THE SURVEY OF INCOME AND
PROGRAM PARTICIPATION

A REVIEW OF THE USE OF ADMINISTRATIVE RECORDS IN THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

No. 43

C. Bowie and D. Kasprzyk
Bureau of the Census

U. S. Department of Commerce BUREAU OF THE CENSUS

Survey of Income and Program Participation

A Review of the Use of Administrative Records in the Survey of Income and Program Participation

by

Chester Bowie and Daniel Kasprzyk
Bureau of the Census

No. 8721 — 43

November 1987

Table of Contents

A Review of the Use of Administrative Records in the Survey of Income and Program Participation

Chester Bowie and Daniel Kasprzyk

INTRODUCTION

The Survey of Income and Program Participation (SIPP) is intended to provide comprehensive information on the economic resources of the American people and on how public transfer and tax programs affect their financial circumstances. The data from the SIPP are expected to provide government policy makers with an information base for studying the efficiency of government tax and transfer programs, for estimating future program costs and coverage, and for assessing the effects of proposed policy changes.

The SIPP provides comprehensive information about annual and sub-annual income and participation in public and private transfer programs for the household population in the United States; it also devotes considerable attention to measuring various kinds of economic resources other than current cash income. The most important elements of this broader perspective are the SIPP data on assets, debts, and non-cash resources, such as means-tested housing benefits, publicly and privately provided health insurance, pension coverage, and other employee fringe benefits.

The SIPP arose in response to the recognition that the principal source of information on the distribution of household and personal income in the United States--the March Income Supplement of the Current Population Survey (CPS)--had limitations which could only be rectified by making substantial changes in the survey instrument and procedures. One of the limitations of the CPS was the inability to provide linkages to administrative record data for statistical purposes. Recognizing this limitation of the CPS and the analytic usefulness of linking survey data to administrative records, the designers of the SIPP explicitly stated the ultimate SIPP data system should be a combination of data from administrative records and household surveys linked through the Social Security Number. The goals of the SIPP as described by Lininger (1980) state that administrative records will be used to:

1. increase sampling efficiency for certain subpopulations (e.g., Old Age, Survivors and Disability Insurance recipients or Supplemental Security Income recipients);

2. compare with survey data for validation studies of items common to both sources; and

3. supplement survey reported data with administrative record data for items difficult to obtain in a survey (e.g., earnings and program benefit histories).

These goals manifest themselves first in the SIPP development program and then in the SIPP.

This paper describes the SIPP's continuing commitment to the use of administrative records for statistical purposes: Section I reviews the work of the research and development program preceding the SIPP with regard to the use of administrative record data; Section II describes the design and content of the SIPP and its program to obtain accurate reporting of the Social Security Numbers (SSN) to facilitate linkages between survey reported data and administrative record sources; Section III describes five areas of applications where the SIPP project has initiated work involving the use of administrative records for statistical purposes; and Section IV provides a few additional examples of potential administrative data--SIPP linkage possibilities.

I. The Use of Administrative Records in the ISDP

The Income Survey Development Program (ISDP), authorized in 1975, was a program whose goal was to develop methods and a survey design to overcome underreporting and misclassification problems in the CPS (Ycas and Lininger, 1981). Furthermore, the ISDP also developed procedures and methodology for improving the collection of SSNs. The philosophy, attitudes, and plans of the ISDP strongly reflected the work of Scheuren and his colleagues (Scheuren et al, 1975) in the development of the 1973 Exact Match File (Kilss and Scheuren, 1978). 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). A brief summary of this work will provide a context for the discussion of the SIPP experience and the plans and potential for the statistical uses of administrative records in the SIPP.

In the ISDP the collection and accurate reporting of the Social Security Number (SSN) from each person in sample was deemed essential to the program. By emphasizing the collection of the SSN and then developing a system to validate and correct reported SSN's, 95.5% of the total cases were identified as having a correct SSN (Kasprzyk, 1983). The system served as a prototype for the SIPP system which is described in the next section.

The ISDP consisted of four experimental field tests which were conducted to examine different concepts, procedures and questionnaires. One aspect of each of these field tests was the use of an administrative record frame for sampling purposes. Even though the principal thrust of this approach was to increase sampling efficiency for selected subpopulations through the use of multiple frame estimators, the most important result was that these feasibility studies provided an opportunity for the survey planners to understand the administrative, methodological and operational difficulties in using administrative sources for sampling.

During the ISDP the following administrative record sources were used: 1) the Aid to Families with Dependent Children (AFDC) master file maintained by the Texas State Department of Welfare[1/], the Supplemental Security Record (SSR)[2/], the Master Beneficiary Record (MBR)[3/], the Basic Educational Opportunity Grant (BEOG) applicant file [4/], the Veterans Administration Pension and Compensation file[5/], the Internal Revenue Service Individual Master File [6/], and State record files for Unemployment Insurance and Workers Compensation.

The ISDP also effectively used administrative records to clarify misreporting and nonreporting of program benefits by comparing the survey reported data with the administrative data. Vaughan (1978) and Goudreau, Oberheu and Vaughan (1981, 1984) report on ISDP studies which led to redesigning questionnaires in order to reduce errors in classifying sources of income.

Finally, although the ISDP intended to create a data base augmented with administrative data which were difficult to obtain in a household survey it never did. A planned match to the Summary Earnings Record [7/] was never implemented because of higher priority projects.

The next section provides a very general summary of the SIPP survey design and content, followed by a description of the program established to collect and validate Social Security Numbers.

II.1. SIPP Design Features

The primary goals in designing the SIPP were to improve reporting of income and other program-related data and to do it in a way that would allow the analysis of changes over time at a microlevel. The design also had 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 of age or older) at households selected for a sample panel are interviewed about their income and other topics once every 4 months for approximately 2 1/2 years. 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 longer than a few months and reduces the number of questions asked in one interview.

To further enhance the estimates of change, particularly year-to-year change, a new sample panel is introduced every year instead of at the

conclusion of a panel. Consequently, two or sometimes thre panels are in the field concurrently. The overlapping panel desi al s cross-sectional estimates to be produced from a larger, co ne. sample that is about double in size when 2 panels overlap d triple with 3 overlapping panels.

The first SIPP panel, designated as the 1984 Panel but f lded in October 1983, started with approximately 20,000 intervie ed households. The second panel, i.e., the 1985 Panel, began in February 1985 with around 14,000 interviewed households. Panels of about 12,300 interviewed households are expected to be fielded every February. The sample size changes in each wave of a panel due to losses through attrition and gains from following movers to new households.

The reference period for the primary survey items is the 4 months preceding the interview; for example, in February, 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 each month instead of one large work load every 4 months, 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 called a "wave."

II.2. SIPP Survey Content

Each interview is planned to take about 30 minutes of a respondent's time and includes content that is divided into three main groups of questions. The substance of two of these groups should be essentially the same for each wave and for each panel. The third group of questions covers topics that will change in each wave of a panel. This allows for the inclusion of some new content in each panel, although many of the topics will be repeated across all the panels. Each rotation group in a wave is administered the same set of questions although the reference period is different as explained above.

The first group of questions are control card items. The control card is a separate document from the questionnaire and serves several important functions. The control card is used to list every person residing at an address and to record basic social and demographic characteristics (age, race, sex, and so forth) for each person at the time of the initial interview. Some information relating to the housing unit or household also is collected; e.g., number of units in the structure, tenure, and so forth. The card is reused at subsequent

interviews to record changes in characteristics such as age, educational attainment, and marital status, and to record the dates when persons enter or leave the household. Finally, during each interview, information on each source of income received and the name of each job or business is transcribed to the card so that this information can be used in the updating process at the next interview.

The second major group of questions form the core portion of the questionnaire, which is divided into 5 sections. The core set of questions is asked at the first interview and then updated in each subsequent interview. The first section of the core collects the basic labor force participation data for the 4 reference months.

In addition, this first section of the core collects much of the information on the receipt of income from various sources during the 4 month reference period. This includes income from government sources such as Aid to Families with Dependent Children, Supplemental Security Income, General Assistance, and Workmen's Compensation. Respondents are also asked about both Social Security and other retirement income including Railroad Retirement, pension from company or union, and civil service retirements, as well as others. The receipt of miscellaneous sources of income such as alimony, child support, interest from savings, income for foster child care, and educational assistance is also identified. In addition, questions on major sources of noncash benefits such as food stamps, WIC (Women, Infants, and Children Nutrition Program), Medicaid, Medicare, and health insurance coverage are included in this section.

The second section of the SIPP core questionnaire collects information associated with wage and salary earnings. This section includes information on industry and occupation as well as hourly earnings for up to two jobs.

The third section of the core collects data on self-employment earnings and specific information about the kind of self-employment--whether it was incorporated, sole proprietorship, or partnership--and the profits and losses from the business. Again, space is provided for two self-employment jobs.

The fourth section is identified as the general amounts section. This section of the questionnaire collects monthly amounts received from the income sources identified in the first section. That is, the first section identifies the receipt of income during the 4 month reference period, while amounts of income received are collected in the fourth section of the questionnaire. Space is provided for amounts from up to six income sources.

The fifth and last section of the core questionnaire collects amounts of income earned from asset holdings. Asset sources include savings accounts, bonds, stocks, and rental property, as well as others. Information is collected for the 4 month reference period on both individual and joint recipiency.

The third major question grouping consists of the various supplements or topical modules that are included in waves following the initial interview. A wide variety of topics are covered under the aegis of the topical module concept. The breadth of these data ensure that SIPP will be a widely used and powerful data base serving multiple purposes. The administration of a module is possible in Waves 2 through 8 (or 9 in 1984) because less time is required to update the core information after the first interview. Depending on the time available and length of the modules, more than one may be administered in the same wave. The topical modules cover areas that do not require examination every 4 months and may use a different reference period than the core questions. Some modules are assigned to only one wave of a panel, while other modules may be repeated in more than one wave. The modules provide a broader context for analysis by obtaining information on a variety of topics not covered in the core portion of the questionnaire. The module data may be analyzed independently or in conjunction with the control card items or core data. Frequently, a module is administered at the same time in concurrent panels so that the data may be combined to improve reliability.

II.3. The 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. This reduces respondent burden by using other data sources for difficult-to-obtain information. Interview responses can be supplemented by information from program files such as the earnings and benefit records of the Social Security Administration (SSA). This allows, for example, analysis of the long-term impact of various Social Security benefit formulas.

To make these linkages accurate, Social Security Numbers (SSN) are required for sample individuals. The SSN is obtained for each household member in SIPP and recorded on the control card. It is identified as a critical survey data item requiring completion to make the interviewers aware of its importance. These numbers are then verified and corrected to maximize the number of accurate linkages to other record systems.

The verification and correction process builds on the work of the development program (Kasprzyk, 1983). At the conclusion of each month's interviewing during the first wave of a SIPP panel, a special extract file is prepared by the Census Bureau for the SSA. This file

contains a small number of key variables (SSN, name, date of birth, sex) for all original sample persons who report a SSN, including children, in a format appropriate for machine validation. Persons who report that they do not have a number or have a number but cannot supply it are handled separately in a clerical (manual) procedure. Persons who refuse to provide a SSN are not included in the search process. The SSA identifies (by machine validation) incorrectly reported numbers then clerically resolves these cases along with cases not reporting a SSN. This work is completed by the fourth wave interview, at which time a field followup is conducted to obtain missing SSNs (provided they are not "refusals") and to reconcile inconsistencies in SSN or demographic data generated by the computer match or the clerical resolution.

Social Security Numbers of persons who enter the sample after Wave 1 (because they start living with original sample people) are validated at the start of the next panel. For example, information on new panel members (nonsample persons) from Waves 2 through 5 of the 1984 Panel was held and submitted for computer validation with Wave 1 of the 1985 Panel. Likewise, information on nonsample persons from Waves 6 through 8 of the 1984 Panel and Waves 2 through 4 of the 1985 Panel were held and submitted for computer validation with Wave 1 of the 1986 Panel.

The following summarizes the SSN validation results for the 1984 Panel Wave 1 sample:

53,588	Total Wave 1 sample persons
- 1,674	Persons who refused to provide a SSN and were excluded from the validation process
51,914	Persons eligible for SSN validation
-42,128	Persons who reported a usable SSN and were eligible for computer validation
9,786	Persons who did not report a SSN and were eligible for the manual search (mostly children)

44,172	Validated SSNs (85% of eligible)
7,742	Unvalidated SSNs (mostly children who have no SSN)
51,914	Eligible for SSN validation

Based on these results, Sater (1986) has concluded that the SSN acquisition rate for persons who have a SSN is between 93 and 97 percent.

The next section briefly describes five areas of application where work has begun to use the survey data and administrative record data in some capacity: 1) SIPP/SSA data linkage project; 2) Employer Provided Benefits Study; 3) SIPP Record Check Study; 4) the use of Administrative Records in SIPP Estimation; and 5) merging economic data with SIPP demographic data.

III.1. SSA/SIPP Data Linkage Project

SSA's interest in a data set which merges administrative data with household survey data follows closely the intended uses of SIPP at its inception. A merged data set would enable the SSA to:

1. Estimate future program costs -- The SSA is responsible for projecting program costs for all major SSA programs including: The Old Age, Survivor, and Disability Program, the Supplemental Security Income Program and Aid to Families with Dependent Children. In order to improve the accuracy of the projection methods, the SIPP panel data can be linked with a number of years of SSA data so that inflows and outflows can be analyzed in addition to point-in-time prevalence estimates of SSA program participation. The relationship between program participation and underlying individual characteristics can then be used to estimate future program costs and growth thus providing the SSA an early forecasting capability.

