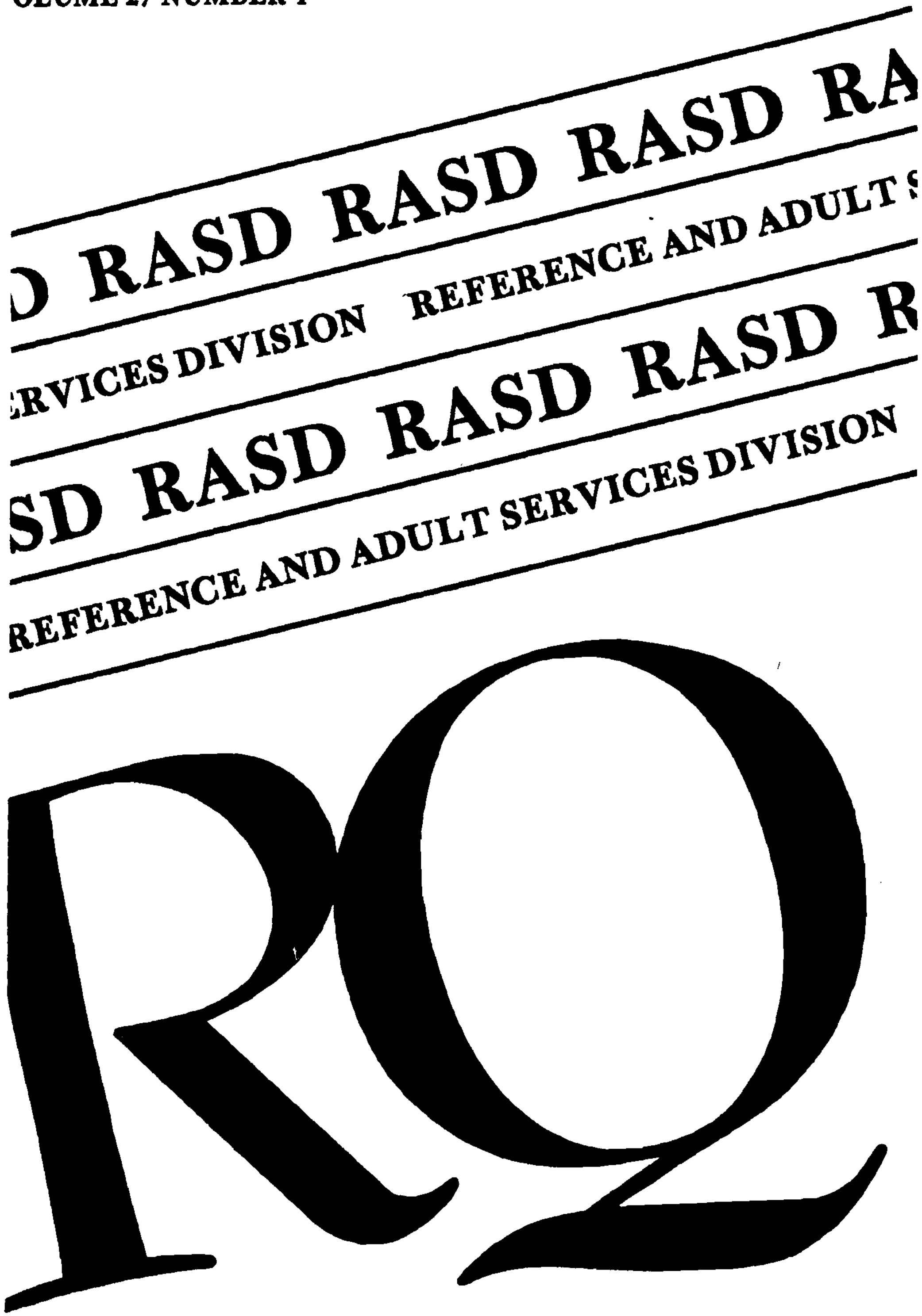
Robbin, A., & David, M. (1988). SIPP ACCESS: Information tools improve access to national longitudinal panel surveys. *Research Quarterly*, *27*, 499-515. (Also IRP Reprint Series #580.)

UMMER 1988 ISSN 0033-7072
OLUME 27 NUMBER 4


) RASD RASD RASD RA
RVICES DIVISION REFERENCE AND ADULT S
SD RASD RASD RASD R
REFERENCE AND ADULT SERVICES DIVISION
RQ

SIPP ACCESS: Information Tools Improve Access to National Longitudinal Panel Surveys

SIPP ACCESS represents an innovation in providing services for statistical data. A computer-based, integrated information system incorporates both the data and information about the data. SIPP ACCESS systematically links the technologies of laser disk, mainframe computer, microcomputer, and electronic networks and applies relational technology to create great efficiencies and lower the costs of storing, managing, retrieving, and transmitting data and information about complex statistical data collections. This information system has been applied to national longitudinal panel surveys. The article describes the reasons why SIPP ACCESS was created to improve access to these complex surveys and provides examples of tools that facilitate access to information about the contents of these large data sets.

Alice Robbin and Martin David

Alice Robbin is Associate Scientist, and Martin David is Professor of Economics at the Institute for Research on Poverty, University of Wisconsin–Madison. Submitted for review January 5, 1988; revised and accepted for publication February 6, 1988.

Information is costly and resources are finite; consequently, librarians perform a benefit-cost analysis to manage their resources more effectively. They consider, for example, utility, contribution to other information resources located in the library, and use. They weigh these benefits against the costs of acquisition, processing, description, access, and use.[1] An analogous benefit-cost ratio will be calculated when substantial investments are required for collecting statistical data.[2] A funding agency and data producer may accept the high costs of collecting, processing, and releasing statistical data if they are projected to yield new knowledge through scientific discoveries once the data are placed in the public domain. Data libraries will accept the high costs of acquisition, processing, description, control, distribution, and preservation for certain data files because they are deemed to have significant subsequent value for researchers and students.

We have some evidence to show, however, that widespread exploitation by the research community has not followed a substantial public and private investment in collecting and processing *national longitudinal panel surveys.*[3] Underutilization has been the norm, even when governmental agencies and research centers have *an explicit policy of data sharing* and the data are *publicly available* in data libraries and other repositories.[4] Despite the scope and significance of longitudinal panel studies and their conceptual, theoretical, and methodological contributions to the social sciences and public policy, use has been very low when compared to other available statistical data files in the public domain.[5]

Why has the enormous investment in these data not realized comparable widespread use? As consumers and service providers of statistical data, over the years we have noted great interest among colleagues and graduate students in the subject matter of these panel surveys but have seen that only a few experts had the resources to mine these data.

