self-employment - - e.g., the type of business (farm or nonfarm), earnings, whether it was incorporated, the profits and losses from the business - - for up to two self-employment jobs;

- Data associated with nonearned income - - e.g., Aid to Families with Dependent Children, supplemental security income, general assistance workmen's compensation, social security and other retirement income, miscellaneous sources of income (such as alimony, child support, income for foster child care, and educational assistance), and noncash benefits (such as food stamps, Women, Infants, and Children Nutrition Program, medicaid, medicare, and health insurance coverage);

- Data associated with asset holdings - - e.g., income from savings accounts, bonds, stocks, and rental property - - for the 4-month reference period and both individual and joint recipiencies.

A wide variety of topics not covered in the core portion of the questionnaire are collected in *topical modules*. The module data may be analyzed independently of, or in conjunction with, the core data. The topics include many subjects, such as wealth, taxes, health, and personal histories (e.g., lifetime work experience, marriage, and education).

Collection and validation of social security numbers in the SIPP

The SIPP data system has always been thought of as a combination of data from administrative records and household surveys. To make these linkages accurate, social security numbers (SSN) are obtained for sample individuals. These numbers are then verified and corrected to maximize the number of accurate linkages to other record systems.

Persons who refuse to provide an SSN are not included in the verification process. The Social Security Administration identified (by machine validation) incorrectly reported numbers and then clerically resolves these cases along with those cases not reporting an SSN. This work is completed by the fourth wave interview, at which time a field followup is conducted to obtain missing SSN's (provided they are not refusals) and to reconcile inconsistencies in SSN or demographic data generated by the computer match or by the clerical resolution.

The following summarizes the SSN validation results from the wave 1 sample of the 1984 panel: Total wave 1 sample persons, 53,588; persons who refused to provide an SSN and were excluded from the validation process, 1,674; persons eligible for SSN validation, 51,914; validated SSN's (85 percent of those eligible), 44,172; unvalidated SSN's (mostly children who have no SSN), 7,742. Sater (1986) concluded that the SSN acquisition rate for persons who have SSN is between 93 and 97 percent.

Enhancing SIPP Data

SSA-SIPP data linkage project

This section briefly describes several areas in which matched data sets can extend the analytical potential of the SIPP. Interest in a data set that matches SSA administrative data with household survey data follows closely the intended uses of SIPP at its inception. A matched data set would enable researchers; (1) To estimate future costs for programs, such as the Old Age Survivor, and Disability Insurance (OASDI) Program and the Supplemental Security Income (SSI) Program; (2) to assess the effects of program policy changes on the economic well-being of participant families; (3) to describe nonprogrammatic characteristics of program participants; and (4) to test social science theories as they relate to the dynamics of social security programs.

In essence, the SSA-SIPP data linkage project involves a maximum linkage with SIPP. For each SIPP panel, modules - can be linked to extracts of the basic SSA program records, including the following: The Master Beneficiary Record, which contains eligibility and benefit histories of the OASDI program; the Supplemental Security Record, which contains eligibility and benefit histories for the SSI program; and the Summary Earnings Record, which contains a history of covered earnings for each worker. SSA records will be updated periodically so that each SIPP panel's files will contain additional years of the SSA's program data. We may also want to link SIPP to new disability administrative files that are now being developed at the SSA. All initial and subsequent linkages will be by mutual agreement between the SSA and the Bureau of the Census.

Employer-provided benefits feasibility study

Employer contributions to health insurance plans, retirement plans, and life insurance plans have recently been the focus of national attention by Congress, other policymakers, and researchers. SIPP collects information on whether a person is covered by health insurance and whether the employer makes contributions for health insurance, but it stops short of obtaining amounts for either the respondent's contribution or the employer's contribution. For life insurance, information is obtained on coverage, face value, and whether the policies are provided through an employer. Amounts of employee payments and employer contributions are not obtained.