2. Assess the effects of program policy changes -- An SSA-SIPP linkage will contain family, income and SSA benefit data. This combination of information will permit the SSA to estimate the programmatic costs of policy changes that depend on these factors and to assess the effects of policy changes on the economic well-being of program participants.

3. Describe non-programmatic characteristics of program participants -- The SSA is frequently asked by Congress and others to provide information about program participants that is not routinely captured by administrative record systems. In the past, the SSA has used a series of widely spaced and usually one-time surveys to provide such information. Since the prospects for a new round of special purpose surveys are not good, an ongoing SSA-SIPP data link would provide relatively up-to-date data on a routine basis.

4. Test social science theories as they relate to Social Security programs -- The longitudinal component of the SIPP's research design and the wealth of data captured in core questions and topical modules provide data that will be sufficiently rich to test many social and economic theories of program participation,

thus making a significant contribution to the basic research that must accompany any dynamic social program.

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

The merged data set will reside on the computer at the Census Bureau and will be used only for general statistical research. Only Census Bureau staff and SSA employees who are designated as Census Bureau Special Sworn Employees will have access to the file. The SSA may publish statistical data in a summary form that does not permit the identification of a household, family, or individual.

The primary tasks in the linkage project are:

1. Verification of, and searching for, Social Security Numbers (SSN) -- This task is already a part of the SIPP project activities and was described earlier. In particular, the vast majority of SSN's for the 1984 SIPP panel have been processed by SSA staff.

2. Obtaining SSA administrative records -- As mentioned above, this project involves matching the MBR, SSR and SER to the SIPP. Decisions will have to be made about the content of the data extracts from these files that would be included in the match.

3. Merging administrative records with SIPP survey data -- The matching tasks are not one-time activities. Instead we anticipate a number of data processing operations for each SIPP panel.

4. Weighting, imputation and sampling error estimation -- We will have to consider and develop schemes for weighting and imputation that take into account non-matched SIPP records. Both cross-sectional and longitudinal weights will be required. The SSA would also need the capability for estimating sampling errors.

5. Development of documentation for the matched files -- Documentation for a matched file would include tape description and utilization information, the SIPP questionnaires and descriptions of the SSA administrative records, a sampling statement, imputation descriptions and any other information required for estimation or analysis.

III.2. 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 on the part of Congress, other policy makers, and researchers in areas such as health care, the elderly, and tax reform. SIPP collects information on whether a person is covered by health insurance and whether the employer makes contributions, but 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 policies are provided through an employer. Amounts of employee payments and employer contributions are not obtained.

This study involves obtaining a signed release from the respondent at the interview and contacting the respondent's employer and asking the employer to fill out a short questionnaire to obtain data on both the employer's and employee's contributions to health insurance plans, pension plans, and life insurance plans. Information provided by the employer would supplement the SIPP data.

A half sample of one rotation group's households was selected for the study. The test was done in August 1987, (rotation group 4) for households in Wave 8 of the 1985 Panel. This was the last interview for these households.

The test included only employed persons, 18 years old and 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 (33 percent) refused to sign, and 337 proxy or telephone respondents (25 percent) did not return the authorization form that was left/mailed to them. We did not conduct a followup of the refused or non-return cases.

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

III.3. SIPP Record Check Study

Another area of research with respect to administrative record systems is the development of validation studies of items common to both the survey and administrative records. The purpose of the study is to investigate response quality issues in SIPP through a case-by-case comparison of SIPP data and administrative record information. The ultimate goal is the improved understanding of the quality of the SIPP data and, ultimately, the development of quantitative estimates of response and nonresponse errors for the purposes of adjusting survey data or modifying survey procedures to obtain better quality survey data.

An overview and progress report of the study can be found in Moore and Marquis (1987). Simply put the study intends to address the following questions:

1. The quality of the respondent reports of receipt of program benefits for a variety of state and Federally administered transfer programs;

2. The quality of benefit dollar amount reporting for these programs;

3. Demographic correlates of report quality;

4. Extent of misclassification errors;

5. The (nonexperimental) effects of self-proxy respondent status on report quality; and

6. Between wave recipiency turnover effects (The "seam" problem (Burkhead and Coder, 1985; Moore and Kasprzyk, 1984)).

The questions will be addressed by using administrative record information for receipients of each of nine government transfer programs in four states--Florida, New York, Pennsylvania, and Wisconsin. These are four state-administered programs (Aid to Families with Dependent Children, food stamps, unemployment compensation, and worker's compensation) and five Federally-administered programs (Civil-Service Retirement, Pell Grants, Old Age Survivors and Disability Insurance (OASDI), Supplemental Security Income, and Veterans' Pensions and Compensation) which will be studied. The project has obtained a great deal of information on acquiring administrative record systems, learning about each systems idiosyncrasies, and generalized matching procedures at the Census Bureau. Some very preliminary results are now available in Moore and Marquis (1987).

III.4. Use of Administrative Records in SIPP Estimation

Information on the effect of sample reductions on the variance of estimates and on our ability to measure changes in differences in the number of statistics have created serious concerns. These concerns have caused us to increase our exploration of ways to reduce the variance. One approach is through the use of administrative records for post-stratification. Currently, cross-section estimation procedures for SIPP make use of a second-stage adjustment to increase the precision of estimates by ratio adjusting collection month and reference month estimates to population estimates. However, the Census Bureau has access to some Internal Revenue Service and Social Security Administration files which can be used to produce detailed age, race, and sex distributions by adjusted gross income. The issue, which we have just begun to explore, is how these administrative data can be used for post-stratification to improve estimates of mean and median personal and household income as well as the estimates of the deciles of the personal and household income distribution. Furthermore, a basic question which will be considered is how much reduction in the variances of these estimates can be achieved through such a procedure. These issues will be investigated next year.

The first phase of this research (Huggins, 1987) will estimate the reductions in variances of SIPP estimates by using the IRS data as auxilliary variables in the estimation procedures. The procedure being studied has been advocated by Herriot (1985) and Scheuren (1983). In the SIPP study the estimation method will involve a ratio adjustment of SIPP estimates at the second stage of estimation in cells defined by age + race + sex + "income" where "income" is adjusted gross income as reported to the Internal Revenue Service.

Controls are prepared from a 1% sample of 1984 IRS file matched with age, race, and sex characteristics from the Summary Earnings Record; adjusted gross income from the 100% IRS file is matched to a file of SIPP data. The SIPP cases are then reweighted by controlling to the 1984 IRS controls; that is, a factor f_j, which is the ratio of IRS control in cell$_j$ to the SIPP estimate of persons matched to IRS data with 1984 IRS income in cell$_j$, is applied to persons who fall in cell$_j$ based on the IRS data. Estimates and variances of selected SIPP characteristics will be obtained using the newly created weights and with the weights which do not use this procedure.

III.5. Merging Economic Data with SIPP Demographic Data

During the first two years of the SIPP program a good deal of background research was completed on the potential for augmenting SIPP data with micro-level establishment and enterprise data from the economic census and other data files maintained by the Bureau of the

Census (Haber, Ryscavage, Sater, Valdisera, 1984). Haber (1985) has described the analytic potential of matching economic data to the demographic data for individuals in the SIPP. Haber suggests that new insights are possible in the following areas: the relationship between capital and wage rates, the study of labor mobility between low and high-wage employees, studying implications of the transition from goods-producing to a service economy, and analyzing the effects of unions on the labor market. A pilot project was initiated to investigate methodologies for merging individuals in the SIPP (who report their employer's name) to the employer data in the economic census, testing the methodology to identify problem areas and solutions, and conduct the match for a pilot sample. Sater (1985) describes the project, and problems encountered. Unfortunately, due to costs, higher priorities, and staffing limitations, this project was never completed.

IV. Potential Linkage to Other Administrative Data Sets

The SIPP is a relatively new continuous survey, collecting a comprehensive socio-economic portrait of the U.S. household population. As mentioned above, the SIPP also gives substantial attention to the correct reporting of Social Security Numbers. These two elements together provide the principle reasons for the power of the data set. In the future, the good match variable (SSN) which the SIPP provides could be used in matching the survey data to the Health Care Financing Administration's Health Insurance Master File (Medicare) to study the relationship between hospital use, health status, employment and income. Similarly, the SSN will allow linkage of deceased respondents to the National Death Index. In the latter case, numerous SIPP panels would be necessary to have sufficient sample for analysis. Nevertheless, the potential for such linkages exist. In fact any linkage with an administrative record system which uses the SSN as the primary identifier is possible. The principal difficulties, however, are the costs for such projects and the difficulty of sharing matched administrative-survey data with all researchers.

FOOTNOTES

1/ The Aid to Families with Dependent Children (AFDC) master file is an administrative system maintained by each individual State and containing data on benefit amounts, payment history, demographic characteristics, and other information needed to administer the program.

2/ The Supplemental Security Record (SSR) is the national master administrative file for data on Supplemental Security Income (SSI) benefit amounts, payment history, and demographic data.

3/ The Master Beneficiary Record (MBR) is the national master administrative file for data on the Old Age, Survivors, and Disability Insurance Program (Title II); it contains current and historical program information on claimants for Title II benefits, past and present cash beneficiaries, disallowed claimants, and denied claims.

4/ The Basic Educational Opportunity Grant (BEOG) file is an administrative file maintained by the Department of Education. It contains data for all applicants of a given academic year, including ineligibles, eligibles who did not use their grant, and eligibles who used their grant. The BEOG program is now called the Pell Grant program.

5/ The Veteran's Administration Pension and Compensation File is a national master file containing records of benefits provided as disability compensation, dependency and indemnity compensation, disability pension, death pension, or burial allowance.

6/ The Internal Revenue Service Individual Master File (IMF) is a national file of selected income and tax information from all individual Income Tax Returns pertaining to wages, dividend and interest income, taxes paid, and exemptions.

7/ The Summary Earnings Record (SER) is a file containing the lifetime covered earnings (up to the maximum for each employer) and quarters of social security coverage of the individual. It is used to determine entitlement to benefits and calculation of benefit amounts. Individuals are identified in this file by their Social Security Number.

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, 351-356

GOUDREAU, K., OBERHEU, H. and VAUGHAN, D. (1981). An Assessment of the Quality of Survey Reports of Income from the Aid to Families with Dependent Children (AFDC) Program. Proceedings of the Section on Survey Research Methods, American Statistical Association, 377-382.

GOUDREAU, K., OBERHEU, H. and VAUGHAN, D. (1984). An Assessment of the Quality of Survey Reports of Income from the Aid to Families with Dependent Children (AFDC) Program. Journal of Business and Economic Statistics,179-186.

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 6. U.S. Government Printing Office, Washington, D.C. 20402.

HABER, S., RYSCAVAGE, P. SATER, D., and VALDISERA, V. (1984). Matching Economic Data to the Survey of Income and Program Participation: A Pilot Study. Proceedings of the Social Statistics Section, American Statistical Association, Washington, D.C., 529-533.

HABER, S. (1985). Applications of a Matched File Linking the Bureau of the Census Survey of Income and Program Participation and Economic Data. SIPP Working Paper Series No. 8502, U.S. Bureau of the Census, Washington, D.C.

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.

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

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

KILSS, B. and SCHEUREN, F. (1978). The 1973 CPS-IRS-SSA Exact Match Study. Social Security Bulletin, vol. 41, No. 10, 14-22.

LININGER, C. (1980). The Goals and Objectives of the Survey of Income and Program Participation. Proceedings of the Section on Survey Research Methods, American Statistical Association, 480-485.

MOORE, J. and KASPRZYK, D. (1984). Month-to-Month Recipiency Turnover in the ISDP. Proceedings of the Section on Survey Research Methods, American Statistical Association, 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 Administration Records, November 23-25, 1987, Ottawa, Canada.

SATER, D. K. (1985). Enhancing Data from the Survey of Income and Program Participation with Data from Economic Census and Surveys. SIPP Working Paper Series No. 8505, U.S. Bureau of the Census, Washington, D.C.

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

SCHEUREN, F., HERRIOT, R., VOGEL, L, VAUGHAN, D, KILSS, B., TYLER, B., COBLEIGH, C. and ALVEY, W. (1975). Report No. 4: Exact Match Research using the March 1973 Current Population Survey--Initial States. Studies from Interagency Data Linkages. U.S. Department of Health, Education, and Welfare, Social Security Administration, Office of Research and Statistics, Department of Health, Education and Welfare, publication No. SSA 76-11750.

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

VAUGHAN, D. (1978). Errors in Reporting Supplemental Security Income Recipiency in a Pilot Household Survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 288-293.

YCAS, M. and LININGER, C. (1981). The Income Survey Development Program: Design Features and Initial Findings. Social Security Bulletin, Vol. 44, No. 11, 13-19.