A brief answer is that these studies magnify all the problems that data libraries face with other data files. Their large scale and complexity and the lack of analysis tools for understanding the data meant that most researchers have been unable to exploit these surveys. For example, reference services are an important part of any library, but the lack of bibliographic access tools, standards, and control make it difficult and time-consuming to ascertain the contents of these data files. Lack of quality control over data products, either in the design, collection, or processing stages, makes it difficult for the user to access and retrieve data in an easy, efficient manner. Computer technology, data structures, and substantive knowledge of social science call for special expertise in the social sciences, library management, data management, and data processing, thus demanding personnel who have special skills not ordinarily obtained in the formal requisites for a degree in library science, computer science, or a social science subject area.[6]

This article is about strategies that we have devised to solve the access, retrieval, and information problems posed by longitudinal panel surveys. The context of the discussion is SIPP ACCESS, a data center and research network for the Survey of Income and Program Participation (SIPP) and Income Survey Development Program (ISDP). SIPP ACCESS was funded for thirty months in late 1984 by the National Science Foundation (NSF); in July 1987 the project received a new twenty-four-month grant. By early 1987 we saw that considerable progress had been made in transferring mainframe technology to the microcomputer. This resulted in a proposal to the Sloan Foundation for seed money to design and distribute microcomputer data and information products that derive from our NSF grant work.[7] Although this project has multiple objectives, our discussion concentrates on one aspect: *how to apply relational database technology to facilitate access to information about large-scale statistical data collections.*

In the first section we identify some of the characteristics of national longitudinal surveys that impede access. The second section provides an overview of the conceptual design and physical components of the facility. SIPP ACCESS uses a relational database management system (RDBMS)—Ingres—to organize, store, retrieve, and describe the *1984 Survey of Income and Program Participation.*[8] In the third section we illustrate the RDBMS improvement in the organization of information about and access to the SIPP data. The data and metadata assists that SIPP ACCESS provides reflect the analysis of the problems with longitudinal panel surveys that we describe in part one.

WHY LONGITUDINAL PANEL SURVEYS ARE HARD TO USE

The barriers to using longitudinal panel surveys include their very high access and retrieval costs, a lack of information about the data source, and the inability to deal with the complexities of the design. We discuss the central problems of such surveys.

Size and Scope

Longitudinal panel studies typically interview many thousands of people and contain thousands of variables. The problems faced by the researcher are in identifying and selecting the subpopulation of interest and retrieving variables to solve the data user's problem. Most file-handling software used to restrict the sample and retrieve selected data employs sequential processing, an inefficient method of reading and retrieving data. Many magnetic tapes must be mounted for each interview. Complicated data management tasks often entail hiring a full-time, experienced programmer for at least a year to write extensive code.

Replicated Measures

The longitudinal survey has a core set of questions that are regularly asked, so that changes in the life of the sample individuals can be examined. The problem of identifying which measurements are replicates and which are unique is often complicated by the fact that different names are assigned to the same measurement over the life of the panel.

Data Structure

Longitudinality itself implies a complex data structure. A sample is drawn, and people are interviewed for the first time. Then they are reinterviewed periodically. People enter and leave the survey. Analysis entails finding out if the sample is large enough and keeping track of everyone over the period of the survey—both are difficult. Programmer assistance is likely to be needed to extract responses from one interview to the next. Matching people across interviews to create a longitudinal sample requires expensive data processing on a mainframe. The way that the data are released also complicates the analysis. Public use files are released for each interview period, creating a series of cross-sectional observations that impede understanding the longitudinality of the measurements. Two central problems, therefore, are identifying those who are in the sample at one time and determining how long they remain in it over the length of the panel.

Missing Data

Sometimes, the interviewee does not answer all the questions, and this results in missing data that must be accounted for when statistical analysis is carried out. The data producer will alert the researcher to the existence of missing data with special codes. Sometimes, the nonresponse is replaced by some value in a procedure known as *imputation.*[9] The problems for the researcher are identifying missing data, in order to establish which data are real and when they are imputed, and linking the original measurements to their imputation flag, so that they can decide how to handle the nonresponse when carrying out statistical analysis.

Conditioning of Survey Questions

Understanding the questionnaire design and who is "sample-relevant" is extremely important. Many of the questions asked by the interviewer are often not relevant for every person participating in the survey, so that "not-in-universe" (NIU) or "not applicable" codes are generated for a particular questionnaire item. The interviewer will be instructed to skip over a question or set of questions if they do not apply to a particular respondent. It becomes very important to know under what conditions a question is asked and when it is skipped, i.e., for whom a particular question is relevant. Complex survey designs often incorporate many conditioning statements, and it becomes hard to follow the "skip patterns" that are generated as a result.

Conditioning becomes even more complex when examined from a longitudinal perspective. After a previous interview many items are updated by noting changes since the last interview. Failure to appreciate this can lead to erroneous conclusions. For example, in the *1984 SIPP* one might conclude that the number of persons receiving Medicare is small, because only the persons attaining age sixty-five since the last interview will be asked about their eligibility. Furthermore, the same question may be asked in five separate interviews, but the question order can change.

In other words, one problem that the researcher faces is knowing the context for the question, that is, the relationship of one question to a previously asked question. Another problem is knowing for whom a particular question is relevant. This latter problem is related to drawing the correct subpopulation, which we identified in the "Size and Scope" paragraph above.

Description of the Design

These problems of data structure make access to and retrieval of longitudinal panel surveys problematic, even for experts. In most cases, hundreds of pages of description are needed to explain survey design, drawing of the sample, questionnaire structure, field work and interviewing, coding decisions, data processing, and so on.

Nevertheless, a great deal of the understanding required to use these data can be handled, with appropriate description and careful organization of the data, in a way that reveals their underlying structure. Serious lacunae in the written material (documentation) describing most longitudinal data sets have meant, however, that people not originally associated with the data collection (secondary analysts) have been at a substantial disadvantage in understanding the complexity of the data.[10] The central problem faced by the data producer is how to communicate knowledge about the survey design, both in descriptive materials and in the data structure of the public-use files.

Summary

Size and scope of the data collection, multiple interviews, periodic measurements, mobility of the sample, and structure of the survey instrument impede information and data seeking, understanding of the design, and retrieval of selected data for solving a research problem. This discussion about the attributes of longitudinal panel surveys provides the framework for the proposal we made to NSF. Our description of the complexity of longitudinal panel surveys also explains our decision to locate the data in a relational database management environment in order to improve retrieval and clarify the underlying meaning of these panel data. Finally, it serves as essential background information for understanding the information tools we designed to improve access to the contents of longitudinal panel surveys.

DESIGN OF SIPP ACCESS

Conceptual Framework

SIPP ACCESS represents an evolution in statistical data and reference services provided by data libraries and archives since the 1960s; however, its design and method of facilitating access to the *1984 SIPP* suggest a different model of a data library.[11] The SIPP ACCESS model specifies

Integrated information system environment for data and information about data;

High level of staff expertise and responsibility for (a) developing the conceptual and technical tools that facilitate access to the data and information about the data and (b) training novices to understand the data and the information system;

Involvement with a discrete number of complex statistical databases that have underlying structural, theoretical, or methodological similarities;

Enlargement of the information service provider's typical set of relationships to include the data producer, data center staff, and researcher and to provide a vital communications link between and among them; and,

Use of recent developments in computer and communications technologies, including remote access computing, relational database management systems, computer networking, optical archival storage on laser disks, and microcomputers.

SIPP ACCESS shares many of the attributes of the highly successful *observatory* concept in the natural and physical sciences. It is a national facility that supports retrieval from a complex database with a capacity for archiving and retrieving data obtained through periodic measurements; it is organized around a set of technologies maintained by a permanent staff; and it tailors its activities to the research needs of a community of investigators that uses the observatory for a variety of specialized studies.

