RESEARCH REPORT SERIES
(Survey Methodology #2010-10)

Using Administrative Record Data to Evaluate the Quality of Survey Estimates

Jeffrey C. Moore
Kent H. Marquis

Statistical Research Division
U.S. Census Bureau
Washington, D.C. 20233

Originally published in the Proceedings of the Statistics Canada Symposium on Statistical Uses of Administrative Data in November of 1987.

Report made available online: June 9, 2010

Proceedings:
Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

USING ADMINISTRATIVE RECORD DATA TO EVALUATE THE QUALITY OF SURVEY ESTIMATES

JEFFREY C. MOORE and KENT H. MARQUIS[1]

ABSTRACT

The Survey of Income and Program Participation (SIPP) is a new Census Bureau panel survey designed to provide data on the economic situation of persons and families in the United States. Each SIPP household is interviewed eight times - every four months - over the two-and-one-half year life of the panel.

The basic datum of SIPP is monthly income, which is reported for each month of the four-month reference period preceding the interview month. The SIPP Record Check Study uses administrative record data to estimate the quality of SIPP estimates for a variety of income sources and transfer programs. The project uses statistical matching techniques to identify SIPP sample persons in four states who are on record as having received payments from any of nine state or Federal programs, and then compares survey-reported dates and amounts of payments with official record values. The paper describes basic considerations in designing the project and presents some early findings.

1. INTRODUCTION

This paper addresses issues concerning the use of records to evaluate the quality of survey estimates and describes a specific application to the Survey of Income and Program Participation (SIPP) in the United States.

Matching administrative records to survey observations on a case-by-case basis, which we call a "record check," provides useful information to survey users and designers. A record check enables the analyst to make a full range of measurement error parameter estimates for evaluation purposes. These estimates, in turn, facilitate two basic kinds of activities:

1. adjusting subject-matter estimates such as means, proportions, correlation coefficients, and multivariate regression coefficients to correct for the measurement errors; and
2. deriving more efficient survey designs that directly address, for example, the trade-offs between measurement quality and costs.

[1] Jeffrey C. Moore and Kent H. Marquis, Center for Survey Methods Research, U.S. Bureau of the Census, Room 2737 FB 3, Washington, DC 20233, U.S.A.

1.1. Basic Terms

Our focus here will be on measurement or response errors, although the record check method can be extended to evaluate other nonsampling and sampling errors also. This is not a technical exposition, but we do need to define some of our basic terms first. We assume that the survey observation from sample element i can be expressed as the sum of the true value and an error, e: $Survey_i = True_i + e_i$.

The average bias in a set of N survey observations, which we call the response bias or survey bias, is $\bar{e} = \Sigma_i / N$ and the response error variance is just Var e.

Similarly the measurement model for the administrative record observation is: $Record_i = True_i + u_i$, so that record bias is $\bar{u}$ and record error variance is Var u.

1.2. Comparison of Evaluation Approaches

The capabilities of the record check approach can be contrasted to other methods of evaluation such as reinterviews and experiments. Reinterviews and other repeated measures designs aim at estimating a very limited set of measurement error parameters, usually something called the simple response variance or the response error variance. These approaches implicitly make strong assumptions about true change over time and about either the true score or bias parameter (Marquis (1986)).

One frequently attempted remedy is to create a criterion measurement as part of the reinterview program, for example by reconciling discrepant answers with a knowledgeable respondent or by asking much more detailed and specific questions during the reinterview. But the validity of these criterion measures is suspect. Both Bailar (1968) and Koons (1973) have shown, for example, that reconciled reinterview responses are biased. And while detailed, specific questioning is often preferred to a more global approach, there is no independent evidence that it reduces measurement biases to zero - or at all. Record checks potentially provide higher quality criterion information requiring much weaker (and perhaps more realistic) assumptions for purposes of estimating survey data quality.

A different method of evaluating aspects of surveys is the experiment, such as a fully-crossed factorial design or an interpenetrated design for assigning interviewers. Analysts compare experimental groups with respect to statistics such as subject-matter means or proportions and draw conclusions about which treatment produces more or less reporting of the subject-matter of interest. What is controversial, however, is determining which treatment is "better" in a measurement sense, a difficulty that is much reduced when criterion data are available, such as administrative records.