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, D.C. 20233

March 11, 1986

MEMORANDUM FOR Roger A. Herriot
Senior Demographic and
Housing Analyst

From: Douglas Sater
Chief, Revenue Sharing and Administrative
Records Staff

Subject: SSN Response Rates and Results of SSN Validation/
Improvement Operation

We recently received results of the SSN validation/improvement operation for the 1984 SIPP Wave 1 respondents in the March 7, 1986 memorandum from Chet Bowie. In the context of the CNSTAT and other exact match proposals, I have outlined below some summary figures. Also attached is a more detailed table.

Of the 53,588 persons (including children) in the 1984 Wave 1 SIPP;

72.8% reported a valid SSN
5.8% reported a invalid SSN
18.3% reported no SSN
3.1% refused SSN

After all phases of SSN validation and improvement were done, the rates were improved to;

82.4% with a valid SSN
14.4% with no SSN
3.1% refusals

However, please note this. These are raw numbers of presence or absence of SSN's for all persons, including children. This does not translate to a match status because the large majority of the 14.4% with no SSN are not "false nonmatches" but are "true nonmatches" -- because they should have no SSN. In this context, what we really want to know is "Of those persons that have an SSN, how many have we obtained in this process."

Unfortunatly, we do not know, but we can come up with an upper-bound and a lower-bound. The upper-bound is, of course, 96.6% (all persons less refusals). I have calculated a lower-bound by:

1. Assuming all adults have an SSN.

2. Assuming that 80% of the "no names" (see the Bowie memo for a description of these) are children that do not have a SSN and, therefore, are legitimate no-finds ("true nonmatches").

The lower-bound under these assumptions becomes 93.3%.

Thus, I think we can safely say that the SSN acquisition rate for persons who have a SSN is between 93% and 97%.

For comparison purposes (and as a proxy estimate for the CPS) if we did not conduct the validation and SSN improvement process, I estimate the SSN acquisition rate for persons who have a SSN to be at least 85%, some of which are invalid. The distribution would look like this:

79% reported and valid
6% reported and invalid
3% refused
12% not reported

Attachment

cc: J. Gates
D. Kasprzyk
D. Sater
Chron

POP:DSater:djd

ATTACHMENT

SSN Response/Validation/Search Rates

Total Wave 1 sample persons (including children)	53,588	100.0%	-	
SSN Refused	1,674	3.1%	-	
SSN Reported	42,128	78.6%	100.0%	
Valid	39,017	72.8%	92.6%	
Invalid	3,111	5.8%	7.4%	100.0%
Found in manual search	2,199	4.1%	5.2%	70.7%
Not found (can't find or bad microfilm)	912	1.7%	2.2%	29.3%
SSN Not Reported	9,786	18.3%	100.0%	
Found in manual search	1,764	3.3%	18.0%	
Not found in manual search	8,022	15.0%	82.0%	
No names (mostly children)	7,268	13.6%	74.3%	
Unable to locate	754	1.4%	7.7%	
Total eligible for field followup	8,934	100.0%	-	
New SSN collected	1,361	15.2%	100.0%	
Valid	1,192	13.3%	87.6%	
Invalid	169	1.9%	12.4%	
Final SSN Rates (All Persons)				
Total Wave 1 sample persons (including children)	53,588	100.0%		
SSN available and valid	44,172	82.4%		
SSN not available	9,416	17.6%		
Refused	1,674	3.1%		
Unable to locate	7,742	14.4%		
Estimated SSN Rates [1/]				
Total Wave 1 sample persons (including children)	53,588	100.0%		
SSN available and valid or person has no SSN	49,986	93.3%		
SSN missing and person has a SSN	3,602	6.7%		
SSN missing	1,928	3.6%		
SSN refused	1,674	3.1%		

1/ This is adjusted to presume that no finds on SSN's for children is equivalent to a valid SSN, that all adults should have a SSN (so any without are no SSN's) and uses a rash assumption (the old rabbit trick) that 80% of the "no names" are children.

A Review of the Use of Administrative Records in the Survey of Income and Program Participation

Dan K

Chester Bowie and Daniel Kasprzyk

INTRODUCTION

The Survey of Income and Program Participation (SIPP) is intended to provide comprehensive information on the economic resources of the American people and on how public transfer and tax programs affect their financial circumstances. The data from the SIPP are expected to provide government policy makers with an information base for studying the efficiency of government tax and transfer programs, for estimating future program costs and coverage, and for assessing the effects of proposed policy changes.

The SIPP provides comprehensive information about annual and sub-annual income and participation in public and private transfer programs for the household population in the United States; it also devotes considerable attention to measuring various kinds of economic resources other than current cash income. The most important elements of this broader perspective are the SIPP data on assets, debts, and non-cash resources, such as means-tested housing benefits, publicly and privately provided health insurance, pension coverage, and other employee fringe benefits.

The SIPP arose in response to the recognition that the principal source of information on the distribution of household and personal income in the United States—the March Income Supplement of the Current Population Survey (CPS)—had limitations which could only be rectified by making substantial changes in the survey instrument and procedures. One of the limitations of the CPS was the inability to provide linkages to administrative record data for statistical purposes. Recognizing this limitation of the CPS and the analytic usefulness of linking survey data to administrative records, the designers of the SIPP explicitly stated the ultimate SIPP data system should be a combination of data from administrative records and household surveys linked through the Social Security Number. The goals of the SIPP as described by Lininger (1980) state that administrative records will be used to:

1. increase sampling efficiency for certain subpopulations (e.g., Old Age, Survivors and Disability Insurance recipients or Supplemental Security Income recipients);

2. compare with survey data for validation studies of items common to both sources; and

3. supplement survey reported data with administrative record data for items difficult to obtain in a survey (e.g., earnings and program benefit histories).

These goals manifest themselves first in the SIPP development program and then in the SIPP.

This paper describes the SIPP's continuing commitment to the use of administrative records for statistical purposes in the following four sections: Section I reviews the work of the research and development program preceding the SIPP with regard to the use of administrative record data; Section II describes the design and content of the SIPP and its program to obtain accurate reporting of the Social Security Numbers (SSN) to facilitate linkages between survey reported data and administrative record sources; Section III describes six areas of applications where the SIPP project has initiated work involving the use of administrative records for statistical purposes; and Section IV provides a few additional examples of potential administrative data—SIPP linkage possibilities.

I. The Use of Administrative Records in the ISDP

The Income Survey Development Program (ISDP), authorized in 1975, was a program whose goal was to develop methods and a survey design to overcome underreporting and misclassification problems in the CPS (Ycas and Lininger, 1981). Furthermore, the ISDP also developed procedures and methodology for improving the collection of SSNs. The philosophy, attitudes, and plans of the ISDP strongly reflected the work of Scheuren and his colleagues (Scheuren et al, 1975) in the development of the 1973 Exact Match File (Kilss and Scheuren, 1978). 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). A brief summary of this work will provide a context for the discussion of the SIPP experience and the plans and potential for the statistical uses of administrative records in the SIPP.

In the ISDP the collection and accurate reporting of the Social Security Number (SSN) from each person in sample was deemed essential to the program. By emphasizing the collection of the SSN and then developing a system to validate and correct reported SSN's, 95.5% of the total cases were identified as having a correct SSN (Kasprzyk, 1983). The system served as a prototype for the SIPP system which is described in the next section.

The ISDP consisted of four experimental field tests which were conducted to examine different concepts, procedures and questionnaires. One aspect of each of these field tests was the use of an administrative record frame for sampling purposes. Even though the principal thrust of this approach was to increase sampling efficiency for selected subpopulations through the use of multiple frame estimators, the most important result was that these feasibility studies provided an opportunity for the survey planners to understand the administrative, methodological and operational difficulties in using administrative sources for sampling.

During the ISDP the following administrative record sources were used: 1) the Aid to Families with Dependent Children (AFDC) master file maintained by the Texas State Department of Welfare[1/], the Supplemental Security Record (SSR)[2/], the Master Beneficiary Record (MBR)[3/], the Basic Educational Opportunity Grant (BEOG) applicant file[4/], the Veterans Administration Pension and Compensation file[5/], the Internal Revenue Service Individual Master File[6/], and State record files for Unemployment Insurance and Workers Compensation.

The ISDP also effectively used administrative records to clarify misreporting and nonreporting of program benefits by comparing the survey reported data with the administrative data. Vaughan (1978) and Goudreau, Oberheu and Vaughan (1981, 1983) report on ISDP studies which led to redesigning questionnaires in order to reduce errors in classifying sources of income.

Finally, although the ISDP intended to create a data base augmented with administrative data which were difficult to obtain in a household survey it never did. A planned match to the Summary Earnings Record[7/] was never implemented because of higher priority projects.

The next section provides a very general summary of the SIPP survey design and content, followed by a description of the program established to collect and validate Social Security Numbers.

II.1. SIPP Design Features

The primary goals in designing the SIPP were to improve reporting of income and other program-related data and to do it in a way that would allow the analysis of changes over time at a microlevel. The design also had 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 of age or older) at households selected for a sample panel are interviewed about their income and other topics once every 4 months for approximately 2 1/2 years. 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 longer than a few months and reduces the number of questions asked in one interview.

To further enhance the estimates of change, particularly year-to-year change, a new sample panel is introduced every year instead of at the conclusion of a panel. Consequently, two or sometimes three panels are in the field concurrently. Since portions of the sample are the same from one year to the next, year-to-year change estimates can be based in part on a direct comparison across 2 years for the same individuals. This design gives a more precise estimate of change than a design involving interviews 1 year apart with two different groups of individuals in which greater sampling variability obscures the actual change. This overlapping panel design also allows cross-sectional estimates to be produced from a larger, combined sample that is about double in size when 2 panels overlap and triple with 3 overlapping panels.

The first SIPP panel, designated as the 1984 Panel but fielded in October 1983, started with approximately 20,000 interviewed households. The second panel, i.e., the 1985 Panel, began in February 1985 with around 14,000 interviewed households. Panels of about 12,300 interviewed households are expected to be fielded every February. The sample size changes in each wave of a panel due to losses through attrition and gains from following movers to new households.

The reference period for the primary survey items is the 4 months preceding the interview; for example, in February, 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 each month instead of one large work load every 4 months, 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 called a "wave."

II.2. SIPP Survey Content

Each interview is planned to take about 30 minutes of a respondent's time and includes content that is divided into three main groups of questions. The substance of two of these groups should be essentially the same for each wave and for each panel. The third group of questions covers topics that will change in each wave of a panel. This allows for the inclusion of some new content in each panel, although many of the topics will be repeated across all the panels. Each rotation group in a wave is administered the same set of questions although the reference period is different as explained above.

The first group of questions are control card items. The control card is a separate document from the questionnaire and serves several important functions. The control card is used to list every person residing at an address and to record basic social and demographic characteristics (age, race, sex, and so forth) for each person at the time of the initial interview. Some information relating to the housing unit or household also is collected; e.g., number of units in the structure, tenure, and so forth. The card is reused at subsequent interviews to record changes in characteristics such as age, educational attainment, and marital status, and to record the dates when persons enter or leave the household. Finally, during each interview, information on each source of income received and the name of each job or business is transcribed to the card so that this information can be used in the updating process at the next interview.

The second major group of questions form the core portion of the questionnaire, which is divided into 5 sections. The core set of questions is asked at the first interview and then updated in each subsequent interview. The first section of the core collects the basic labor force participation data for the 4 reference months.

In addition, this first section of the core collects much of the information on the receipt of income from various sources during the 4 month reference period. This includes income from government sources such as Aid to Families with Dependent Children, Supplemental Security Income, General Assistance, and Workmen's Compensation. Respondents are also asked about both Social Security and other retirement income including Railroad Retirement, pension from company or union, and civil service retirements, as well as others. The receipt of miscellaneous sources of income such as alimony, child support, interest from savings, income for foster child care, and educational assistance is also identified. In addition, questions on major sources of noncash benefits such as food stamps, WIC (Women, Infants, and Children Nutrition Program), Medicaid, Medicare, and health insurance coverage are included in this section.

The second section of the SIPP core questionnaire collects information associated with wage and salary earnings. This section includes information on industry and occupation as well as hourly earnings for up to two jobs.

The third section of the core collects data on self-employment earnings and specific information about the kind of self-employment--whether it was incorporated, sole proprietorship, or partnership--and the profits and losses from the business. Again, space is provided for two self-employment jobs.

The fourth section is identified as the general amounts section. This section of the questionnaire collects monthly amounts received from the income sources identified in the first section. That is, the first section identifies the receipt of income during the 4 month reference period, while amounts of income received are collected in the fourth section of the questionnaire. Space is provided for amounts from up to six income sources.

The fifth and last section of the core questionnaire collects amounts of income earned from asset holdings. Asset sources include savings accounts, bonds, stocks, and rental property, as well as others. Information is collected for the 4 month reference period on both individual and joint recipiency.

The third major question grouping consists of the various supplements or topical modules that are included in waves following the initial interview. A wide variety of topics are covered under the aegis of the topical module concept. The breadth of these data ensure that SIPP will be a widely used and powerful data base serving multiple purposes. The administration of a module is possible in Waves 2 through 8 (or 9 in 1984) because less time is required to update the core information after the first interview. Depending on the time available and length of the modules, more than one may be administered in the same wave. The topical modules cover areas that do not require examination every 4 months and may use a different reference period than the core questions. Some modules are assigned to only one wave of a panel, while other modules may be repeated in more than one wave. The modules provide a broader context for analysis by obtaining information on a variety of topics not covered in the core portion of the questionnaire. The module data may be analyzed independently or in conjunction with the control card items or core data. Frequently, a module is administered at the same time in concurrent panels so that the data may be combined to improve reliability.