The observatory acts as a central node in a scientific network of researchers who analyze the SIPP. Information is exchanged among researchers, data producers, and SIPP ACCESS staff, and scientific discoveries are made. Feedback to the data producer from the staff and community of researchers results in corrections to the instrument.[12] SIPP ACCESS applies the successful experiences of the scientists' electronic communications networks and establishes a research network where applications of the data and discoveries of users are linked.[13]

The staff take active roles as experts, teachers, and intermediaries. The SIPP ACCESS staff manage the instrument—the panel survey data known as the *1984 SIPP* conducted by the U.S. Bureau of the Census—and the information needed to understand the survey data. They take responsibility for training researchers to use the technologies associated with the observatory, so that they make efficient and correct use of the instrument when conducting their "experiments."

Computer and Communications Technologies

Applying state-of-the-art computer and communications technologies to improve access to, clarify the meaning of, store, manage, retrieve, and transmit the results of discoveries about statistical data has been a principal objective of SIPP ACCESS. The systematically linked technologies of laser disk, mainframe computer, and microcomputer create great efficiencies and lower the costs of these activities for researchers who use the data and staff who manage the facility. Electronic networks speed communications—between data producer, SIPP ACCESS staff, and researchers—and transfer of data from one site to another in a distributed computer networking environment. A "clone" of the mainframe database will be available on microcomputer to serve as a "cost-free" environment for preparing and debugging command files, preparing data and text files to be uploaded to the mainframe, and downloading extracts from the mainframe database that are subsequently used for statistical analysis.

Data from the public use files are stored in a relational database on a VAX/VMS computer at the University of Wisconsin Physical Sciences Laboratory. Among the characteristics that lend themselves to many data management tasks carried out by researchers, the relational database makes it possible to store data on the basis of clear semantic principles, assures quality control of the data, and provides users with a shared data storage and common referencing system. Another important attribute that responds to the description problem we identified in part one is the RDBMS's self-documenting capability: a description of the structure and contents of the database is permanently maintained in its system catalogs, and all work that is carried out can be permanently recorded. The RDBMS offers a powerful capability for easy, inexpensive reorganization of the data, a task that requires high data processing costs and a great deal of programming time in the more traditional file-handling environment.[14]

The RDBMS uses artificial intelligence to carry out efficient retrievals. Random access of the data by the variable's (*attribute*) name means rapid retrieval of data in contrast to sequential access of the magnetic tape environment. For example, small subsets of the population pertaining to rare events, like job loss or marital disruption, can be located efficiently. The capacity of the RDBMS to retrieve small portions of the data by direct addressing and other "tricks" of data storage permits extremely cost-effective access of these populations. The query language, based on set theory, is simple for users to learn.[15] In addition, the entire database is portable because it consists of rectangular files of rows and columns and is easily transported to other sites and RDBMS.

The *1984 SIPP* is a very large database. The complete data set stored in the computer would fill four VAX hard disks, 1.5 gigabytes of data. Recent developments in laser disk technology, however, make it possible to store the complete survey (called *SIPPRUN*) on optical archive disk. The Physical Sciences Laboratory's engineers and computer scientists have designed a completely automated process for accessing data sorted on laser disks. SIPP ACCESS provides software for researchers to retrieve any part of the database stored on the optical disk with only three commands. The ROM available on the disk partitions the data set into small and manageable components.

Carrying out data management tasks interactively on the enormous set of data (the complete sample of sample units and their associated household family, and person records) would be prohibitively expensive and time-consuming. Instead, SIPP ACCESS created a 2 percent sample of sample units (called *SIPPTEST*), which is maintained online twenty-four hours a day. In this way users can interactively learn about the *SIPP,* experiment with the data, debug their command files, and test hypotheses very inexpensively.[16] The 2 percent sample of sample units contains all their household, family, and person records and all the variables associated with these record types. The *SIPPTEST* and *SIPPRUN* databases are structured in exactly the same way, so that work done on the 2 percent sample can be applied to the complete sample: after completing interactive work on the 2 percent sample, data management tasks on the complete sample can be submitted for low-cost, overnight batch processing.

These data are available to researchers around the world through remote access and via the electronic networks of BITNET and ARPANET. Electronic communication is cost-effective and efficient: exchanging information via messages and bulletin boards is less expensive than conventional mail, delivery is speeded, and information can be shared by a number of people. SIPP ACCESS created a friendly, online consultant known as SIPPASSIST, who responds to electronic mail requests for assistance and information. Computer networking makes it possible for users at one node to work on alternate hosts. Distributed computing facilitates the appropriate allocation of scarce computer resources.

Integrated Information System Concept

We have interpreted this concept to mean a capacity to inform analysts about the scientific design of SIPP and the collecting of data in the field during the interview process. In addition, the information system provides a facility for producing statistical output that includes both data and software. The data are both statistical and textual.

The statistical data are the original data provided by the Bureau of the Census *and* restructured data prepared by SIPP ACCESS to clarify the meaning of the original measurements and to help researchers carry out time series and event history analysis. For example, we commented earlier in our discussion about the difficulty of knowing who is in the sample, for what length of time, and whether interviews took place (see "Size and Scope" and "Data Structure" paragraphs in part one). One of the most important investments we made was in development of a file (a *table* or *relation* in relational database terminology) to record entrance, exit, and duration of stay in the panel. We call this table *Retention,* and it has become essential for all work related to drawing a longitudinal sample and knowing whether the personal data reported in the file are real or imputed.

Project-developed software provides users with information about computing and database resource costs, retrieving data from the database and optical archive storage, linking the VAX operating system and the Ingres database, and current status of changes to the database. We also automated the collection of statistical information on use of the SIPP ACCESS facility and the *SIPPTEST* and *SIPPRUN* databases, and this information will be used in the evaluation phase of the project beginning in July 1988.

The VAX has easy-to-use, menu-driven, help libraries. SIPP ACCESS designed *SIPPHELP* to introduce users to the structure of SIPP ACCESS and the contents of the databases and to broadcast important news about user and staff discoveries regarding the data and new developments in the database. This help library acts as a permanent "memory" of the discoveries that have been made, so that the learning period required to make correct use of the data is reduced.

The simple directory structure of the VAX operating system is used for organizing the storage of textual information about the database and about techniques for using RDBMS. For example, users will find online files of the workshop materials with which we teach the SIPP design, use of relational database, the VAX/VMS operating system, applications of RDBMS to the SIPP, and so on. Users will also find files of the data dictionaries (codebooks), reference materials, policies, error status reports, and users notes. These materials are regularly updated, and users are notified of changes in the reference libraries through electronic mail and broadcast messages.

It is, as noted earlier, very important for users to have a better understanding of the scientific design of SIPP and its implications for the meaning of observations. For that reason, we have focused the power of the RDBMS on *metadata,* information about the measurements that are the subject of statistical study. This focus is not a new one for an information system: pathbreaking work by the Zentralarchiv at the University of Cologne[17] and by J. J. Card[18] led them to construct a link between the questions and the measurements. What is new about the efforts made by SIPP ACCESS, however, is that the linkage has been constructed on relational principles and consideration given to the context in which questions are asked. (These points will become clearer in the next section when we describe the *varname* and *predecessor* tables.)