Without criterion data, it is often necessary for the analyst to resort to strong assumptions about measurement errors such as:

- more reporting is better reporting;
- forgetting of meaningful material increases with the passage of time;
- unbounded interviews contain overreports, bounded interviews don't;
- reporting performance decays with length of interview or time-in-sample;
- people tend to be lazy and devious - they will lie to avoid being asked a detailed set of questions; and
- self reports are better than proxy reports.

Indeed, these assumptions have become part of the folklore of survey design in the western world. And yet, it is difficult to find any support for any of these assumptions

from appropriately designed record checks. Experiments and related arrangements are excellent approaches to pinpointing the sources of variation, and to untangling estimation problems of collinearity, but are often unnecessary and seldom sufficient for evaluating an existing measurement process.

In sum, these other evaluation approaches are forced to make strong assumptions about: (1) the independence of the original and evaluation measures when they are clearly dependent; (2) the relationship of the original measure to a criterion when no objective, external link exists; and/or (3) cognitive processes not supported by research.

Record checks also employ assumptions in evaluating measurements. For example, the usual way of estimating the response bias is to assume no record bias ($\bar{u} = 0$) and take the average of the differences between the matched survey and record observed values: Estimated Survey Bias $= \Sigma(S_i - R_i) / N$. While one cannot directly support the no record bias assumption, one can conduct meaningful sensitivity tests of the effects of possible violations of the assumption on evaluation conclusions. (At a later date the SIPP Record Check Study will employ these tests and other analyses to examine errors in the records.)

1.3. Issues in Designing Record Checks

Several issues merit consideration in designing a record check to evaluate survey measurement. We comment on some of the main issues here: incomplete observation designs, matching errors, record errors, true score differences, and absence of repeated measures or experimental design features.

1.3.1. Incomplete Observation Designs

Past record checks have often used one-directional or partial designs for data collection, such as when we survey people about owning library cards and check the records for those who claim to have one, or when we sample from a list of people with a diagnosed chronic disease and survey them to see if they report it in a survey questionnaire. Because these partial designs do not observe the full range of response errors in the correct proportions, they yield biased estimates of such classical measurement error parameters as the response bias and the response error variance. One-directional designs can fail to detect some or all of the true survey bias, can cause the analyst to interpret up to one-half of the response error variance as response bias, and can predetermine the sign of the estimated response bias if the measured variable is binary (Marquis (1978)). Full designs are a necessary (albeit not sufficient) condition for obtaining unbiased estimates of the desired response errors.

1.3.2. Matching Errors

The essence of the record check is a one-to-one matching of survey and record observations. This is difficult to do correctly, and matching errors (false matches, false nonmatches) will potentially bias the measurement error estimates of interest. Neter et al. (1965) show that when there are no unmatched cases, the mismatches will bias the estimates of response error variance upward. In terms of the reliability of a dichotomous measure (which is a function of the response error variance), the estimate will be attenuated by exactly the match error rate (Marquis et al. (1986)). It is therefore desirable to keep match errors to a minimum and to know something about the errors that remain.

1.3.3. Administrative Record Errors

As noted earlier, one usually has confidence that the records in a record check study are very good measures of the trait of interest. If the implied assumptions about record measurement bias and record measurement error variance are violated, this can cause the response error estimates to be biased away from zero. For example, bias in the record observations can appear as bias in the survey observations but with the opposite sign. Feather (1972) describes this effect in a record check of physician visits in Saskatchewan, in which an apparently large survey overreporting rate was due to the record's recording a complete treatment procedure rather than the individual visits for the diagnosis. Similarly, the presence of measurement error variance in the record can cause inflated estimates of response error variance in the survey (Marquis (1978)).

1.3.4. True Score Differences

Problems arise when the survey and record systems use different definitions. This is often the case in "aggregate comparisons" of population parameter estimates made separately by each source. A common difference is in the scope of the populations covered, such as when the survey frame is limited to the civilian, noninstitutionalized population and the record includes everybody. Case-by-case matching can minimize the threats posed by differential coverage, but even estimates derived from these studies can still be plagued by differences in the concepts or the attributes of the concept. For example, our administrative records often contain the date a check was written for a transfer payment and SIPP survey respondents tell us when they received the payment. Such differences can threaten our time-related estimates of such things as telescoping response errors.