In addition to the data collected by the survey questionnaire, the content may be supplemented with administrative record data that are difficult for respondents to recall such as lifetime earnings and program benefit histories. To facilitate future linkages with administrative records, steps have been taken in the SIPP, as in the ISDP, to ensure that the Social Security Number is obtained for as many persons as possible.

II.3. The 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. This reduces respondent burden by using other data sources for difficult-to-obtain information. Interview responses can be supplemented by information from program files such as the earnings and benefit records of the Social Security Administration (SSA). This allows, for example, analysis of the long-term impact of various Social Security benefit formulas.

To make these linkages accurate, Social Security Numbers (SSN) are required for sample individuals. The SSN is obtained for each household member in SIPP and recorded on the control card. It is identified as a critical survey data item requiring completion to make the interviewers aware of its importance. These numbers are then verified and corrected to maximize the number of accurate linkages to other record systems.

The verification and correction process builds on the work of the development program (Kasprzyk, 1983). At the conclusion of each month's interviewing during the first wave of a SIPP panel, a special extract file is prepared by the Census Bureau for the SSA. This file contains a small number of key variables (SSN, name, date of birth, sex) for all original sample persons who report a SSN, including children, in a format appropriate for machine validation. Persons who report that they do not have a number or have a number but cannot supply it are handled separately in a clerical (manual) procedure. Persons who refuse to provide a SSN are not included in the search process. The SSA identifies (by machine validation) incorrectly reported numbers then clerically resolves these cases along with cases not reporting a SSN. This work is completed by the fourth wave interview, at which time a field followup is conducted to obtain missing SSNs (provided they are not "refusals") and to reconcile inconsistencies in SSN or demographic data generated by the computer match or the clerical resolution.

Social security numbers of persons who enter the sample after Wave 1 (because they start living with original sample people) are validated at the start of the next panel. For example, information on new panel members (nonsample persons) from Waves 2 through 5 of the 1984 Panel was held and submitted for computer validation with Wave 1 of the 1985 Panel. Likewise, information on nonsample persons from Waves 6 through 8 of the 1984 Panel and Waves 2 through 4 of the 1985 Panel were held and submitted for computer validation with Wave 1 of the 1986 Panel.

The following summarizes the SSN validation results for the 1984 Panel Wave 1 sample:

53,588	Total Wave 1 sample persons
− 1,674	Persons who refused to provide a SSN and were excluded from the validation process
51,914	Persons eligible for SSN validation
−42,128	Persons who reported a usable SSN and were eligible for computer validation
9,786	Persons who did not report a SSN and were eligible for the manual search (mostly children)

44,172	Validated SSNs (85% of eligible)
7,742	Unvalidated SSNs (mostly children)
51,914	Eligible for SSN validation

the next

This section briefly describes six areas of application where work has begun to use the survey data and administrative record data in some capacity: 1) SIPP/SSA data linkage project; 2) SSA Disability Survey ADD-on to the SIPP; 3) Employer Provided Benefits Study; 4) SIPP Record Check Study; 5) the use of Administrative Records in SIPP Estimation; and 6) merging economic data with SIPP demographic data.

III.1. SSA/SIPP Data Linkage Project

SSA's interest in a data set which merges administrative data with household survey data follows closely the intended uses of SIPP at its inception. A merged data set would enable the SSA to:

1. Estimate future program costs — The SSA is responsible for projecting program costs for all major SSA programs including: The Old Age, Survivor, and Disability Program, the Supplemental Security Income Program and Aid to Families with Dependent Children. In order to improve the accuracy of the projection methods, the SIPP panel data can be linked with a number of years of SSA data so that inflows and outflows can be analyzed in addition to point-in-time prevalence estimates of SSA program participation. The relationship between program participation and underlying individual characteristics can then be used to estimate future program costs and growth thus providing the SSA an early forecasting capability.

2. Assess the effects of program policy changes — An SSA-SIPP linkage will contain family, income and SSA benefit data. This combination of information will permit the SSA to estimate the programmatic costs of policy changes that depend on these factors and to assess the effects of policy changes on the economic well-being of program participants.

3. Describe non-programmatic characteristics of program participants — The SSA is frequently asked by Congress and others to provide information about program participants that is not routinely captured by administrative record systems. In the past, the SSA has used a series of widely spaced and usually one-time surveys to provide such information. ~~The information that can be obtained is often out of date and~~ since the prospects for a new round of special purpose surveys are not good, an ongoing SSA-SIPP data link would provide relatively up-to-date data on a routine basis.

4. Test social science theories as they relate to Social Security programs — The longitudinal component of the SIPP's research design and the wealth of data captured in core questions and topical modules provide data that will be sufficiently rich to test many social and economic theories of program participation, thus making a significant contribution to the basic research that must accompany any dynamic social program.

   In essence, the project involves a maximum linkage with SIPP. For each SIPP panel, all waves of data, including core questions and topical modules will be linked to extracts of the basic SSA program records: The Master Beneficiary Record (MBR) which contains eligibiliity and benefit histories of the OASDI program, the Supplemental Security Record (SSR) which contains eligibility and benefit histories for the SSI program, and the Summary

Earnings Record (SER) which contains a history of covered earnings for each worker. ~~We will update the~~ SSA records will be updated periodically so that each panel's files will contain additional years of the SSA's program data. We may also want to link to new disability administrative files that are now being developed at the SSA on a regular basis. All initial and subsequent linkages will be by mutual agreement between the SSA and the Bureau of the Census.

The merged data set will reside on the computer at the Census Bureau and will be used only for general statistical research. Only Census Bureau staff and SSA employees who are designated as Census Bureau Special Sworn Employees will have access to the file. The SSA may publish statistical data in a summary form that does not permit the identification of a household, family, or individual.

The primary tasks in the linkage project are:

1. Verification of, and searching for, Social Security numbers (SSN) — This task is already a part of the SIPP project activities and was described earlier. In particular, the vast majority of SSN's for the 1984 SIPP panel have been processed by SSA staff.

2. Obtaining SSA administrative records — As mentioned above, this project involves matching the MBR, SSR and SER to the SIPP. Decisions will have to be made about the content of the data extracts from these files that would be included in the match. ~~Specialized computer programs may be required to complete the task~~.

   In the future, we want to consider adding data from additional SSA administrative records. We may also want to make arrangements with the Health Care Financing Administration to obtain Medicare utilization and cost data for a SIPP linkage.

3. Merging administrative records with SIPP survey data — ~~We do not view~~ the matching tasks are not one-time activities. Instead we anticipate a number of data processing operations for each SIPP panel.

4. Weighting, imputation and sampling error estimation — We will have to consider and develop schemes for weighting and imputation that take into account non-matched SIPP records. Both cross-sectional and longitudinal weights will be required. The SSA would also need the capability for estimating sampling errors.

5. Development of documentation for the matched files — Documentation for a matched file would include tape description and utilization information, the SIPP questionnaires and descriptions of the SSA administrative records, a sampling statement, ~~editorial~~ imputation descriptions and any other information required for estimation or analysis.

III.2. SSA Disability Survey ADD-On to the SIPP

The SSA needs information describing ~~and explaining~~ the economic and labor force status of disability insurance (DI) program participants. This information is not currently available. Social Security administrative data alone are not suitable for these analytical purposes. These records do not contain information on the program participants' family income and earnings before and after contact with the program. The data bases of national surveys are inadequate because of the insufficient sample of disability program participants to support reliable analysis of pertinent program issues. As a first step toward remedying this information gap, the SSA proposes to conduct a feasibility study of the use of the Census Bureau's SIPP as a suitable data collection mechanism.

There are two goals of the study. First, in order to plan for an SSA-sponsored data collection effort in the future, the SSA is testing the effectiveness of using SIPP to gather data for SSA. The study will determine whether a small SSA sample can be integrated into regular SIPP field work and data processing activities. The SSA will investigate whether they can draw a sample of allowances from the newly developed national disability data system, obtain current addresses, and forward the information to the Census Bureau to meet the SIPP interview schedules. The Census Bureau will test its ability to administer special questions to the SSA sample persons without significant dislocation to standard SIPP field practices. The Census Bureau will also test whether the data can be accurately processed along with the regular SIPP effort in a timely manner.

The second goal of the study is to provide results which will describe and explain the economic and labor force status of newly awarded DI beneficiaries. The information on family income, labor force participation, job search, job offers, and employer accommodations bear directly on work incentives and employment. This information describes the processes of family economic adjustment and return to work of disabled persons. With these data, the SSA can target beneficiary subpopulations that can best benefit from work incentive program modifications. In addition, the SSA will have data for establishing a context within which to evaluate its forthcoming series of work incentive and vocational rehabilitation demonstration projects.

This project will sample 1,200 households with newly awarded DI beneficiaries as an add-on to the Census Bureau's 1988 SIPP Panel. Sampled persons will be under age 45 and will be selected from specific geographic areas. Household members will be interviewed during the first three waves of the 1988 SIPP panel. They will answer the regularly administered SIPP questions, together with an additional set of special SSA project questions. The results of this study will determine the feasibility of conducting a full scale nationally representative add-on to the SIPP in the 1990s.

III.3. 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 on the part of Congress, other policy makers, and researchers in areas such as health care, the elderly, and tax

reform. While SIPP collects information on a respondent's contribution to retirement plans, it does not collect information on the employer's contribution. Moreover, SIPP collects information on whether a person is covered by health insurance and whether the employer makes contributions, but 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 policies are provided through an employer. Amounts of employee payments and employer contributions are not obtained.

This study involves obtaining a signed release from the respondent at the interview and contacting the respondent's employer and asking the employer to fill out a short questionnaire to obtain data on both the employer's and employee's contributions to health insurance plans, pension plans, and life insurance plans. Information provided by the employer would supplement the SIPP data and improve data quality.

A half sample of one rotation group's households was selected for the study. The test was done in August 1987, (rotation group 4) for households in Wave 8 of the 1985 Panel. This was the last interview for these households.

The test included only employed persons, 18 years old and 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 (33 percent) refused to sign, and 337 proxy or telephone respondents (25 percent) did not return the authorization form that was left/mailed to them. We did not conduct a followup of the refused or non-return cases.

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 will be undertaken next year, together with an assessment of the future prospects for a study of this type on the complete sample.

III.4. SIPP Record Check Study

Another area of research with respect to administrative record systems is the development of validation studies of items common to both the survey and administrative records. The purpose of the study is to investigate response quality issues in SIPP through a case-by-case comparison of SIPP data and administrative record information. The ultimate goal is the improved understanding of the quality of the SIPP data and, ultimately, the development of quantitative estimates of response and nonresponse errors for the purposes of adjusting survey data or modifying survey procedures to obtain better quality survey data.

An overview and progress report of the study can be found in Moore and Marquis (1987). Simply put the study intends to address the following questions:

1. The quality of the respondent reports of receipt of program benefits for a variety of state and Federally administered transfer programs;

2. The quality of benefit dollar amount reporting for these programs;

3. Demographic correlates of report quality;

4. Extent of misclassification errors;

5. The (nonexperimental) effects of self-proxy respondent status on report quality; and

6. Between wave recipiency turnover effects (The "seam" problem (Burkhead and Coder, 1985; Moore and Kasprzyk, 1984)).

The questions will be addressed by using administrative record information for receipients of each of nine government transfer programs in four states—Florida, New York, Pennsylvania, and Wisconsin. These are four state-administered programs (Aid to Families with Dependent Children, food stamps, unemployment compensation, and worker's compensation) and five Federally-administered programs (Civil-Service Retirement, Pell Grants, Old Age Survivors and Disability Insurance (OASDI), Supplemental Security Income, and Veterans' Pensions and Compensation) which will be studied. The project has obtained a great deal of information on acquiring administrative record systems, learning about each systems idiosyncrasies, and generalized matching procedures at the Census Bureau. Some very preliminary results are now available in Moore and Marquis (1987).

III.5. Use of Administrative Records in SIPP Estimation

Information on the effect of sample reductions on the variance of estimates and on our ability to measure changes in differences in the number of statistics have created serious concerns. These concerns have caused us to increase our exploration of ways to reduce the variance. One approach is through the use of administrative records for post-stratification. Currently, cross-section estimation procedures for SIPP make use of a second-stage adjustment to increase the precision of estimates by ratio adjusting collection month and reference month estimates to population estimates. However, the Census Bureau has access to some Internal Revenue Service and Social Security Administration files which can be used to produce detailed age, race, and sex distributions by adjusted gross income. The issue, which we have just begun to explore, is how these administrative data can be used for post-stratification to improve estimates of mean and median personal and household income as well as the estimates of the deciles of the personal and household income distribution. Furthermore, a basic question which will be considered is how much reduction in the variances of these estimates can be achieved through such a procedure. These issues will be investigated next year.

The first phase of this research (Huggings, 1987) will estimate the reductions in variances of SIPP estimates by using the IRS data as auxilliary variables in the estimation procedures. The procedure being studied has been advocated by Herriot (1985) and Scheuren (1983). In the SIPP study the estimation method will involve a ratio adjustment of SIPP estimates at the second stage of estimation in cells defined by age + race + sex + "income" where "income" is adjusted gross income as reported to the Internal Revenue Service.