The employer-provided benefits feasibility study involved obtaining a signed release from the respondent at the interview, contacting the respondent's employer, and asking the employer to fill out a short questionnaire. The questionnaire was designed to obtain data on both the employer's and the employee's contributions to the firm's health insurance, pension, and life insurance plans. One-half of one rotation group's households were selected for the study. The study was done in August 1987 using rotation group 4 in wave 8 of the 1985 panel. This was the last interview for these households.

The study included only employed persons, 18 years old or older, for whom a wave 8 interview questionnaire was completed. Of the 1,352 persons eligible for the test, 569 persons (42 percent) signed the authorization form, 446 persons (38 percent) refused to sign, and 337 proxy or telephone respondents (25 percent) did not return the authorization form that was left with or mailed to them. We did not conduct a followup of the refused or nonreturn cases, since the primary purpose of the test was to evaluate the process of collecting the information from the employer.

Of the 569 questionnaires that were mailed to an employer, 548 (96 percent) were completed and returned. A more detailed evaluation of the data collected in this study and an assessment of the future prospects for a study of this type on the complete sample will be undertaken next year.

Adding "contextual" variables

Summary information from the decennial census offers another way to enhance the SIPP data. Although SIPP offers a very rich set of data about persons, the only contextual information collected concerns their living arrangements (household and family characteristics) and the areas in which they live (state, urban, etc.).

Because SIPP's addresses are computer readable, it is possible to geocode them into 1990 census geography (census block, tract, and city) using the new "Tiger" geocoding system. With such identifiers added to the SIPP files, one could augment the data with various contextual variables that have been created from the census. For example, labor force analysis could be enhanced by knowledge of the unemployment rate in a labor market, and migration analysis could be improved by the inclusion of per capita income in formation in out-migration and in-migration areas. Adding such variables would permit the testing of additional hypotheses that could not otherwise be examined. Besides those variables relating to labor markets and income, numerous others relating to spatial areas and demographic, social, and educational variables could be incorporated into the SIPP data. We believe that adding such variables would greatly enhance the usefulness of the data, but there are no formal plans to add such variables at present. However, if there is sufficient interest, the capability to create such data exists.

Merging Economic Data with SIPP Demographic Data

During the first 2 years of the SIPP program, a good deal of background research was completed on the potential for augmenting SIPP data with microlevel establishment and enterprise data from the economic census and other data files maintained by the Bureau of the Census. The analytic potential of SIPP suggests the desirability of augmenting it with these data. Additionally, the marginal cost of merging these data with SIPP is relatively small, and the potential gain in knowledge is very large.

Besides the substantive knowledge to be gained by merging SIPP demographic data with economic data, merging these data sets makes it possible to verify the accuracy of the estimates given by respondents in survey data (for example, one could verify the respondent's estimate of the employment size of the firm for which he or she worked). An additional, indirect benefit of linking SIPP and economic data stems from the fact that the former is a representative sample of the working population. Accordingly, the probability of an establishment's employee being included in SIPP is inversely related to its employment size; estimates of the population of establishments in each establishment-size group can be derived from the number of SIPP respondents employed in each group and the SIPP respondent population weights. These advantages, plus the manageable size of the SIPP sample, should result in valuable insights into how the size distribution of establishments is changing over time and the economic implications stemming from this change.

SIPP and the economic data files

In merging SIPP demographic data and economic data, it is necessary to know the information contained in the various files to be linked and how each file is constructed. Three data sets that might be incorporated into a SIPP-economic data file are the Standard Statistical Establishment List (SSEL), the Longitudinal Research Database (LRD) file, and the enterprise statistics (ES) file.

The SSEL is a complete directory of establishments in single-establishment and multiestablishment enterprises with one employee or more, irrespective of industry. The SSEL links parent companies, subsidiaries, and their establishments. It contains information on approximately 4.7 million enterprises and 5.7 million establishments.

The SSEL is important because it is a current file containing a complete list of establishments and companies that have paid employees. Although the SSEL contains only a narrow range of economic data, these data impart information not found elsewhere. For example, the SSEL contains the addresses of the physical locations of establishments; this information is useful for merging the demographic and economic data, since the addresses are a primary way of identifying an individual's place of work. Employment and payroll figures yield an estimate of average annual earnings, thereby indicating whether an employer is a low-or high-wage employer.