1.3.5. Absence of Experiments and Reinterviews

Evaluation record checks can detect errors but are not good at evaluating the remedies for the errors. To know how well a different survey design might perform, one must usually either test the alternative design options or arrange to estimate parameters of an underlying model from which survey designs can be derived (e.g., a model of forgetting effects). For example, an evaluation record check design can estimate and compare response errors for self and proxy respondents. Without heroic assumptions it cannot, however, suggest how the measurement error parameters would change if the survey's respondent rule were changed (say, to allow only self-response).

Similarly, a record check without a reinterview or another set of independent measures is limited in the number of basic error parameters it can estimate. For example, our initial definitions mentioned three parameters: true score, survey error, and record error. Without a reinterview (or other independent measure) there are only two measures with which to estimate the three unknowns. An additional measure such as a reinterview can help identify the estimates of the parameters in the model.

2. CHARACTERISTICS OF SIPP

Here we briefly describe features of SIPP as a prelude to discussing the record check evaluation design.

2.1. Overview of SIPP Contents

The purpose of SIPP is to provide improved information on the economic situation of people and households in the United States. It collects comprehensive longitudinal data on cash and noncash income, eligibility for and participation in Government transfer programs, assets and liabilities, labor force participation, and a host of related topics. SIPP data assist the evaluation of the cost and effectiveness of current Federal Government programs, of the potential impacts of proposed program changes, and of the actual impacts of changes when implemented. In general, the Census Bureau and other Government agencies which have fostered and supported the development of SIPP expect it to be an invaluable tool for domestic policy planning (Nelson et al. (1985)).

Core SIPP questions - repeated in each wave of interviewing - cover labor force participation and amounts and types of income received, including transfer payments and noncash benefits from various programs for each month of the reference period. The core questions cover nearly 50 sources of income, including Government transfer payments from retirement, disability and unemployment benefits, and welfare programs such as Aid to Families with Dependent Children. Information is also gathered on noncash programs such as food stamps, Medicare, and Medicaid; private transfers such as pensions from employers, alimony, and child support; ownership of assets that produce income, such as interest, dividends, rent, and royalties; and on miscellaneous sources of income, such as estates.

2.2. SIPP Data Collection Design

SIPP started in October 1983 with a sample of approximately 25,000 designated housing units (the "1984 Panel") selected to represent the noninstitutional population of the United States. In February 1985 a new and slightly smaller panel was introduced. Additional panels are to be introduced each February throughout the life of the survey. Due to budget reductions, the sample size for new panels is currently about 15,000 households.

Each sample household is interviewed by personal visit once every four months for 2-½ years, resulting in a total of eight interviews per household. The reference period for each interview is the four months preceding the interview month. At each visit to the household, each person fifteen years of age or older is asked to provide information about himself/herself. Proxy reporting is permitted for household members not available at the time of the visit. Information concerning proxy response situations is recorded and is available for analytical purposes.

To facilitate field operations, each sample panel is divided into four subsamples ("rotation groups") of approximately equal size, one of which is interviewed each month. Thus, one "wave" or cycle of interviewing is conducted over a period of four months for each panel. This design produces steady field and processing workloads, but it also means that each rotation group uses a different four month reference period.

Beginning with the second wave of interviewing in the 1984 panel, SIPP includes reinterviews with a small sample of households about a subset of items (including program participation). These data are used primarily to check for interviewer falsifications, but may also be of some use in estimating response inconsistencies.

3. RECORD CHECK DESIGN

The purpose of the record check is to provide an evaluation of some of the data gathered in SIPP. We highlight important features of the design of the record check next, covering the samples, the administrative records, the matching approach, and the analysis.

3.1. Record Check Samples

The SIPP record check uses a "full" rather than a one-directional design; that is, the records we have allow us to validate all observed values in the survey. Design options we did not choose include: (1) checking records only for people who claimed to be participating in a program, or (2) drawing a sample of known recipients and interviewing them to determine how truthfully they report. Both of the latter designs are incomplete and will result in biased estimates of the response error parameters.

The Record Check Study restricts attention to a subset of available SIPP data from the 1984 Panel. First, the sample of people is restricted to households in four target states: Florida, New York, Pennsylvania, and Wisconsin. In the 1984 Panel this translates to approximately 5,000 households. Second, the study's sample of calendar time periods includes only the first two waves of the 1984 Panel. Figure 1 illustrates the wave, rotation group, interview month, and reference period structure for the target survey data.