We noted (in "Description of the Design" in the first section) that a central problem facing researchers was lack of information about survey design and results of data analysis. One way that SIPP ACCESS addresses this problem is by serving as a central library for documentation and for published and unpublished articles on SIPP and ISDP. SIPP ACCESS also archives internal, unpublished technical memoranda produced during the design, fieldwork, and data processing stages by the Bureau of the Census. The technical memoranda are reviewed; important information is extracted and maintained in the reference libraries and also broadcast to the research network. Citations from the articles and technical memoranda are entered in tables in the *SIPPTEST* database and may be retrieved with the same query language used for the numeric data.[19] The *author* table is the authority file for all authors and corporate bodies. The author, publications, and technical memoranda tables thus provide researchers with several access routes for locating a document.[20]

Metadata tools inside the database provide information about the sample and its data, survey instruction, data dictionaries, and location of the data in the database. The relational database system catalog of attribute names serves as the key linking device for retrieving information from the tables that contain a computer mapping of the survey instrument, data dictionaries, and attribute locations in the database. We illustrate some of these metadata tools in the next section.

SOLUTIONS: METADATA TOOLS IN A RELATIONSHIP DATA BASE ENVIRONMENT

SIPP ACCESS has created four files that clarify the structure of the scientific design, implementation of RDBMS, and context for each variable in the database. Because these relations are themselves part of the database, the same retrieval language applied to the variables (*attributes* in relational parlance) is available for understanding the basic tools of the questionnaire and the data dictionary, locating attributes in the database, and retrieving imputation flags associated with the original measurement. Print and online documentation outside the database explain how to use these tables.

We illustrate relational solutions to the problems of size and scope, replicated measures, missing data, and conditioning of questions that we discussed in the first section. The retrievals also reflect multiple access routes that researchers take to a data set: via the data dictionary, the survey instrument, and directly into the data set. We use the example of a researcher's interest in knowing about food stamp recipiency. The appendix includes an excerpt from the portion of the second interview questionnaire where the food stamp coverage questions are located.

The Varlist Table

In part one we described the aspects of national longitudinal surveys that make access and retrieval tasks more complicated than they are in other data collections. One of the problems we noted was that their size and scope complicate the task of locating variables of interest. For example, the *1984 SIPP* contains about 13,000 variables that are replicated measurements. Ordinarily, it is very time-consuming to locate replicates. Attribute names in the database must be unambiguously linked to their synonyms in other materials related to the data collection and design. Further, the location of the attributes must be identified, so that the researcher can retrieve them.

The relational technology performs these functions easily in a table known as *varlist.* This relation, a concordance between the variable names that appear in the Bureau of the Census data dictionaries (column 1) and the SIPP ACCESS database (columns 2,3) for the replicated measures, provides information on the location of the attribute in the database (column 4) and the wave in which the attribute appears (column 5). The relational structure compactly organizes this description in a narrow table of about 1,300 rows. Linking the table root and the wave (interview) number provides the unique location of every attribute-wave in the data base.

Figure 1 shows the retrieval of the Bureau of the Census data dictionary variable name, sc3100: *"Were all the people living here covered under _____'s food stamp allotment?"* Note the simple syntax of the query language. We also developed conventions for naming the tables.[21] (An explanation of

```
*retrieve (varlist.all) where varlist.bureau="sc3100"

  (1)          (2)           (3)         (4)         (5)
|bureau      |attname      |attrange|tableroot  |waves     |
|--------------------------------------------------------- |
|sc3100      |sc3100       |        |ga_atbl    |123456789 |
|--------------------------------------------------------- |
(1 row)
```

Fig. 1. A Retrieval from the Varlist Table

these conventions is available online in the *SIPPHELP* library.)[22] Although we have entered the complete table name, we could also give that table a one-character alias to reduce data entry.

The Implist Table

We noted previously the importance of knowing which values are real (derive from the responses of the interviewee) and which have been imputed by the data producer. The second metadata simplification is a relation, *implist,* that cross-references the imputation flags with the attributes to which they refer. The same naming conventions are used to parameterize the relation name associated with these flags, so that replication is again obvious.

We illustrate this with a retrieval of the questionnaire item on the total dollar amount of food stamps received in the last month (prior to the interview), item sc3124. From another retrieval, we learned that the Bureau of the Census did not impute item sc3100 (above); however, an examination of the *implist* table (see figure 2) shows that dollar amounts are imputed if a respondent indicated an income receipt but either did not know or refused to say how much.

The researcher could also obtain information about how sc3124 was coded without leaving the database. The statement, "doc sc3124" retrieves both the row from the *varlist* table and the section of the data dictionary that applies to sc3124 (see figure 3).

The Predecessor Table

In our discussion of "Conditioning of Survey Questions" in the first section, we noted the importance of understanding the conditions that cause particular items to be omitted and that estimates from the population for whom the question was asked could not be meaningfully constructed without this understanding. The *predecessor* relation is a mapping of the questionnaire instrument. The relation encodes the instructions to the interviewers that determine which questions are asked.

Predecessor contains four essential attributes: "*attname,*" "*value,*" "*result,*" and "*waves.*" "Attname" is the attribute name in the RDBMS that is used for a particular response in a particular questionnaire instrument. "Attname" and "waves" uniquely identify a questionnaire item. "Value" indicates the response recorded for that attname. "Result" shows the location of the next relevant response location, whenever that location is not the immediately following questionnaire item. Other attributes in the table provide textual information on the nature of the conditioning, describe whether the information is provided by the respondent or transcribed by the enumerator, and define the ordering of the particular item in the flow of questions; these are omitted in our example.

In our next example, we retrieve those questionnaire items that would lead to a skip over questionnaire item sc3100. We do this by including a statement containing the name of the attribute and the wave with which it is associated. The computer first returns the name of the variable and the wave, and then it displays the questionnaire items that lead the interviewer not to ask question item sc3100. The database retrieved ten questionnaire items that would lead to a skip over sc3100 when we retrieved items for wave two. We then performed the retrieval on item sc3100 for wave four (see figure 4). The relevant pages of the questionnaire instrument are displayed in the appendix A.

The questionnaire item sc3100 would not be asked if the respondent had an-

```
*retrieve (implist.all) where implist.attname="sc3124"

|attname   |end_range|imp_flag |tableroot |waves    |
|---------------------------------------------------|
|sc3124    |         |g1imp10  |ga_atbl   |123456789|
|---------------------------------------------------|
(1 row)
```

Fig. 2. A Retrieval from the Implist Table

```
*doc sc3124

|bureau     |attname    |attrange|tableroot  |waves     |
|-------------------------------------------------------|
|sc3124     |sc3124     |        |ga_atbl    |123456789 |
|-------------------------------------------------------|

|description                                      |entry|row    |
|-----------------------------------------------------------------|
|SC3124                                           | 1176|      1|
|What was the total amount                        | 1176|      2|
|last month                                       | 1176|      3|
|Range = 0,800.  In dollars.                      | 1176|      4|
|U Household members, including children,         | 1176|      5|
|  covered by food stamp allotment                | 1176|      6|
|V 000 .Not in universe                           | 1176|      7|
|-----------------------------------------------------------------|
```

Fig. 3. A Retrieval from the Data Dictionary inside the Database

```
* pred sc3100 2\g
Executing . . .

|attname  |value |result  |waves      |
|---------------------------------------|
|sc1704   |     2|sc4800  |123 56 89  |
|sc3002   |     2|sc3138  |123456789  |
|sc3014   |     1|sc3200  |123456789  |
|sc3032   |     3|sc3200  |123456789  |
|sc3056   |     2|sc3200  |123456789  |
|sc3058   |     1|sc3200  | 234       |
|sc3060   |   999|sc3200  |123456789  |
|sc3068   |     2|sc3200  |12345      |
|sc3086   |     1|sc3200  |123456789  |
|---------------------------------------|
continue

*pred sc3100 4\g
Executing . . .

|attname  |value |result   |waves     |
|---------------------------------------|
|sc1704   |     2|tm4a8200|    4      |
|. . .                                  |
|---------------------------------------|
```

Fig. 4. Two Retrievals from the Predecessor Table

swered no (code of 2) to questionnaire item sc1704 for the retrieval on wave two. The remaining skips apply when this sequence of questions is used to elicit data other than food stamp income. For example, item sc3002, code of 2, refers to the WIC income source, and therefore would have generated a skip over the food stamp questions in this series. The response to item sc3058 is conditioned on information received in a previous interview; a response of yes (code of 1) would generate a skip over item sc3100. Note that three attribute-result pairs were not asked in all nine interviews (waves) of the panel. Note that sc1704 does not result in a skip to sc4800 in every interview (wave) of the panel, and that item sc3058 with a code of 1 generates a skip to sc3200 in only the second, third, and fourth interviews of the survey.