Controls are prepared from a 1% sample of 1984 IRS file matched with age, race, and sex characteristics from the Summary Earnings Record; adjusted gross income from the 100% IRS file is matched to a file of SIPP data. The SIPP cases are then reweighted by controlling to the 1984 IRS controls; that is, a factor f_j, which is the ratio of IRS control in $cell_j$ to the SIPP estimate of persons matched to IRS data with 1984 IRS income in $cell_j$, is applied to persons who fall in $cell_j$ based on the IRS data. Estimates and variances of selected SIPP characteristics will be obtained using the newly created weights and with the weights which do not use this procedure.

III.6. Merging Economic Data with SIPP Demographic Data

During the first two years of the SIPP program a good deal of background research was completed on the potential for augmenting SIPP data with micro-level establishment and enterprise data from the economic census and other data files maintained by the Bureau of the Census (Haber, Ryscavage, Sater, Valdisera, 1984). Haber (1985) has described the analytic potential of matching economic data to the demographic data for individuals in the SIPP. Haber suggests that new insights are possible in the following areas: the relationship between capital and wage rates, the study of labor mobility between low and high-wage employees, studying implications of the transition from goods-producing to a service economy, and analyzing the effects of unions on the labor market. A pilot project was initiated to investigate methodologies for merging individuals in the SIPP (who report their employer's name) to the employer data in the economic census, testing the methodology to identify problem areas and solutions, and conduct the match for a pilot sample. Sater (1985) describes the project, and problems encountered. Unfortunately, due to costs, higher priorities, and staffing limitations, this project was never completed.

IV. Potential Linkage to Other Administrative Data Sets

The SIPP is a relatively new continuous survey, collecting a comprehensive socio-economic portrait of the U.S. household population. As mentioned above, the SIPP also gives substantial attention to the correct reporting of Social Security Numbers. These two elements together provide the principle reasons for the power of the data set. In the future, the good match variable (SSN) which the SIPP provides could be used in matching the survey data to the Health Care Financing Administration's Health Insurance Master File (Medicare) to study the relationship between hospital use, health status, employment and income. Similarly, the SSN will allow linkage of deceased respondents to the National Death Index. In the latter case, numerous SIPP panels would be necessary to have sufficient sample for analysis. Nevertheless, the potential for such linkages exist. In fact any linkage with an administrative record system which uses the SSN as the primary identifier is possible. The principal difficulties, however, are the costs for such projects and the difficulty of sharing matched administrative-survey data with all researchers.

FOOTNOTES

1/ The Aid to Families with Dependent Children (AFDC) master file is an administrative system maintained by each individual State and containing data on benefit amounts, payment history, demographic characteristics, and other information needed to administer the program.

2/ The Supplemental Security Record (SSR) is the national master administrative file for data on Supplemental Security Income (SSI) benefit amounts, payment history, and demographic data.

3/ The Master Beneficiary Record (MBR) is the national master administrative file for data on the Old Age, Survivors, and Disability Insurance Program (Title II); it contains current and historical program information on claimants for Title II benefits, past and present cash beneficiaries, disallowed claimants, and denied claims.

4/ The Basic Educational Opportunity Grant (BEOG) file is an administrative file maintained by the Department of Education. It contains data for all applicants of a given academic year, including ineligibles, eligibles who did not use their grant, and eligibles who used their grant. The BEOG program is now called the Pell Grant program.

5/ The Veteran's Administration Pension and Compensation File is a national master file containing records of benefits provided as disability compensation, dependency and indemnity compensation, disability pension, death pension, or burial allowance.

6/ The Internal Revenue Service Individual Master File (IMF) is a national file of selected income and tax information from all individual Income Tax Returns pertaining to wages, dividend and interest income, taxes paid, and exemptions.

7/ The Summary Earnings Record (SER) is a file containing the lifetime covered earnings (up to the maximum for each employer) and quarters of social security coverage of the individual. It is used to determine entitlement to benefits and calculation of benefit amounts. Individuals are indentified in this file by their social security number.

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, 351-356

GOUDREAU, K., OBERHEU, H. and VAUGHAN, D. (1981). An Assessment of the Quality of Survey Reports of Income from the Aid to Families with Dependent Children (AFDC) Program. Proceedings of the Section on Survey Research Methods, American Statistical Association, 377-382.

GOUDREAU, K., OBERHEU, H. and VAUGHAN, D. (1984). An Assessment of the Quality of Survey Reports of Income from the Aid to Families with Dependent Children (AFDC) Program. Journal of Business and Economic Statistics,179-186.

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 6. U.S. Government Printing Office, Washington, D.C. 20402.

HABER, S., RYSCAVAGE, P. SATER, D., and VALDISERA, V. (1984). Matching Economic Data to the Survey of Income and Program Participation: A Pilot Study. Proceedings of the Social Statistics Section, American Statistical Association, Washington, D.C., 529-533.

HABER, S. (1985). Applications of a Matched File Linking the Bureau of the Census Survey of Income and Program Participation and Economic Data. SIPP Working Paper Series No. 8502, U.S. Bureau of the Census, Washington, D.C.

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.

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

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

KILSS, B. and SCHEUREN, F. (1978). The 1973 CPS-IRS-SSA Exact Match Study. Social Security Bulletin, vol. 41, No. 10, 14-22.

LININGER, C. (1980). The Goals and Objectives of the Survey of Income and Program Participation. Proceedings of the Section on Survey Research Methods, American Statistical Association, 480-485..

MOORE, J. and KASPRZYK, D. (1984). Month-to-Month Recipiency Turnover in the ISDP. Proceedings of the Section on Survey Research Methods, American Statistical Association, 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 Administration Records, November 23-25, 1987, Ottawa, Canada.

SATER, D. K (1985). Enhancing Data from the Survey of Income and Program Participation with Data from Economic Census and Surveys. SIPP Working Paper Series No. 8505, U.S. Bureau of the Census, Washington, D.C.

SCHEUREN, F., HERRIOT, R., VOGEL, L, VAUGHAN, D, KILSS, B., TYLER, B., COBLEIGH, C. and ALVEY, W. (1975). Report No. 4: Exact Match Research using the March 1973 Current Population Survey---Initial States. Studies from Interagency Data Linkages. U.S. Department of Health, Education, and Welfare, Social Security Administration, Office of Research and Statistics, Department of Health, Education and Welfare, publication No. SSA 76-11750.

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

VAUGHAN, D. (1978). Errors in Reporting Supplemental Security Income Recipiency in a Pilot Household Survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, 288-293.

YCAS, M. and LININGER, C. (1981). The Income Survey Development Program: Design Features and Initial Findings. Social Security Bulletin, Vol. 44, No. 11, 13-19.

Statistics Canada / Statistique Canada
Ottawa, Canada

December 8, 1987

Dr. Dan Kasprzyk
Population Division
U.S. Bureau of the Census
Rooøm 2024, FOB3
Washington, D.C. 20233
U.S.A.

Dear Dan Dr. Kasprzyk;

I want to offer my thanks and congratulations to you for your participation and excellent presentation of your paper at the Symposium. Your cooperation throughout the preparatory phase made our task much easier and resulted in the smooth running of the entire program. Judging from many reactions and opinions expressed by various participants it was a very successful Symposium.

As mentioned at the Symposium, we are planning to publish the proceedings containing both invited and contributed papers. As you known, we have received manuscripts that are in different stages of preparation towards the final version. In case the manuscript of your paper with us, is the final version, please let me know to that effect, so that we can proceed with its processing immediately. Otherwise the final version of your paper should be sent to us at the latest by mid-January 1988 in order to facilitate timely publication of the proceedings.

In all cases we would greatly appreciate receiving the diskette, if the paper has been typed in machine readable form, indicating the type of software used (a list of software available with us is attached, along with guidelines for manuscript preparation.) Please note that except for some minor editing, no formal review of these papers will be undertaken for inclusion in the proceedings.

.../2

Further, a selected number of papers will be reviewed for publication in the Statistics Canada's SURVEY METHODOLOGY JOURNAL. Please indicate if you would like your paper to be considered for the Journal.

no problem. Finally, in case you do have the facilities to get your paper translated into French, we would appreciate receiving copies (and the diskette) in both the official languages of Canada.

Once again thank you very much for your cooperation.

Yours sincerely,

M.P. Singh
Co-Chairperson, Symposium Organizing Committee
4-C2, Jean Talon Building
Ottawa, Ontario
K1A 0T6
Tel: 951-9894

SOFTWARE USED IN SOCIAL SURVEY METHODS DIVISION

XEROX 860

XEROX WRITER II/III

MICROPRO WORDSTAR

TANDA WORD WAND/L'EDITEXTE

PELADA COMMUNIQUE

MICROSOFT WORD

WORD PERFECT

LIAISON

L'ACCENT

MULTIMATE

PC WRITE

IMPRESSIONS

IF THE SOFTWARE PACKAGES LISTED ABOVE ARE NOT AVAILABLE CONVERT FILE TO ASCII (EXTENDED)

GUIDELINES FOR MANUSCRIPTS

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1 inches on all sides.

1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.

1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).

3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.

4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

1. Présentation

1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.

1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.

1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.

1.4 Les remerciements doivent paraître à la fin du texte.

1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

3.1 Éviter les notes au bas des pages, les abréviations et les sigles.

3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.

3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.

3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.

3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0; 1, l).

3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).

4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5. Bibliographie

5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).

5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

A Review of the Use of Administrative Records in the Survey of Income and Program Participation

Daniel Kasprzyk and Chester Bowie

First paragraph needs polishing -- I suggest exerpt from the recent budget write up

INTRODUCTION

In October 1983, the Bureau of the Census conducted the first interviews of the Survey of Income and Program Participation (SIPP). The SIPP is a nationally representative household survey intended to provide detailed information on all sources of cash and noncash income, eligibility and participation in various government transfer programs, disability, labor force status, assets and liabilities, pension coverage, taxes, and many other items. Data from the survey will provide a multiyear perspective on changes in income, and their relationship to participation in government programs, changes in household composition, and so forth. In general, the SIPP data system is designed to directly measure elements of tax and transfer system in a comprehensive data base.

The SIPP arose in response to the recognition that the principal source of information on the distribution of household and personal income in the United States--the March Income Supplement of the Current Population Survey (CPS)--had limitations which could only be rectified by making substantial changes in the survey instrument and procedures. The limitations of the CPS ~~extended to its~~ inability to provide linkages to administrative record data for statistical purposes. Recognizing this limitation of the CPS and the analytic usefulness of linking survey data to administrative records, the designers of the SIPP explicitly stated the SIPP data system should be a combination of data from administrative records and household surveys linked ~~by the~~ Social Security Number. The goals of the SIPP as described by Lininger (1980) state that administrative records will be used to:

1. increase sampling efficiency for certain subpopulations (e.g., Old Age, Survivors and Disability Insurance recipients or Supplemental Security Income recipients);

2. compare with survey data for validation studies of items common to both sources; and

3. supplement survey reported data with administrative record data for items difficult to obtain in a survey (e.g., earnings and program benefit histories).

The purpose of this paper is to 1) review the experience of the research and development program preceding the SIPP with regard to administrative record data usage, 2) review the ~~design and~~ content of the SIPP and its program to obtain accurate reporting of the Social Security Numbers (SSN) to facilitate linkages between survey reported data and administrative record sources, and 3) review the potential applications of administrative records in the SIPP.

The Use of Administrative Records in the ISDP

The Income Survey Development Program (ISDP), authorized in 1975, was a program whose goal was to develop methods and a survey design to overcome underreporting and misclassification problems in the CPS (Ycas, and Lininger, 1981). Furthermore, the ISDP also developed procedures and methodology for improving the collection of SSNs. The philosophy, attitudes, and plans of the ISDP strongly reflected the work of Scheuren and his colleagues (Scheuren et al, 1975) in the development of the 1973 Exact Match File (Kilss and Scheuren 1978). 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). A brief summary of this work will provide a context for the discussion of the SIPP experience, the and plans and potential for the statistical uses of administrative records in the SIPP.

In the ISDP the collection and accurate reporting of the Social Security Number (SSN) from each person in sample was deemed essential to the program. By emphasizing the collection of the SSN and then developing a system to validate and correct reported SSN's, 95.5% of the total cases were identified as having a correct SSN (Kasprzyk 1983). The system developed served as a prototype for the SIPP system which is described in the next section.

The ISDP consisted of four experimental field tests which were conducted to examine different concepts, procedures and questionnaires. One aspect of each of these field tests was the use of an administrative record frame for sampling purposes. Even though the principal thrust of this approach was to increase sampling efficiency for selected subpopulations through the use of multiple frame estimators, the most important result was that these feasibility studies provided an opportunity for the survey planners to understand the administrative, methodological and operational difficulties in using administrative sources for sampling. — need to expand on this -- basically, what was learned?

During the ISDP the following administrative record sources were used: 1) the Aid to Families with Dependent Children (AFDC) master file maintained by the Texas State Department of Welfare [1], the Supplemental Security Record (SSR) [2], the Master Beneficiary Record (MBR) [3], the Basic Educational Opportunity Grant (BEOG) applicant file [4], Veterans Administration Pension and Compensation file, the Internal Revenue Service, State record files for Unemployment Insurance and Workers Compensation. this isn't a "file" - the rest are

The ISDP also effectively used administrative records to clarify misreporting and nonreporting of program benefits by comparing the survey reported data with the administrative data. Vaughan (1978) and Goudreau, Oberhau and Vaughan (1981, 1983) report on ISDP studies which led to redesigning questionnaires in order to reduce errors in misclassifying sources of income.