Sales and employment figures provide a proxy measure of productivity. Operational status information can be utilized to identify those establishments that have become inactive. Additionally, the SSEL contains longitudinal information. Currently, establishment and company data are carried for 2 years in the SSEL .

The LRD is a longitudinal micro database containing data at the establishment level from the Annual Survey of Manufactures and the Census of Manufactures. The LRD provides a broader range of information about establishments than the SSEL. For each manufacturing establishment, value added per production worker, which is a proxy for labor productivity, can be calculated. For the larger establishments with 250 workers or more, information is available on depreciable assets and rented machinery so that capital-to-labor ratios can be computed. Additionally, a measure of labor compensation, including fringe benefits, can be obtained.

Like the Census of Manufactures, the ES data re collected every 5 years. These data cover enterprises whose primary activities are in in-scope industries. For each enterprise, the data are consolidated from all operating units. The information contained in the ES is similar to that in the Census of Manufactures; however, the ES contains fringe benefit, asset, and related data only for companies with 500 workers or more. Haber (1985) has presented a detailed accounting of the economic files-their universe restrictions, data content, and applications-when merged with the SIPP data.

Some applications of microdemographic and economic data

In this section, two applications of a SIPP-economic data file are discussed to illustrate the use of this data set.

Low-wage workers and low-wage firms. - Although survey data, such as data from the Current Population Survey, provide insights into the characteristics of low-wage workers, they provide little information about low-wage firms. A number of hypotheses have been formulated about how production is organized in low-wage and high-wage firms. For example, to the extent that high-wage firms are capital intensive, their need for trained workers is likely to be greater than that of low-wage firms (Oi 1983). To reduce turnover, which disrupts the production process, high-wage firms are also more likely to substitute future benefits in the form of pensions for current benefits in the form of wages (Lazer 1981). A SIPP-economic data file would permit verification of these, and related, hypotheses.

There are two other questions that could be explored. To what extent are the differences in individual earnings in low-and high-paying firms due to the characteristics of the workers and of the capital employed in each type of firm? And, to what extent are workers with a similar characteristics (i.e., skills or training) remunerated in the same way in each type of firm?

Structural unemployment. - An issue of long standing is what happens to workers who are displaced from their jobs as a result of structural changes. How long do they remain unemployed

vis-a-vis other workers who separate from an employer? What sources of income, including cash and non-cash government transfers, do they draw on when they are unable to find a new job? Whey they find a new job, how do the earnings in the new job compare with those in the old one? If there is an earnings loss, how much of this loss is recouped after, for example, 2 years?

One way of identifying structurally unemployed workers is to ascertain whether a firm has closed down or has undergone a substantial decline in employment. A SIPP-economic data file would enable one to determine the extent to which firms are subject to sever, long-term shocks, as evidenced by plant closures or substantial reductions in employment, and to determine how such shocks affect their work forces.

Methodological issues in matching SIPP demographic and economic data

In this section, attention is focused on two methodological problems. The first problem deals with procedures for linking worker data to establishment and company data. The second related to the estimation of data-in particular, asset and fringe benefit data that are collected for large establishments and companies, but generally not for small ones.

Essential to the creation of a SIPP-economic data file is the ability to determine the establishment and company in which a person is employed. The most promising, and least expensive, way of accomplishing the SIPP-economic data link is to use the employer identification number (EIN). A promising source of EIN information for SIPP respondents is the employer-provided benefits questionnaire discussed previously. In the feasibility test of this questionnaire, the EIN was asked for and was well reported.