We then performed the same retrieval on item sc3100 for wave four. In part 2 of figure 4, we include only one row and exclude all rows where wave four appears when we retrieved items sc3100 for wave two. We find that the item sc1704 results in a different skip, to another questionnaire item, tm4a1820. This information informs us that a code of 2 would lead the enumerator to ask a question in a supplementary part of the questionnaire known as a *topical module*.

Varname Table

The *varname* table was constructed to respond to a number of issues related to retrieving information from longitudinal panel surveys. The size of the data collection has meant difficulties in identifying variables with the same content (topic). The varname simplifies the task by recording an English label (column 3) for each attribute (column 1). In addition, it displays a number of properties of the attribute that cannot always be ascertained from reading the data dictionary or the questionnaire. These include the population for which the attribute is relevant (column 2), the periodicity of the measure (column 4); whether the measure is constructed by the data producer or represents the original response given at the time of the interview (column 5); whether the measure sets an income-source flag that conditions subsequent questions asked by the interviewer (column 6); and the referent (unit of analysis), when the referent is other than the person being interviewed (column 7).

The English label is constructed to classify observations into a variety of subject matter domains. We created a controlled vocabulary that is unique to the *1984 SIPP*. This classification can be searched to produce a list of attributes that refer to a common topic and can also be used as the basis for a thesaurus.

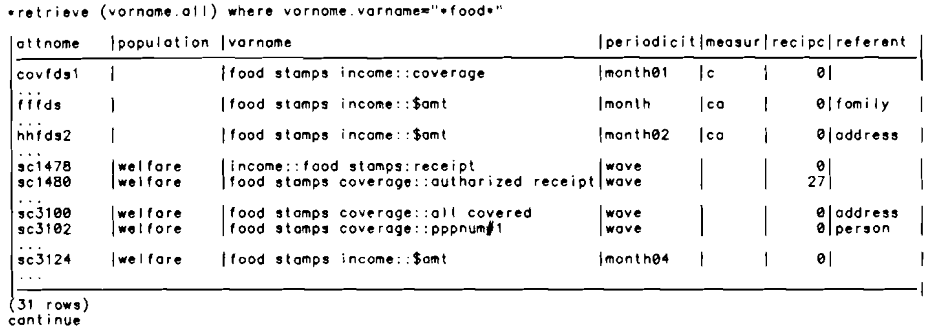
In figure 5, we retrieved all attributes that contain the term *food stamps*. Thirty-one attributes were returned, of which we show only eight.

The example illustrates a number of different aspects of the SIPP. The varname attribute on which the query was performed provides an opportunity for multiple access routes to information. In addition to supplying information about food stamps, we also are informed about recipiency, dollar amounts, and coverage of the respondent and members of the unit. This vocabulary has been standardized for all sources of income that are described in SIPP.

The *attname* and *measure* attributes indicate that, in addition to the original questionnaire items (coded with a prefix of *sc*) contained in the public use files, the bureau constructed variables from the original responses. For example, *fffds* is constructed (code of *c*) and also represents an aggregated measure (code of *a*) of all the members of the family unit. Item sc1478 represents an original questionnaire item.

The *population* attribute indicates that five of the variables are limited to members of the welfare population. The *periodicity* attribute shows us that the variables actually refer to different time periods: any month or a specific month for the period that the enumerator obtains answers from a respondent. For example, *covfds1* refers to the first month of the series of months in a reference period; whereas, item sc1480 is relevant for the interview month (*wave*). A *recipcode* of 27 for item SC1480 is a "flag" indicating that the enumerator will ask the respondent about food stamp recipiency, including who is covered, for what period, and dollar amounts later on in the interview.

Note that four of the questions asked of the respondent actually refer to five different referents (units of analysis), including the respondent. For example, item sc3100 refers to the address where the interview took place, whereas item sc3124 refers to the respondent (a blank column represents the default for the respondent). The unit of analysis of the attribute *fffds* is the family, whereas item sc3124 is any person who is part of the food stamp unit.



```
*retrieve (varname.all) where varname.varname="*food*"

|attname    |population |varname                                   |periodicit|measur|recipc|referent |
|-------------------------------------------------------------------------------------------------------|
|covfds1    |           |food stamps income::coverage              |month01   |c     |     0|         |
|...
|fffds      |           |food stamps income::$amt                  |month     |ca    |     0|family   |
|...
|hhfds2     |           |food stamps income::$amt                  |month02   |ca    |     0|address  |
|...
|sc1478     |welfare    |income::food stamps:receipt               |wave      |      |     0|         |
|sc1480     |welfare    |food stamps coverage::autharized receipt  |wave      |      |    27|         |
|...
|sc3100     |welfare    |food stamps coverage::all covered         |wave      |      |     0|address  |
|sc3102     |welfare    |food stamps coverage::pppnum#1            |wave      |      |     0|person   |
|...
|sc3124     |welfare    |food stamps income::$amt                  |month04   |      |     0|         |
|...
|-------------------------------------------------------------------------------------------------------|
(31 rows)
continue
```

Fig. 5. A Retrieval from the Varname Table

CONCLUDING REMARKS

SIPP ACCESS takes a systematic approach to the problem of retrieving data from a national longitudinal panel survey, the *1984 Survey of Income and Program Participation*. We have developed a computer-based model of an integrated information system that incorporates both the data and the scientific understanding of those measurements. The system employs a parsimonious computer language for describing the data collection and the results of scientific investigation. This parsimony is made possible by incorporating all the information in a relational database management system. We believe this approach represents an innovation for providing data and reference services and will be applied to other large, complex data collections.

There is an important aspect of this innovation that we have not addressed but is implicit in our discussion. In the future social measurement will depend more on linguistics and computer and information science. These disciplines will be explicitly applied to designing data structures that clarify the complex empirical world captured in the social research process and to the organization and retrieval of information about these data.[23]

Postscript: July 1, 1988.