Finally, the ISDP never was able to provide a data base augmented with administrative data which were difficult to obtain in a household survey. A planned match to the Summary Earnings Record [5] was never implemented because of competing projects.

Need to give some idea of why -- technical problems or just time + resource problems

"Canada Paper"

This should be part of the Intro... probably after your 1st paragraph

To: DAN
FROM: CHET

SIPP OVERVIEW

Design Features

The primary goals in designing the SIPP were to improve reporting of income and other program-related data and to do it in a way that would allow the analysis of changes over time at a microlevel. The design also had 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 of age or older) at households selected for a sample panel are interviewed about their income and other topics once every 4 months for approximately 2 1/2 years. 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 longer than a few months and reduces the number of questions asked in one interview.

To further enhance the estimates of change, particularly year-to-year change, a new sample panel is introduced every year instead of at the conclusion of a panel. Consequently, two or sometimes three panels are in the field concurrently. Since portions of the sample are the same from one year to the next, year-to-year change estimates can be based in part on a direct comparison across 2 years for the same individuals. This design gives a more precise estimate of change than a design involving interviews 1 year apart with two different groups of individuals in which greater sampling variability obscures the actual change. This overlapping panel design also allows cross-sectional estimates to be produced from a larger, combined sample that is about double in size when 2 panels overlap and triple with 3 overlapping panels.

The first SIPP panel, designated as the 1984 Panel but implemented in October 1983, started with approximately 20,000 interviewed households. The second panel, i.e., the 1985 Panel, began in February 1985 with around 14,000 interviewed households. Panels of about 12,300 interviewed households are expected to be fielded every February. The sample size changes in each wave of a panel due to losses through attrition and gains from following movers to new households.

The reference period for the primary survey items is the 4 months preceding the interview; for example, in February, 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 each month instead of one large work load every 4 months, 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 called a "wave."

Survey Content

Each interview is planned to take about 30 minutes of a respondent's time and includes content that is divided into three main groups of questions. The substance of two of these groups should be essentially the same for each wave and for each panel. The third group of questions covers topics that will change in each wave of a panel. This ~~will~~ allows for the inclusion of some new

content in each panel, although many of the topics will be repeated across all the panels. Each rotation group in a wave is administered the same set of questions although the reference period is different as explained above.

The first group of questions are control card items. The control card is a separate document from the questionnaire and serves several important functions. The control card is used to list every person residing at an address and to record basic social and demographic characteristics (age, race, sex, and so forth) for each person at the time of the initial interview. Some information relating to the housing unit or household also is collected; e.g., number of units in the structure, tenure, and so forth. The card is reused at subsequent interviews to record changes in characteristics such as age, educational attainment, and marital status, and to record the dates when persons enter or leave the household. Finally, during each interview, information on each source of income received and the name of each job or business is transcribed to the card so that this information can be used in the updating process at the next interview.

The second major group of questions form the core portion of the questionnaire, which is divided into 5 sections. The core set of questions is asked at the first interview and then updated in each subsequent interview. The first section of the core collects the basic labor force participation data for the 4 reference months.

In addition, this first section of the core collects much of the information on the receipt of income from various sources during the 4 month reference period. This includes income from government sources such as Aid to Families with Dependent Children, Supplemental Security Income, General Assistance, and Workmen's Compensation. Respondents are also asked about both Social Security and other retirement income including Railroad Retirement, pension from company or union, and civil service retirements, as well as others. The receipt of miscellaneous sources of income such as alimony, child support, interest from savings, income for foster child care, and educational assistance is also identified. In addition, questions on major sources of noncash benefits such as food stamps, WIC (Women, Infants, and Children Nutrition Program), Medicaid, Medicare, and health insurance coverage are included in this section.

The second section of the SIPP core questionnaire collects information associated with wage and salary earnings. This section includes information on industry and occupation as well as hourly earnings for up to two jobs.

The third section of the core collects data on self-employment earnings and specific information about the kind of self-employment--whether it was incorporated, sole proprietorship, or partnership--and the profits and losses from the business. Again, space is provided for two self-employment jobs.

The fourth section is identified as the general amounts section. This section of the questionnaire collects monthly amounts received from the income sources identified in the first section. That is, the first section identifies the receipt of income during the 4 month reference period, while amounts of income received are collected in the fourth section of the questionnaire. Space is provided for amounts from up to six income sources.

The fifth and last section of the core questionnaire collects amounts of income earned from asset holdings. Asset sources include savings accounts, bonds, stocks, and rental property, as well as others. Information is collected for the 4 month reference period on both individual and joint recipiency.

The third major question ~~group~~ grouping consists of the various supplements or topical modules that are included in waves following the initial interview. The administration of a module is possible in Waves 2 through 8 (or 9 in 1984) because less time is required to update the core information after the first interview. Depending on the time available and length of the modules, more than one may be administered in the same wave. The topical modules cover areas that do not require examination every 4 months and may use a different reference period than the core questions. Some modules are assigned to only one wave of a panel, while other modules may be repeated in more than one wave. The modules provide a broader context for analysis by obtaining information on a variety of topics not covered in the core portion of the questionnaire. The module data may be analyzed independently or in conjunction with the control card items or core data. Frequently, a module is administered at the same time in concurrent panels so that the data may be combined to strengthen the analyses.

There are two types of topical modules: fixed and variable. The fixed topical modules are designed to be conducted on a regular basis to augment the core data. They are considered necessary to meet the survey's goals and objectives. Although the topics are "fixed," the questions in these modules may be modified from time-to-time to accommodate conceptual changes or to make improvements in collecting these data. An example of a fixed topical module is the annual "round-up" module on earnings and benefits.

The variable topical modules are designed to satisfy the special programmatic needs of ~~other~~ Federal agencies. Time is set aside for variable modules in several waves to allow the flexibility to add content to meet special needs that develop as the survey continues. An example of a variable topical module is the child care topical module administered in the 1984 Panel. It was developed to obtain information about child care arrangements, such as who provides the care, the number of hours of care per week, where the care is provided, and the cost of the care. These data will be useful to other agencies because child care expenses are frequently deductible for program eligibility purposes. Variable topical modules may be repeated in subsequent waves or panels as necessary.

A wide variety of topics are covered under the aegis of the topical module concept. The breadth of these data ensure that SIPP will be a widely used and powerful data base serving multiple purposes.

In addition to the data collected by the survey questionnaire, the content may be supplemented with administrative record data that are difficult for respondents to recall such as lifetime earnings and program benefit histories. To facilitate future linkages with administrative records, steps have been taken in the SIPP to ensure that the social security number is obtained for as many persons as possible.

be consistent thru-out -- either l.c. or u.c.

Links to Administrative Records and the Validation of Social Security Numbers

Background

The SIPP data system has always been thought of as a combination of data from administrative records and household surveys. This reduces respondent burden by using other data sources for difficult-to-obtain information. Interview responses can be supplemented by information from program files such as the earnings and benefit records of the Social Security Administration (SSA). This allows, for example, analysis of the long-term impact of various Social Security benefit formulas.

compared to what. - not sure this sentence is needed.

To make these linkages accurate, social security numbers (SSN) are required for sample individuals. The SSN is obtained for each household member in SIPP and recorded on the control card. It is identified as a critical survey data item requiring completion to make the interviewers aware of its importance. These efforts should result in improved accuracy of the survey-reported social security numbers. These numbers are then verified and corrected to maximize the number of accurate linkages to other record systems.

The verification and correction process builds on the work of the development program (the Income Surveys Development Program) in which SSNs were obtained and verified for more than 95 percent of the adult (15+) sample (see "Social Security Number Reporting, the Use of Administrative Records, and the Multiple Frame Design in the Income Survey Development Program" by D. Kasprzyk in David, M. (ed.), 1983). At the conclusion of each month's interviewing during the first wave of a SIPP panel, a special extract file is prepared by the Census Bureau for the SSA. This file contains a small number of key variables (SSN, name, date of birth, sex) for all original sample persons who report a SSN, including children, in a format appropriate for machine validation. Persons who report that they have a number but cannot supply it ~~or that they do not have a number~~ are handled separately in a clerical (manual) procedure. Persons who refuse to provide ~~a~~ an SSN are not included in the search process. The SSA identifies (by machine validation) incorrectly reported numbers then ~~also~~ clerically resolves these cases along with cases with a missing SSN. This work is completed ~~by~~ before the timing for the fourth wave interview, at which time a field followup is conducted to obtain missing SSNs (provided they are not "refusals") and to reconcile inconsistencies in SSN or demographic data generated by the computer match or the clerical resolution.

- just use a reference not full title

Social security numbers of persons who enter the sample after Wave 1 (because they start living with original sample people) are validated at the start of the next panel. ~~For example, information on new panel members (nonsample persons) from Waves 2 through 5 of the 1984 Panel was held and submitted for computer validation with Wave 1 of the 1985 Panel. Likewise, information on nonsample persons from Waves 6 through 8 of the 1984 Panel and Waves 2 through 4 of the 1985 Panel will be held and submitted for computer validation with Wave 1 of the 1986 Panel.~~

too much detail

The following summarizes the SSN validation results for the 1984 Panel Wave 1 sample:

53,588	Total Wave 1 sample persons
- 1,674	Persons who refused to provide a SSN and were excluded from the validation process
51,914	Persons eligible for SSN validation
-42,128	Persons who reported a usable SSN and were eligible for computer validation
9,786	Persons who did not report a SSN and were eligible for the manual search (mostly children)
44,172	Validated SSNs (85% of eligible)
7,742	Unvalidated SSNs (mostly children)
51,914	Eligible for SSN validation

← Is this the beginning of the "potential applications" section? Please add a section title because also subsections corresponding to 1, 2, & 3 on the first page would help. Or at least an intro. that discusses the general uses.

SSA Disability Survey ADD-On to the SIPP

The SSA needs information describing and explaining the economic and labor force status of disability insurance (DI) program participants in the periods immediately following the key programmatic events of awards, terminations, and denial of benefits. This information is not currently available. Social Security administrative data are not suitable for these analytical purposes because these records do not contain information on the program participants' family income and earnings before and after contact with the program. The data bases of national surveys are inadequate because they have not sampled enough disability program participants to support ~~accurate~~ reliable analysis of pertinent program issues. As a first step toward remedying this information gap, SSA proposes to conduct a feasibility study of the use of the Census Bureau's SIPP as a suitable data collection mechanism.

There are two goals of the study. First, in order to plan for an SSA-sponsored data collection effort in the future, SSA is testing the effectiveness of using SIPP to gather data for SSA.

The study will determine whether a national SSA panel can be integrated into regular SIPP field work and data processing activities. The SSA will investigate whether they can draw a sample of allowances from the newly developed national disability data system, obtain current addresses, and forward the information to the Census Bureau to meet the SIPP interview schedules. The Census Bureau will test its ability to administer special questions to the SSA sample persons without significant dislocation to standard SIPP field practices. The Census Bureau will also test whether the data can be accurately processed along with the regular SIPP effort in a timely manner. Second, the study results will describe and explain the economic and labor force status of newly awarded DI beneficiaries. The information on family income, labor force participation, job search, job offers, and employer accommodations bear directly on work incentives and employment. This information describes the processes of family economic adjustment and return to work of disabled persons. With these data, the SSA can target beneficiary subpopulations that can best benefit from work incentive program modifications. In addition, the SSA will have data for establishing a context within which to evaluate its forthcoming series of work incentive and vocational rehabilitation demonstration projects.

This project will sample 1,200 households with newly awarded DI beneficiaries as an add-on to the Census Bureau's 1988 SIPP. Sampled persons will be under age 45 and will be selected from specific geographic areas. Household members

Must cover confidentiality protections in this section.

will be interviewed during the first three waves of the 1988 SIPP panel. They will answer the regularly administered SIPP protocol, together with an additional set of special SSA project questions.

SSA/SIPP Data Linkage Project

parallels (or corresponds to)

SSA's interest in a data link ~~follow closely~~ the intended uses of SIPP at its inception. A linked data set would enable the SSA to:

1. Estimate future program costs. -- The SSA is responsible for projecting program costs for all major SSA programs including: The Old Age, Survivor, and Disability Program, the Supplemental Security Income Program and Aid to Families with Dependent Children. In order to improve the accuracy of the projection methods, the SIPP panel data can be linked with a number of years of SSA data so that inflows and outflows can be analyzed in addition to point-in-time prevalence estimates of SSA program participation. The relationship between program participation and underlying individual characteristics can then be used to estimate future program costs and growth based on the SIPP data alone from future panels thus providing the SSA an early forecasting capability.

2. Assess the effects of program policy changes. -- An SSA-SIPP linkage will contain family, income and SSA benefit data. This combination of information will permit the SSA to estimate the programmatic costs of policy changes that depend on these factors and to assess the effects of policy changes on the economic well-being of program participants.

3. Describe non-programmatic characteristics of program participants. -- The SSA is frequently asked by Congress and others to provide information about program participants that is not routinely captured by administrative record systems. In the past, the SSA has used a series of widely spaced and usually one-time surveys to provide such information. The information that can be obtained is often out-of-date and the prospects for a new round of special purpose surveys are not good. An ongoing SSA-SIPP data link would provide relatively up-to-date data on a routine basis.