Without the EIN, for employers with only one establishment in an area, the firm name and the employee's address will typically be sufficient to determine where a person is employed. This information is available in SIPP and the SSEL. For companies with more than one establishment in an area, the firm name, address, census industry code, and the respondent's estimate of size establishment and company can be used to identify a person's workplace. In the event that a unique workplace cannot be determined for a multiestablishment firm, an employer's characteristics can be imputed. For example, data from the SSEL on number of employees and on payroll can be averaged over a company's establishments in a local area. When it is not possible to identify a worker's firm by name in the SSEL, imputations can be made by averaging over establishments in the same local area with the same census industry codes as that of the given employer.

Imputation can also be made for variable not contained in the SSEL. For example, the average capital-to-labor ratio for a large firm with a chain of fast-food stores can be used as an estimate of the capital-to-labor ratio for each store in the chain. It should be noted that information on capital stock is not generally available for small establishments. However, such information is available for a large sample of small establishments in manufacturing, and it can be utilized in an economic model to obtain capital stock estimates for all small manufacturing

establishments. For example, it is plausible that an establishment's capital-to-labor ratio is positively related to its use of purchase electricity per employee. The latter ratio could then be used to infer the former.

Match of the decennial census to the economic census and surveys?

We have included a question mark in this title because very little work has been done in this area. However, it appears that advances in the decennial census procedures permit such a possibility. Block-level geographic coding of place of work and automated industry and occupation coding using the company name suggest the possibility of forming labor force statistics from the decennial census for establishments, enterprises, and five-digit Standard Industrial Classification industry codes.

The resulting record would contain the data about the economic unit from the economic surveys and the data about the labor force (e.g., age, race, occupation, and education) in the economic unit from the decennial census.

A large-scale application of this idea is probably not possible for the 1990 census. However, if initial research could begin and if limited studies for use in a particular industry or area could be done, then it would be more likely that a much more ambitious program could be designed in the future. At this point, we need guidance from researchers about the potential uses of this file so that we can concentrate our efforts in the most productive manner.

Issue of Data Access

Given the confidentiality requirements of Title 13, data access to these enhanced files is a major issue. There are several avenues available for a researcher to gain access to such files. One could obtain a permanent staff position at the Census Bureau, or one could obtain a temporary position (for example, when an sabbatical) or a postdoctoral position in the Center for Demographic Studies or the Center for Economic Studies. One could become a research follow at the Bureau to work on a specific proposal involving these data sets. One could enter into a Joint Statistical Arrangement with the Census Bureau and access the data at one of the Bureau's regional offices, or one could access the data through the Data Resource Center (DRC). Options for access are discussed in detail in Gates (1988).

The DRC provides a new capability that we are experimenting with at the Bureau. Its goal is to improve access to data collected by the Census Bureau and, thus, to cilitate analysis and research. The initial activities of the DRC were concerned with assisting Census Bureau personnel with access to SIPP. Its long-term mission will expand to include the following activities: To create and manage enhanced data sets, to provide liaison between internal and external users for access to such data, and to review the outputs for confidentiality. The process for gaining access to DRC data by outside users has been sketched out by Cavanaugh (1987).

The Census Bureau is also exploring new ways to make the information content of enhanced data files publicly available. We are experimenting with new products that could substitute for the original microdata file in cases where the disclosure risk is great. One approach is microaggregation in which individual records are grouped according to specified criteria and responses are replaced with averages for the group (McGuckin and Nguyen 1988). This approach, which is operationally straightforward, has been suggested as a way to provide access to sensitive economic microdata (Govoni and Waite 1985). The primary objection to this approach is that the linking of "like" establishments is dependent on the grouping criteria.

Another aggregation approach that are considering for more general application is the release of summary statistics, such as variance-covariance measures or correlation matrices of the data. Such files would contain all the information needed for linear regression analysis; they would also provide excellent confidentiality protection, since any given covariance matrix can be derived from an infinite number of data sets. The biggest disadvantage to this approach is that different users require different matrices, and a user may require new columns in a matrix as the analysis proceeds.

The Census Bureau confidentiality staff is also currently looking into microaggregation and data transformation as techniques to allow the release of economic microdata.