SIPP ACCESS has consolidated the information retrieval assists into a more powerful form. A sample of the data in a new format suitable for study on the AT-compatible PC is being released. The data set PC-SIPPTEST occupies approximately 12-mb and can be used with the PC version of Ingres. ■■

NOTES AND REFERENCES

1. See Martin M. Cummings, *The Economics of Research Libraries* (Washington, D.C.: Council on Library Resources, 1986).
2. For a discussion of the benefits and costs of data sharing, see Stephen E. Fienberg, Margaret E. Martin, and Miron L. Straf, eds., *Sharing Research Data* (Washington, D.C.: National Academy Pr., 1985).
3. In a longitudinal panel survey the same people are repeatedly interviewed over a period of time. Interviewing may be carried out once every few months, once a year, or once every two years over a time span as short as thirty-six months or over a much longer period, such as twenty years. Well-known national longitudinal panel surveys include *The Panel Study of Income Dynamics; National Longitudinal Survey of the High School Class of 1972; High School and Beyond;* and *National Longitudinal Surveys of Labor Market Experience.*

[4] Discussions on the problems of access to statistical data include Martin H. David, *The Language of Panel Data and Lacunae in Communication about Data,* Center for Demography and Ecology, Working Paper no.85-20 (Madison: Univ. of Wisconsin, May 1985); Martin H. David, "The Frustration and Utopia of the Researcher," *Review of Public Data Use* 8:327-37 (1980); Martin David and Alice Robbin, "The Great Rift: Gaps between Administrative Records and Knowledge Created through Secondary Analysis," *Review of Public Data Use* 9:153-66 (1981); Alice Robbin, "Strategies for Improving Utilization of Computerized Statistical Data by the Social Science Community," *International Journal of Social Science Information Studies* 1:89-110 (1981); Alice Robbin, *Strategies for Increasing the Use of Statistical Data,* Occasional Paper Series (Urbana: Univ. of Illinois Graduate School of Library and Information Science, 1983). For an extensive discussion of the rationale for developing data libraries and the critical issues that confront the library community in providing statistical data services, see Kathleen Heim, ed., *Library Trends* (Winter 1982), a special issue on data libraries and the social sciences that includes a discussion by Alice Robbin on the Data and Program Library Service at the University of Wisconsin-Madison as a case study in organizing special libraries of computer-readable statistical data.

[5] Records at the University of Wisconsin's Data and Program Library Service on data use and counts of bibliographic citations in the literature indicate a low level of use compared to cross-sectional data files, in which a sample was drawn and people were interviewed once. For a discussion of underutilization of national longitudinal panel surveys, see Robert F. Boruch and Robert W. Pearson, *The Comparative Evaluation of Longitudinal Surveys. A Report Submitted to the Measurement Methods and Data Improvement Program of the National Science Foundation by the Working Group on the Comparative Evaluation of Longitudinal Surveys, Social Science Research Council* (New York: Social Science Research Council, Dec. 1985). See also Frank Stafford, "Forestalling the Demise of Empirical Economics: The Role of Microdata in Labor Economics Research," in Orley Ashenfelter and Robert Layard, eds., *Handbook of Labor Economis,* V.1 (Amsterdam: Elsevier, 1986.)

[6] These problems have been widely discussed in the library literature. See Heim (reference 4) for a collection of discussions on how data libraries have addressed these problems. Sue A. Dodd has written extensively on cataloging machine-readable data files, and two full-length books have been published since the early 1980s. See, for example, "Building an On-Line Bibliographic/MARC Resource Data Base for Machine-Readable Data Files," *Journal of Library Automation* 12:6-21 (Mar. 1979) and *Cataloging Machine-Readable Data Files* (Chapel Hill: Univ. of North Carolina, 1983). For another perspective on bibliographic access tools, see Paul E. Peters, "Notes on the Distribution of Labor in a Social Sciences Data Information Network," *IASSIST Newsletter* 2:69-76 (Summer 1978). For problems of data quality, documentation, and needs for standards and control, see Alice Robbin, "Managing Information Accessed through Documentation of the Data Base," *SIGSOC Bulletin* 6:56-68 (Fall/Winter 1974-75).

[7] SIPP ACCESS is supported by the National Science Foundation (grant no.8411785) and by the Sloan Foundation (grant no.B1987-46). The authors are principal investigators and co-directors of SIPP ACCESS. For further information, write the Institute for Research on Poverty, Social Science Building, 1180 Observatory Dr., University of Wisconsin, Madison, WI 53706.

[8] The Survey of Income and Program Participation (SIPP), conducted by the Bureau of the Census, is a series of panel studies beginning in 1984, and a major scientific research tool designed to improve our understanding of fundamental social processes. SIPP monitors short-term changes in the economic situations of persons, households, and families in the United States by gathering data on family formation and dissolution, job changes, earning of income from labor and capital, and government's role in economic well-being and financial security. In the *1984 SIPP* more than 64,000 persons fifteen years and older were interviewed every four months over a thirty-two or thirty-six-month period. SIPP records the events, changes, and contexts of their economic lives—working, seeking income maintenance, and paying taxes. The *1984 SIPP* data file contains about 20,000 variables. The replicated items are released in a "relational format" on twenty-seven magnetic tapes written at 9,600 bpi. Users of the supplementary surveys, called *topical modules,* can expect to acquire twenty-one additional reels of magnetic tape.

[9] For example, the Bureau of the Census uses a statistical procedure to replace the missing response. The bureau inserts a statistically generated value into the data file for that person but informs the user that the original response was missing and the inserted value is not "real" data. The information informing the user is called an *imputation flag.*

[10] For a more extensive discussion of documentation, see Alice Robbin, "Technical Guidelines for Preparing and Documenting Data," in *Reanalyzing Program Evaluation: Policies and Practices for Secondary Analysis of Social and Educational Programs,* ed. Robert F. Boruch and David Cordray (San Francisco: Jossey-Bass, 1981), p.84-143; Martin H. David and others, "Standards for Public Use Files: Lessons from SIPP," paper presented at the Social Science Research Council meeting on the potential of SIPP, Washington, D.C., June 1985.

[11] A more detailed description of the conceptual design of SIPP ACCESS is available in Martin H. David, "Designing a Data Center for SIPP: An Observatory for the Social Sciences," *Proceedings of the American Statistical Association* (Washington, D.C.: American Statistical Assn., 1985); Alice Robbin, "Data Rich But Communication Poor: Creating a Public Utility for the Survey of Income and Program Participation," paper presented at the annual meeting of the American Association of Public Opinion Research, St. Petersburg Beach, Fla., May 16-18, 1986; Alice Robbin, Martin David, and Thomas S. Flory, "Facilitating Complex Data Management Tasks, paper delivered at the annual meeting of the Population Association of America, San Francisco, Apr. 3-5, 1986; and Martin David, Alice Robbin, and Thomas Flory, "Access to Data: Handling the 1984 Survey of Income and Program Participation," paper presented at the annual meeting of the American Statistical Association, San Francisco, Aug. 1987 (to be published in abbreviated form in the *Proceedings of the American Statistical Association, 1987*).

[12] We want to emphasize what we believe is an important objective of SIPP ACCESS: to serve as a communications vehicle that holds the potential for improving the quality of data produced by statistical agencies of the federal government. This objective of serving as a proactive intermediary between data producer and consumer is consistent with the relationship that the research community and the federal government have had for more than forty years. Robert Parke, testifying before a congressional subcommittee in 1982, noted that "The development of quantitative social science in this country is inseparable from the development of government statistics, and the two are mutually dependent now." U.S. Congress, House, Committee on Post Office and Civil Service, *Impact of Budget Cuts on Federal Statistical Programs, Hearings before a Subcommittee of the House Committee on Post Office and Civil Service,* 97th Cong., 2nd sess., 1982. This interdependency has influenced which data are collected and how they are created, described, distributed, preserved, and used.