4. Test social science theories as they relate to Social Security programs. -- The longitudinal component of the SIPP's research design and the wealth of data captured in core questions and topical modules provide data that will be sufficiently rich to test many social and economic theories of program participation, thus making a significant contribution to the basic research that must accompany any dynamic social program.

In essence, SSA wants a maximal linkage with SIPP. For each SIPP panel, they want all waves of data, including core questions and topical modules linked to extracts of the basic SSA program records: The Master Beneficiary Record (MBR) which contains eligibiliity and benefit histories or the OASDI program, the Supplemental Security Record (SSR) which contains eligibility and benefit histories for the SSI program, and the Summary Earnings Record (SER) which contains a history of covered earnings for each worker. They will want to update the SSA records periodically so that each panel's files will contain additional years of the SSA's program data. They may also want to link to new administrative files in the disability area that are now being developed at the SSA on a regular basis. All initial and subsequent linkages will be by mutual agreement between the SSA and the Bureau of the Census.

While it is not possible to specify now, the format of the linked files, they must be constructed in such a way so as to be accessible via remote terminals from both Baltimore and Washington and on both UNIVAC and IBM systems.

The primary tasks in the linkage project are:

1. Verification of, and scouting for, Social Security numbers (SSN). -- This task is already a part of the SIPP project activities. In particular, the vast majority of SSN's for the 1984 SIPP panel have been processed by SSA staff.

2. Obtaining SSA administrative records. -- As mentioned above, the SSA has a definite interest in matching the MBR, SSR and SER to the SIPP. Decisions will have to be made about the content of the data extracts from these files that would be included in the match. Specialized computer programs may be required to complete the task.

   In the future, they want to consider adding data from additional SSA administrative records. They may also want to make arrangements with the Health Care Financing Administration to obtain Medicare utilization and cost data for a SIPP linkage.

3. Merging administrative records with SIPP survey data. -- The SSA does not see the matching tasks as one-time activities. Instead they anticipate a number of data processing operations for each SIPP panel.

   a. Development and execution of an initial matching operation for the panel (including the first 2 or 3 waves).

   b. Updating the matched file for additional waves of interviews and topical modules.

   c. Updating the matched file with later (or new) SSA administrative record extracts.

4. Weighting, imputation and sampling error estimation. -- We will have to consider and develop schemes for weighting and imputation that take into account non-matched SIPP records. Both cross-sectional and longitudinal weights will be required. The SSA would also need the capability for estimating sampling errors. Some sort of half sample replication approach would be most convenient.

   meaningless - drop or expand

5. Development of documentation for the matched files. -- Documentation for a matched file would include tape description and utilization information, the SIPP questionnaires and descriptions of the SSA administrative records, a sampling statement, editorial imputation descriptions and any other information required for estimation or analysis.

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 on the part of Congress, other policy makers, and researchers in areas such as health care, the elderly, and tax reform. While SIPP collects information on a respondent's contribution to retirement plans, it does not collect information

on the employer's contribution. Moreover, SIPP collects information on whether a person is covered by health insurance and whether the employer makes contributions, but 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 policies are provided through an employer. Amounts of employee payments and employer contributions are not obtained.

This study involves obtaining a signed release from the respondent at the interview and contacting the respondent's employer and asking the employer to fill out a short questionnaire to obtain data on both the employer's and employee's contributions to health insurance plans, pension plans, and life insurance plans. Information provided by the employer would supplement the SIPP data and improve data quality.

A half sample of one rotation group's households was selected for the study. The test was done in August 1987, (rotation group 4) for households in Wave 8 of the 1985 Panel. This was the last interview for these households.

The test included only employed persons, 18 years old and older, for whom a Wave 8 interview questionnaire was completed. Of the 1,352 persons eligible for the test, 563 persons (42 percent) signed the authorization form, 446 persons (33 percent) refused to sign, and 343 proxy or telephone respondents (25 percent) did not return the authorization form that was left/mailed to them. We did not conduct a followup of the refused or non-return cases.

Of the 563 questionnaires that were mailed to an employer, 424 (75 percent) were returned after the initial mailing and required no followup.

Need something on future plans,

SIPP Record Check Study -- Validation

Another area of research with respect to administrative record systems is the development of validation studies of items common to both the survey and administrative records. The purpose of the study is to investigate response quality issues in SIPP through a case-by-case comparison of SIPP data and administrative record information. The ultimate goal is the improved understanding of the quality of the SIPP data and, ultimately, the development of quantitative estimates of response and nonresponse errors for the purposes of adjusting survey data or modifying survey procedures to obtain better quality survey data.

An overview and progress report of the study can be found in Moore and Marquis (1987). Simply put the study intends to address the following questions:

1. The quality of recipiency status reporting (jargon-reporting whether a person has received benefits) for a variety of state and Federally administered transfer programs;
2. The quality of benefit dollar amount reporting for these programs;
3. Demographics correlates of report quality;
4. Identify extent of misclassification errors;
5. The (nonexperimental) effects of self-proxy respondent status on report quality; and
6. Between wave recipiency turnover effects (The "seam" problem (Burkhead and Coder, 1985; Moore and Kasprzyk 1984)).

The questions will be addressed by using administrative record information for recipients of each of nine government transfer programs in four states--Florida, New York, Pennsylvania, and Wisconsin. These are four state-administered programs (Aid to Families with Dependent Children, food stamps, unemployment compensation, and worker's compensation) and five Federally-administered

programs (Civil-Service Retirement, Pell Grants, Old Age Surviors and Disability Insurance (OASDI), Supplemental Security Income, and Veterans' Pensions and Compensation) which will be studied. The project has obtained a great deal of information on how to set up acquiring administrative record systems, learning about each systems idiosyncradies, and generalized matching procedures at the Census Bureau. Some very preliminary results are now available in Moore and Marquis (1987).

Use of Administrative Records in SIPP Estimation

~~Recent new statistics~~ Information on the effect of sample reductions on the variance of the estimates and our ability to measure changes in differences in ~~the~~ a number of statistics have created serious concerns. These concerns have caused us to increase our exploration of ways to reduce the variance. One approach is through the use of administrative records for post-stratification. Currently, cross-section estimation procedures for SIPP make use of a second-stage adjustment to increase the precision of estimates by ratio adjusting collection month and reference month estimates to CPS March type population estimates. However, the Census Bureau has access to some Internal Revenue Service and Social Security Administration files which can be used to produce detailed age, race, and sex distribution ~~of~~ by adjusted gross income. The issue, which we have just begun to explore, is how these administrative data can be used for post-stratification to improve estimates of mean and median personal and household income as well as the estimates of the deciles of the personal and household income distribution. Furthermore, a basic question which will be considered is how much reduction in the variances of these estimates can be achieved through such a procedure. These issues will be researched during the next 6 months. Further information concerning this topic will be provided at the meeting.

The first phase of this research ~~being undertaken by Vickie Huggins and Robert Fay of the Census Bureau's Statistical Methods Division~~ (Huggins, 1987) will estimate the reductions in variances of SIPP estimates by using the IRS data as auxilliary variables in the estimation procedures. The procedure being studied has been advorated by Herriot (1985) and Scheuren (1983). In the SIPP study the estimation method will involve a ratio adjustment of SIPP estimates at the second stage of estimation in cells defined by age + race + sex + "income" where "income" is adjusted gross income as reported to the Internal Revenue Serivce.

Controls are prepared from a 1% sample of 1984 IRS file matched with age, race, and sex characteristics from the Summary Earnings Record; adjusted gross income from the 100% IRS file is matched to a file of SIPP data. The SIPP cases are then reweighted by controlling to the 1984 IRS controls; that is, a factor f_j, which is the ratio of IRS control in cell j to the SIPP estimate of persons matched to IRS data with 1984 IRS income in cell j, is applied to persons who fall in cell j based on the IRS data. Estimates and variances of selected SIPP characteristics will be obtained using the newly created weights and with the weights which do not use this procedure.

Economic Data and the SIPP Demographic Data

During the first two years of the SIPP program a good deal of background research was completed on the potential for augmenting SIPP data with micro-level establishment and enterprise data from the economic census and other data files maintained by the Bureau of the Census (Haber, Ryscavage, Sater, Valdisera, 1984). Haber (1985) has described the analytic potentical of matching economic data to the demographic data for individuals in the SIPP. Haber suggests that new insights are possible in the following areas: the relationship between

capitol and wage rates, the study of labor mobility between low and high-wage employees, the study of differences in individual earnings in low and high-paying firms ~~due~~ related to the characteristics of the workers and the degree of capitalization of the employees measuring the effect of minimum wage legislation, studying implications of the transition from a goods producing to a service economy, and analyzing the effects of unions on the labor market. A pilot project was initiated to investigate methodologies for monitoring individuals in the SIPP who report their employees name to the employees data in the economic census, testing the methodology to identify problem areas and solutions, and conducting the match for a pilot sample. Sater (1985) describes the project, and problems encountered. Unfortunately, due to costs, higher priorities, and staffing limitations, this project was never completed.

Potential Linkage to Other Administrative Data Sets

The SIPP is a relatively new continuous survey, collecting a comprehensive socio-economic portrait of ~~individuals selected into sample~~ the household population. As mentioned above, the SIPP also gives substantial attention to the correct reporting of Social Security Numbers. These two elements together provide the principle reasons for the power of the data set. In the future, the good link variable (SSN) which the SIPP provides could be used in matching the survey data to the Health Care Financing Administration's Health Insurance Master File (Medicare) to study the relationship between hospital use, health status, employment and income. Similarly, the SSN will allow linkage of deceased respondents to the National Death Index. In the latter case, numerous SIPP panels would be necessary to have sufficient sample for analysis. Nevertheless, the potential for such linkages exist. In fact any linkage with an administrative record system which uses the SSN as the primary identifier is possible. The principal difficulties, however, are the

costs for such projects and the difficulty of sharing matched administrative-survey data with all researchers. The latter topic goes far beyond the scope of this paper.

Conclusion

"Canada Paper"

To: DAN
FROM: CHET

SIPP OVERVIEW

Design Features

The primary goals in designing the SIPP were to improve reporting of income and other program-related data and to do it in a way that would allow the analysis of changes over time at a microlevel. The design also had 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 of age or older) at households selected for a sample panel are interviewed about their income and other topics once every 4 months for approximately 2 1/2 years. 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 longer than a few months and reduces the number of questions asked in one interview.

To further enhance the estimates of change, particularly year-to-year change, a new sample panel is introduced every year instead of at the conclusion of a panel. Consequently, two or sometimes three panels are in the field concurrently. Since portions of the sample are the same from one year to the next, year-to-year change estimates can be based in part on a direct comparison across 2 years for the same individuals. This design gives a more precise estimate of change than a design involving interviews 1 year apart with two different groups of individuals in which greater sampling variability obscures the actual change. This overlapping panel design also allows cross-sectional estimates to be produced from a larger, combined sample that is about double in size when 2 panels overlap and triple with 3 overlapping panels.

The first SIPP panel, designated as the 1984 Panel but implemented in October 1983, started with approximately 20,000 interviewed households. The second panel, i.e., the 1985 Panel, began in February 1985 with around 14,000 interviewed households. Panels of about 12,300 interviewed households are expected to be fielded every February. The sample size changes in each wave of a panel due to losses through attrition and gains from following movers to new households.

The reference period for the primary survey items is the 4 months preceding the interview; for example, in February, 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 each month instead of one large work load every 4 months, 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 called a "wave."

Survey Content

Each interview is planned to take about 30 minutes of a respondent's time and includes content that is divided into three main groups of questions. The substance of two of these groups should be essentially the same for each wave and for each panel. The third group of questions covers topics that will change in each wave of a panel. This ~~will~~ allows for the inclusion of some new

content in each panel, although many of the topics will be repeated across all the panels. Each rotation group in a wave is administered the same set of questions although the reference period is different as explained above.

The first group of questions are control card items. The control card is a separate document from the questionnaire and serves several important functions. The control card is used to list every person residing at an address and to record basic social and demographic characteristics (age, race, sex, and so forth) for each person at the time of the initial interview. Some information relating to the housing unit or household also is collected; e.g., number of units in the structure, tenure, and so forth. The card is reused at subsequent interviews to record changes in characteristics such as age, educational attainment, and marital status, and to record the dates when persons enter or leave the household. Finally, during each interview, information on each source of income received and the name of each job or business is transcribed to the card so that this information can be used in the updating process at the next interview.

The second major group of questions form the core portion of the questionnaire, which is divided into 5 sections. The core set of questions is asked at the first interview and then updated in each subsequent interview. The first section of the core collects the basic labor force participation data for the 4 reference months.

In addition, this first section of the core collects much of the information on the receipt of income from various sources during the 4 month reference period. This includes income from government sources such as Aid to Families with Dependent Children, Supplemental Security Income, General Assistance, and Workmen's Compensation. Respondents are also asked about both Social Security and other retirement income including Railroad Retirement, pension from company or union, and civil service retirements, as well as others. The receipt of miscellaneous sources of income such as alimony, child support, interest from savings, income for foster child care, and educational assistance is also identified. In addition, questions on major sources of noncash benefits such as food stamps, WIC (Women, Infants, and Children Nutrition Program), Medicaid, Medicare, and health insurance coverage are included in this section.