In conclusions, we hope that this paper will stimulate researchers to investigate new hypotheses and to reexamine old ones. Although the research suggests that the creation of such data sets is feasible, the Bureau will need to work with interested researchers to develop the required work with interested researchers to develop the required data for substantive analyses. We are taking the initial steps of restructuring the Center of Demographic Studies to support such activities and to continue research, but it will take a concerted effort by both producers and users to make the potential a reality.

References

Burkhead, D., and Coder, J. (1985). "Gross Changes in Income Recipiency from the Survey of Income and Program Participation." *Proceedings of the Social Statistics Section, American Statistical Association,* 1985: 351-356.

Cavanaugh, F. (1987). "SIPP as an Initiator of a Data Resource Center at the Census Bureau." *Proceedings of the Statistical Computing Section, American Statistical Association,* 1987: 232-237.

Gates, G. (1988). "Census Bureau Microdata: How to Provide Useful Research Data While Protecting the Anonymity of Respondents." *Proceedings of the Social Statistics Section, American Statistical Association.* Forthcoming.

Govoni, J. and Waite, P.J. (1985). "Development of a Public Use File for Manufacturing." Paper presented at the Joint Statistical Meetings, American Statistical Association, Las Vegas, Nevada, August 1985.

Griffith, J. and Kasprzyk, D. (1980). "The Use of Administrative Records in the Survey of Income and Program Participation." Case study in *Report on Statistical Uses of Administrative Records.* Statistical Policy Working Paper No. 6. Washington, DC: GPO.

Haber, S. (1985). *Applications of a Matched File Linking the Bureau of the Census Survey of Income and Program Participation and Economic Data.* Survey of Income and Program Participation. Working Paper Series, No. 8502, U.S. Bureau of the Census, Washington, DC: GPO.

Herriot, R. (1983). "The Use of Administrative Records in Social and Demographic Statistics." Paper Presented at the meeting of the International Statistics Institute, madrid, Spain, 1983.

Huggins, V. (1987). "Research Plans." Memorandum for the Record, Statistical Methods Division, U.S. Bureau of the Census, April 13, 1987.

McGuckin, R.H., and Nguyen, S.V. (1988). "Use of 'Surrogate Files' to Conduct Economic Studies with Longitudinal Microdata." Paper presented at the 1988 Fourth Annual Research Conference, Arlington, VA, March 20-23, 1983.

Kasprzyk, D. (1983). "Social Security Number Reporting, the Use of Administrative Records, and the Multiple Frame Design in the Income Survey Development Program." *Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program (ISDP).* Edited by M. David. New York: Social Science Research Council, 1983: 123-141.

Lazear, E. (1981). "Agency Earnings Profiles Productivity and Hours Restrictions." *American Economic Review* (September 1981): 606-620.

Moore, J., and Kasprzyk, D. (1984). "Month-to-Month Recipiency Turnover in the ISDP." *Proceedings of the Section of Survey Research Methods, American Statistical Association,* 1984: 726-731.

Moore, J., and Marquis, K. (1987). "Using Administrative Record Data to Evaluate the Quality of Survey Estimates." Paper presented at the International Symposium on the Statistical Uses of Administrative Records, Ottawa, Canada, November 23-25, 1987.

Oi, W. (1983). "The Fixed Employment Costs of Specialized Labor." *The Measurement of Labor Costs.* Edited by Jack E. Triplett. Chicago: University of Chicago Press, for the National Bureau of Economic Research. 1983: 63-116.

Sater, D.K. (1986). "SSN Response Rates and Results of SSN Validation/Improvement Operation." Memorandum for Roger Herriot, Population Division, U.S. Bureau of the Census, March 11, 1986.

Scheuren, F. (1983). "Design and Estimation for Large Federal Surveys Using Administrative Records." *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 1983: 377-381.

Scheuren, F., Herriot, R., Vogel, L., Vaughan, D., Kilss, B., *Exact Match Research Using the March 1973 Current Population Survey - Initial States.* Studies from Interagency Data Linkages, Report No. 4. U.S. Department of Health, Education, and Welfare Social Security Administration, Office of Research and Statistics. Publication No. SSA 76-11750.