[13] See Dennis M. Jennings and others, "Computer Networking for Scientists," *Science* 231, no.28:943-50 (Feb. 1986).

[14] For a more extensive description of the role of RDBMS in complex statistical data, see Martin David, "Managing Panel Data for Scientific Analysis: The Role of Relational Database Management Systems," paper delivered at the International Symposium on Panel Surveys, Washington, D.C., Nov. 20, 1986.

[15] At our workshops we have trained more than 125 nonprogrammer researchers and graduate students to use the Ingres RDBMS.

[16] This 2 percent sample will be available for the microcomputer in July 1988.

[17] Ekkehard Mochmann, *The ZAR* (Cologne: Zentralarchiv für Empirische Sozialforschung, Univ. of Cologne, 1981.)

[18] Josephina J. Card, "Topically Focussed Archives: A New Pardigm for the Codification of Social Science Research," *IASSIST Quarterly* 10:35-43 (1986).

[19] The bibliographies are not linked to the attributes at this time, nor has a thesaurus been developed; however, the framework for developing the linkage between citations and attribute names and building a thesaurus is in place.

[20] The publications and technical memoranda tables are updated regularly, and printed and circulated twice yearly.

[21] We have used census-record nomenclature. In our restructuring, different naming conventions have been designed to make a table's contents explicit.

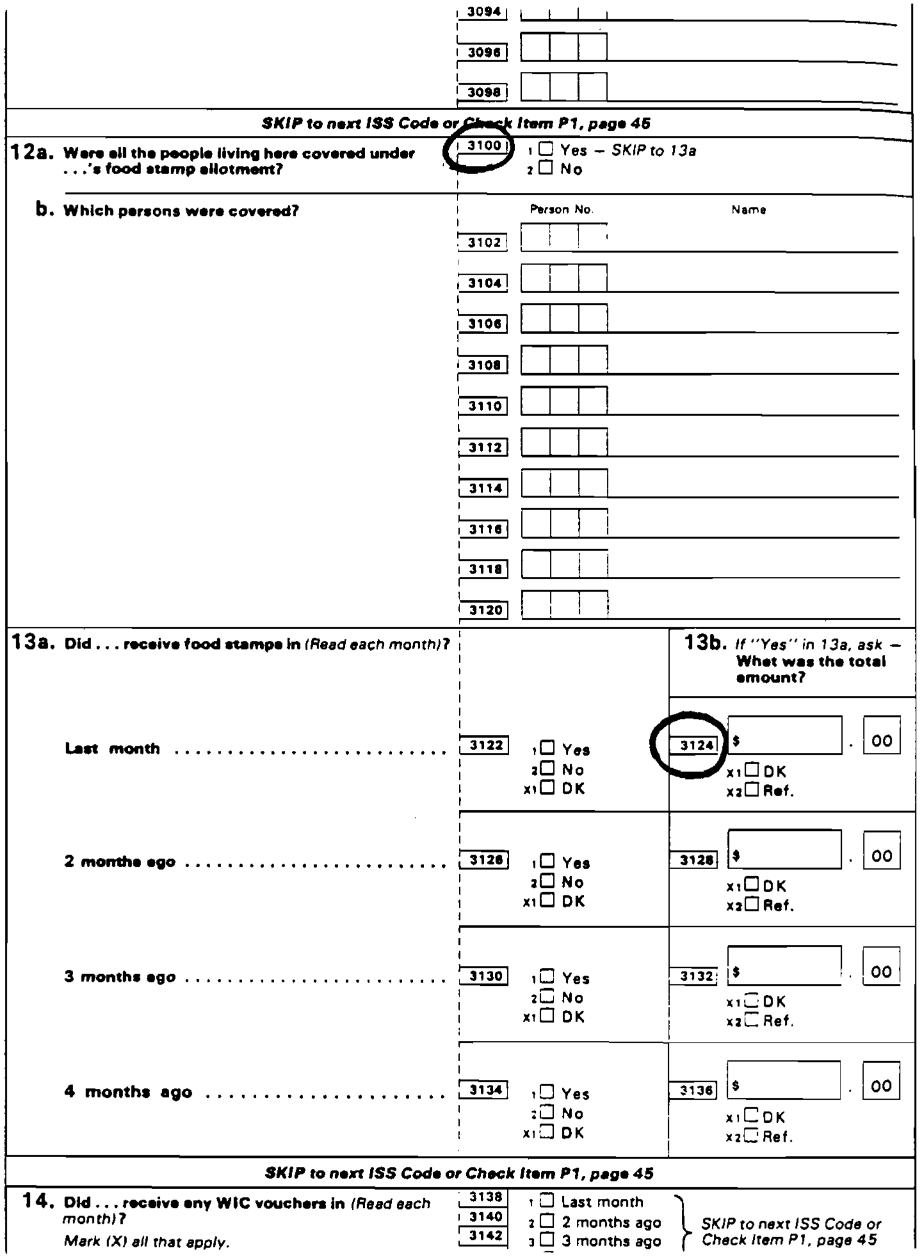
[22] The table name indicates that the item is associated with income recipiency. The only piece of information the user needs to know in advance is the table name *varlist,* available in *SIPPHELP.* Names of the columns can be obtained with the online help inside the database, which provides a description of the table and its contents (see the appendix for an example of a table help).

[23] Research along these lines is already in progress. Zellig Harris, Naomi Sager, and others at the Courant Institute of New York University are applying structural linguistics, information, and computer science to mapping the connections between language and information as applied to the survey instrument. See Naomi Sager and others, "Information Structure in Survey Instruments," *Proceedings of the Survey Research Methods Section, American Statistical Association, 1987* (Washington, D.C.: American Statistical Assn., forthcoming). During the last several years, the National Science Foundation has developed a program to integrate the in-

formation, computer, and social sciences. See Murray Aborn, "Applications of Information Science to Social Measurement," *IASSIST Quarterly* 8:2-17 (Winter 1984).

APPENDIX A. THE VARLIST *HELP* INSIDE THE DATABASE

```
                         VARLIST TABLE

     The varlist table is a concordance of the Bureau of the Census data
dictionary variable names and data base attribute names for the "core"
items (replicates) in the 1984 SIPP.   The table contains the following
attributes: data dictionary name, attribute name, range of the attribute,
location of the attribute in a table, and the waves for which the attribute
is relevant.

     Note that some cases names are repeated because (1) an attribute
appears in more than one table or (2) the Bureau of the Census
variable name underwent a name change (its name is however standardized
in the data base).  For instructions on how to use the varlist table,
see the memo LOCATING_VARIABLES.TXT.


help varlist


Name:                     varlist
Owner:                    sippmanager
Location:                 db_ingres
Type:                     user table
Row width:                47
Number of rows:           1314
Storage structure:        heap
Number of pages:          33
Overflow data pages:      32
Journaling:               disabled
Global Cache:             enabled
Optimizer statistics:     none
Column information:
                                                  key
  column name     type       length   sequence
  bureau            c           10         1
  attname           c           10
  attrange          c            8
  tableroot         c           10
  waves             c            9

Secondary indices:        none
```



3094
3096
3098

SKIP to next ISS Code or Check Item P1, page 45

12a. Were all the people living here covered under . . .'s food stamp allotment? 3100 1 □ Yes – *SKIP to 13a* 2 □ No

b. Which persons were covered? Person No. Name

3102
3104
3106
3108
3110
3112
3114
3116
3118
3120

13a. Did . . . receive food stamps in *(Read each month)?*		**13b.** *If "Yes" in 13a, ask –* **What was the total amount?**	
Last month	3122 1 □ Yes 2 □ No x1 □ DK	3124	$. 00 x1 □ DK x2 □ Ref.
2 months ago	3126 1 □ Yes 2 □ No x1 □ DK	3128	$. 00 x1 □ DK x2 □ Ref.
3 months ago	3130 1 □ Yes 2 □ No x1 □ DK	3132	$. 00 x1 □ DK x2 □ Ref.
4 months ago	3134 1 □ Yes 2 □ No x1 □ DK	3136	$. 00 x1 □ DK x2 □ Ref.