The second section of the SIPP core questionnaire collects information associated with wage and salary earnings. This section includes information on industry and occupation as well as hourly earnings for up to two jobs.

The third section of the core collects data on self-employment earnings and specific information about the kind of self-employment--whether it was incorporated, sole proprietorship, or partnership--and the profits and losses from the business. Again, space is provided for two self-employment jobs.

The fourth section is identified as the general amounts section. This section of the questionnaire collects monthly amounts received from the income sources identified in the first section. That is, the first section identifies the receipt of income during the 4 month reference period, while amounts of income received are collected in the fourth section of the questionnaire. Space is provided for amounts from up to six income sources.

The fifth and last section of the core questionnaire collects amounts of income earned from asset holdings. Asset sources include savings accounts, bonds, stocks, and rental property, as well as others. Information is collected for the 4 month reference period on both individual and joint recipiency.

The third major question grouping consists of the various supplements or topical modules that are included in waves following the initial interview. The administration of a module is possible in Waves 2 through 8 (or 9 in 1984) because less time is required to update the core information after the first interview. Depending on the time available and length of the modules, more than one may be administered in the same wave. The topical modules cover areas that do not require examination every 4 months and may use a different reference period than the core questions. Some modules are assigned to only one wave of a panel, while other modules may be repeated in more than one wave. The modules provide a broader context for analysis by obtaining information on a variety of topics not covered in the core portion of the questionnaire. The module data may be analyzed independently or in conjunction with the control card items or core data. Frequently, a module is administered at the same time in concurrent panels so that the data may be combined to strengthen the analyses.

There are two types of topical modules: fixed and variable. The fixed topical modules are designed to be conducted on a regular basis to augment the core data. They are considered necessary to meet the survey's goals and objectives. Although the topics are "fixed," the questions in these modules may be modified from time-to-time to accommodate conceptual changes or to make improvements in collecting these data. An example of a fixed topical module is the annual "round-up" module on earnings and benefits.

The variable topical modules are designed to satisfy the special programmatic needs of other Federal agencies. Time is set aside for variable modules in several waves to allow the flexibility to add content to meet special needs that develop as the survey continues. An example of a variable topical module is the child care topical module administered in the 1984 Panel. It was developed to obtain information about child care arrangements, such as who provides the care, the number of hours of care per week, where the care is provided, and the cost of the care. These data will be useful to other agencies because child care expenses are frequently deductible for program eligibility purposes. Variable topical modules may be repeated in subsequent waves or panels as necessary.

A wide variety of topics are covered under the aegis of the topical module concept. The breadth of these data ensure that SIPP will be a widely used and powerful data base serving multiple purposes.

In addition to the data collected by the survey questionnaire, the content may be supplemented with administrative record data that are difficult for respondents to recall such as lifetime earnings and program benefit histories. To facilitate future linkages with administrative records, steps have been taken in the SIPP to ensure that the social security number is obtained for as many persons as possible.

Links to Administrative Records and the Validation of Social Security Numbers

Background

The SIPP data system has always been thought of as a combination of data from administrative records and household surveys. This reduces respondent burden by using other data sources for difficult-to-obtain information. Interview responses can be supplemented by information from program files such as the earnings and benefit records of the Social Security Administration (SSA). This allows, for example, analysis of the long-term impact of various Social Security benefit formulas.

To make these linkages accurate, social security numbers (SSN) are required for sample individuals. The SSN is obtained for each household member in SIPP and recorded on the control card. It is identified as a critical survey data item requiring completion to make the interviewers aware of its importance. These efforts should result in improved accuracy of the survey-reported social security numbers. These numbers are then verified and corrected to maximize the number of accurate linkages to other record systems.

The verification and correction process builds on the work of the development program (the Income Surveys Development Program) in which SSNs were obtained and verified for more than 95 percent of the adult (15+) sample (see "Social Security Number Reporting, the Use of Administrative Records, and the Multiple Frame Design in the Income Survey Development Program" by D. Kasprzyk in David, M. (ed.), 1983). At the conclusion of each month's interviewing during the first wave of a SIPP panel, a special extract file is prepared by the Census Bureau for the SSA. This file contains a small number of key variables (SSN, name, date of birth, sex) for all original sample persons who report a SSN, including children, in a format appropriate for machine validation. Persons who report that they have a number but cannot supply it ~~or that they do not have a number~~ are handled separately in a clerical (manual) procedure. Persons who refuse to provide a SSN are not included in the search process. The SSA identifies (by machine validation) incorrectly reported numbers then ~~also~~ clerically resolves these cases along with cases with a missing SSN. This work is completed ~~by~~ the fourth wave interview, at which time a field followup is conducted to obtain missing SSNs (provided they are not "refusals") and to reconcile inconsistencies in SSN or demographic data generated by the computer match or the clerical resolution.

Social security numbers of persons who enter the sample after Wave 1 (because they start living with original sample people) are validated at the start of the next panel. ~~For example, information on new panel members (nonsample persons) from Waves 2 through 5 of the 1984 Panel was held and submitted for computer validation with Wave 1 of the 1985 Panel. Likewise, information on nonsample persons from Waves 6 through 8 of the 1984 Panel and Waves 2 through 4 of the 1985 Panel will be held and submitted for computer validation with Wave 1 of the 1986 Panel.~~

The following summarizes the SSN validation results for the 1984 Panel Wave 1 sample:

53,588	Total Wave 1 sample persons
- 1,674	Persons who refused to provide a SSN and were excluded from the validation process
51,914	Persons eligible for SSN validation
-42,128	Persons who reported a usable SSN and were eligible for computer validation
9,786	Persons who did not report a SSN and were eligible for the manual search (mostly children)
44,172	Validated SSNs (85% of eligible)
7,742	Unvalidated SSNs (mostly children)
51,914	Eligible for SSN validation

SSA Disability Survey ADD-On to the SIPP

The SSA needs information describing and explaining the economic and labor force status of disability insurance (DI) program participants in the periods immediately following the key programmatic events of awards, terminations, and denial of benefits. This information is not currently available. Social Security administrative data are not suitable for these analytical purposes because these records do not contain information on the program participants' family income and earnings before and after contact with the program. The data bases of national surveys are inadequate because they have not sampled enough disability program participants to support ~~accurate~~ reliable analysis of pertinent program issues. As a first step toward remedying this information gap, SSA proposes to conduct a feasibility study of the use of the Census Bureau's SIPP as a suitable data collection mechanism.

There are two goals of the study. First, in order to plan for an SSA-sponsored data collection effort in the future, SSA is testing the effectiveness of using SIPP to gather data for SSA.

The study will determine whether a national SSA panel can be integrated into regular SIPP field work and data processing activities. The SSA will investigate whether they can draw a sample of allowances from the newly developed national disability data system, obtain current addresses, and forward the information to the Census Bureau to meet the SIPP interview schedules. The Census Bureau will test its ability to administer special questions to the SSA sample persons without significant dislocation to standard SIPP field practices. The Census Bureau will also test whether the data can be accurately processed along with the regular SIPP effort in a timely manner. Second, the study results will describe and explain the economic and labor force status of newly awarded DI beneficiaries. The information on family income, labor force participation, job search, job offers, and employer accommodations bear directly on work incentives and employment. This information describes the processes of family economic adjustment and return to work of disabled persons. With these data, the SSA can target beneficiary subpopulations that can best benefit from work incentive program modifications. In addition, the SSA will have data for establishing a context within which to evaluate its forthcoming series of work incentive and vocational rehabilitation demonstration projects.

This project will sample 1,200 households with newly awarded DI beneficiaries as an add-on to the Census Bureau's 1988 SIPP. Sampled persons will be under age 45 and will be selected from specific geographic areas. Household members

will be interviewed during the first three waves of the 1988 SIPP panel. They will answer the regularly administered SIPP protocol, together with an additional set of special SSA project questions.

SSA/SIPP Data Linkage Project

SSA's interest in a data link ~~follow closely~~ parallels (or corresponds to) the intended uses of SIPP at its inception. A linked data set would enable the SSA to:

1. Estimate future program costs. -- The SSA is responsible for projecting program costs for all major SSA programs including: The Old Age, Survivor, and Disability Program, the Supplemental Security Income Program and Aid to Families with Dependent Children. In order to improve the accuracy of the projection methods, the SIPP panel data can be linked with a number of years of SSA data so that inflows and outflows can be analyzed in addition to point-in-time prevalence estimates of SSA program participation. The relationship between program participation and underlying individual characteristics can then be used to estimate future program costs and growth based on the SIPP data alone from future panels thus providing the SSA an early forecasting capability.

2. Assess the effects of program policy changes. -- An SSA-SIPP linkage will contain family, income and SSA benefit data. This combination of information will permit the SSA to estimate the programmatic costs of policy changes that depend on these factors and to assess the effects of policy changes on the economic well-being of program participants.

3. Describe non-programmatic characteristics of program participants. -- The SSA is frequently asked by Congress and others to provide information about program participants that is not routinely captured by administrative record systems. In the past, the SSA has used a series of widely spaced and usually one-time surveys to provide such information. The information that can be obtained is often out-of-date and the prospects for a new round of special purpose surveys are not good. An ongoing SSA-SIPP data link would provide relatively up-to-date data on a routine basis.

4. Test social science theories as they relate to Social Security programs. -- The longitudinal component of the SIPP's research design and the wealth of data captured in core questions and topical modules provide data that will be sufficiently rich to test many social and economic theories of program participation, thus making a significant contribution to the basic research that must accompany any dynamic social program.

In essence, SSA wants a maximal linkage with SIPP. For each SIPP panel, they want all waves of data, including core questions and topical modules linked to extracts of the basic SSA program records: The Master Beneficiary Record (MBR) which contains eligibiliity and benefit histories or the OASDI program, the Supplemental Security Record (SSR) which contains eligibility and benefit histories for the SSI program, and the Summary Earnings Record (SER) which contains a history of covered earnings for each worker. They will want to update the SSA records periodically so that each panel's files will contain additional years of the SSA's program data. They may also want to link to new administrative files in the disability area that are now being developed at the SSA on a regular basis. All initial and subsequent linkages will be by mutual agreement between the SSA and the Bureau of the Census.

While it is not possible to specify now, the format of the linked files, they must be constructed in such a way so as to be accessible via remote terminals from both Baltimore and Washington and on both UNIVAC and IBM systems.

The primary tasks in the linkage project are:

1. Verification of, and scouting for, Social Security numbers (SSN). -- This task is already a part of the SIPP project activities. In particular, the vast majority of SSN's for the 1984 SIPP panel have been processed by SSA staff.

2. Obtaining SSA administrative records. -- As mentioned above, the SSA has a definite interest in matching the MBR, SSR and SER to the SIPP. Decisions will have to be made about the content of the data extracts from these files that would be included in the match. Specialized computer programs may be required to complete the task.

   In the future, they want to consider adding data from additional SSA administrative records. They may also want to make arrangements with the Health Care Financing Administration to obtain Medicare utilization and cost data for a SIPP linkage.

3. Merging administrative records with SIPP survey data. -- The SSA does not see the matching tasks as one-time activities. Instead they anticipate a number of data processing operations for each SIPP panel.

   a. Development and execution of an initial matching operation for the panel (including the first 2 or 3 waves).

   b. Updating the matched file for additional waves of interviews and topical modules.

   c. Updating the matched file with later (or new) SSA administrative record extracts.

4. Weighting, imputation and sampling error estimation. -- We will have to consider and develop schemes for weighting and imputation that take into account non-matched SIPP records. Both cross-sectional and longitudinal weights will be required. The SSA would also need the capability for estimating sampling errors. Some sort of half sample replication approach would be most convenient.

5. Development of documentation for the matched files. -- Documentation for a matched file would include tape description and utilization information, the SIPP questionnaires and descriptions of the SSA administrative records, a sampling statement, editorial imputation descriptions and any other information required for estimation or analysis.

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 on the part of Congress, other policy makers, and researchers in areas such as health care, the elderly, and tax reform. While SIPP collects information on a respondent's contribution to retirement plans, it does not collect information

on the employer's contribution. Moreover, SIPP collects information on whether a person is covered by health insurance and whether the employer makes contributions, but 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 policies are provided through an employer. Amounts of employee payments and employer contributions are not obtained.

This study involves obtaining a signed release from the respondent at the interview and contacting the respondent's employer and asking the employer to fill out a short questionnaire to obtain data on both the employer's and employee's contributions to health insurance plans, pension plans, and life insurance plans. Information provided by the employer would supplement the SIPP data and improve data quality.

A half sample of one rotation group's households was selected for the study. The test was done in August 1987, (rotation group 4) for households in Wave 8 of the 1985 Panel. This was the last interview for these households.

The test included only employed persons, 18 years old and older, for whom a Wave 8 interview questionnaire was completed. Of the 1,352 persons eligible for the test, 563 persons (42 percent) signed the authorization form, 446 persons (33 percent) refused to sign, and 343 proxy or telephone respondents (25 percent) did not return the authorization form that was left/mailed to them. We did not conduct a followup of the refused or non-return cases.

Of the 563 questionnaires that were mailed to an employer, 424 (75 percent) were returned after the initial mailing and required no followup.