SKIP to next ISS Code or Check Item P1, page 45

14. Did . . . receive any WIC vouchers in *(Read each month)?*
Mark (X) all that apply.

3138 1 □ Last month
3140 2 □ 2 months ago
3142 3 □ 3 months ago

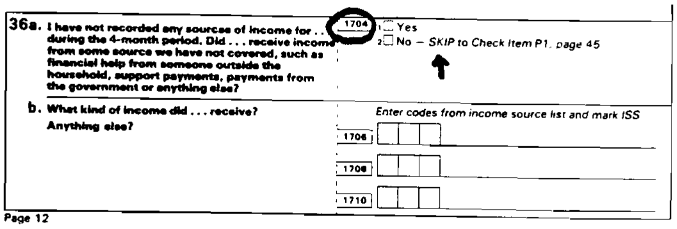
} *SKIP to next ISS Code or Check Item P1, page 45*

Survey of Income and Program Participation
Wave 2 Questionnaire (p, 24)
(Examples 1, 2, 3, 5)

ATTNAME

RESULT

WAVE 2



36a. **I have not recorded any sources of income for . . . during the 4-month period. Did . . . receive income from some source we have not covered, such as financial help from someone outside the household, support payments, payments from the government or anything else?**

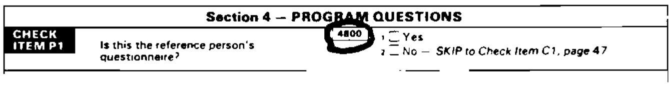
1704 1□ Yes
2□ No – *SKIP to Check Item P1, page 45*

b. **What kind of income did . . . receive?**
Anything else?

Enter codes from income source list and mark ISS

1706 □□□
1708 □□□
1710 □□□

Page 12

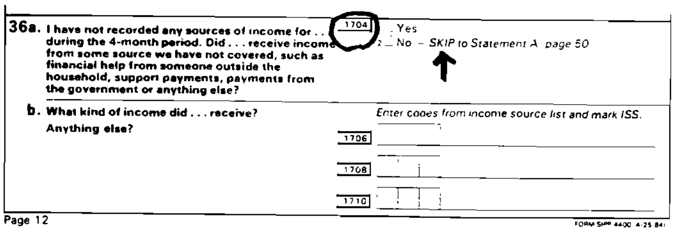


Section 4 – PROGRAM QUESTIONS

CHECK ITEM P1 Is this the reference person's questionneire?

4800 1□ Yes
2□ No – *SKIP to Check Item C1, page 47*

WAVE 4



36a. **I have not recorded any sources of income for . . . during the 4-month period. Did . . . receive income from some source we have not covered, such as financial help from someone outside the household, support payments, payments from the government or anything else?**

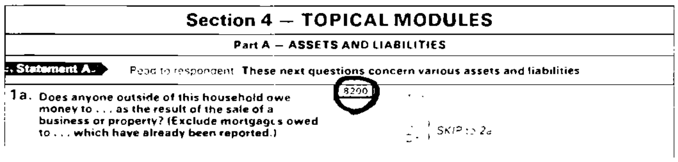
1704 □ Yes
2□ No – *SKIP to Statement A page 50*

b. **What kind of income did . . . receive?**
Anything else?

Enter codes from income source list and mark ISS.

1706
1708
1710

Page 12 FORM SIPP 4400 4-25-84



Section 4 – TOPICAL MODULES

Part A – ASSETS AND LIABILITIES

Statement A. Read to respondent **These next questions concern various assets and liabilities**

1a. **Does anyone outside of this household owe money to . . . as the result of the sale of a business or property? (Exclude mortgages owed to . . . which have already been reported.)**

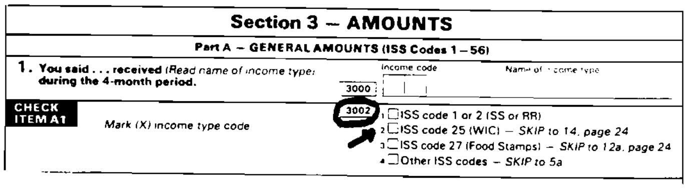
8200

SKIP to 2a

Survey of Income and Program Participation
Waves 2 and 4 Questionnaires
(Example 4)

ATTNAME

RESULT



Section 3 – AMOUNTS

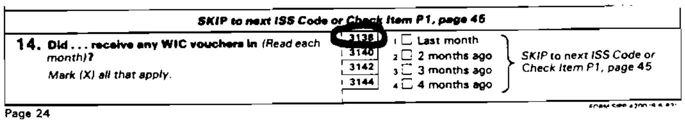
Part A – GENERAL AMOUNTS (ISS Codes 1 – 56)

1. **You said . . . received** *(Read name of income type)* **during the 4-month period.**

Income code Name of income type
3000

CHECK ITEM A1 *Mark (X) income type code*

3002 1□ ISS code 1 or 2 (SS or RR)
2□ ISS code 25 (WIC) – *SKIP to 14, page 24*
3□ ISS code 27 (Food Stamps) – *SKIP to 12a, page 24*
4□ Other ISS codes – *SKIP to 5a*



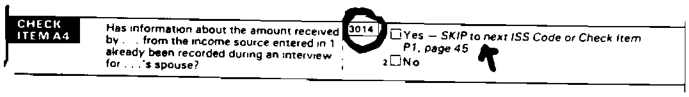
SKIP to next ISS Code or Check Item P1, page 45

14. **Did . . . receive any WIC vouchers in** *(Read each month)***?**
Mark (X) all that apply.

3138 1□ Last month
3140 2□ 2 months ago
3142 3□ 3 months ago
3144 4□ 4 months ago

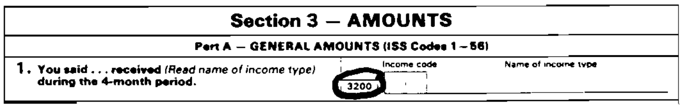
SKIP to next ISS Code or Check Item P1, page 45

Page 24



CHECK ITEM A4 Has information about the amount received by . . . from the income source entered in 1 already been recorded during an interview for . . .'s spouse?

3014 □ Yes – *SKIP to next ISS Code or Check Item P1, page 45*
2□ No



Section 3 – AMOUNTS

Part A – GENERAL AMOUNTS (ISS Codes 1 – 56)

1. **You said . . . received** *(Read name of income type)* **during the 4-month period.**

Income code Name of income type
3200

Survey of Income and Program Participation
Wave 2 Questionnaire
(Example 4)