Figure 1:
Survey Structure for Data Included in the
SIPP Record Check Study

Wave	Rotation Group	Interview Month	Reference Period Months										
			Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr
1	1	Oct 83	X	X	X	X							
	2	Nov 83		X	X	X	X						
	3	Dec 83			X	X	X	X					
	4	Jan 84				X	X	X	X				
2	1	Feb 84					X	X	X	X			
	2	Mar 84						X	X	X	X		
	3	Apr 84							X	X	X	X	
3	4	May 84								X	X	X	X

Third, the SIPP Record Check Study focuses on the quality of recipiency and amount reporting for selected Government transfer programs. We will compare survey reports and administrative records for five Federally administered programs (Federal Civil Service Retirement, Pell Grants, Social Security (OASDI), Supplemental Security Income (SSI), and Veterans' Compensation and Pensions) and four state-administered programs (Aid to Families with Dependent Children (AFDC), food stamps, unemployment compensation, and worker's compensation).

We limited the study to four states - Florida, New York, Pennsylvania, and Wisconsin - in order to keep the study to manageable proportions. Major criteria used to select these states were: (1) the presence of a computerized, accessible, and complete record system for all target programs; (2) a large SIPP sample; (3) reasonable geographic diversity; and (4) a willingness to share individual-level data for purposes of this research. Thus, the

states were selected purposively; no attempt was made to sample states to be representative of the Nation.

We requested from each participating state agency identifying and receipt information for all persons who received income from the target program at any time from May 1983 through June 1984. The identical request was made of the participating Federal agencies, with the exception that only recipients residing in one of the four selected states were to be included in the data extract.

We obtained these administrative records with the understanding that they would be accorded the same confidentiality protection as data gathered by the Census Bureau under Title 13 of the U.S. Code. Thus, the records may be used only by sworn Census Bureau employees engaged in and for the purposes of the Record Check Study. Except in the form of nonindividually-identifiable statistical summary data, the records may not be released or disclosed to any others for any purpose.

Some agencies elected to follow a two-step procedure, initially providing only recipient identifying data, with no (or only minimal) data on program benefit receipt history. Following the matching of the recipient and SIPP files the project will send back to the agency a list of case identifiers for matched persons (plus a sufficient number of nonmatched case identifiers to assure the confidentiality of the SIPP sample). The agency will extract and return to the Census Bureau payment history data for these cases.

As noted earlier, errors in the records can cause problems for record check evaluation studies. Although several of the administrative record files obtained for this project contain very minor deficiencies (for example: not listing a middle initial; no sex designation; age, rather than date of birth; etc.), only three appear at all likely to pose major analytical problems. Two are known to be incomplete in their coverage of recipients: the New York worker's compensation file, and the Veterans' Compensation and Pensions file covering all four states. The former excludes an unknown number of cases which were "closed" (i.e., cases which had already been adjudicated and for which payments by a private insurance carrier had already begun) at the time the data base was created several years ago. The latter excludes the approximately one percent of all recipients whose benefits were sent to a financial or other institution. There are no known coverage problems with any other files. The third problematic file has complete coverage but lacks recipient address information, which can be very useful for matching.

An unavoidable problem which afflicts all of the administrative files to some extent is the discrepancy between payout date and receipt of payment; obviously, the SIPP respondent reports the latter and has no knowledge of the former, and the reverse is true for the program records. Where the payout date is close to the end of a month it may be difficult to distinguish a forward telescoping error from a legitimate difference between month of payment and month of receipt. Where there are definitional discrepancies, such as this payment date issue, our analyses will attempt to model them explicitly.

4. MATCHING

4.1. Introduction

The quality of matching has an important effect on some of the most critical response error estimates such as the response error variance. Ideally, variables used to match survey and record observations are measured without error and are able to identify an individual uniquely. The ideal, of course, is never realized.

However, the variables we have available to match surveys and records should go a long way toward minimizing the match errors. Some, such as social security number (SSN),

uniquely identify an individual even if other information such as address is outdated, garbled, or obliterated or missing. For purposes not directly related to this study (although certainly of benefit to it), the Census Bureau has taken special measures to ensure that SSN information as reported to the SIPP is complete and valid. For all Wave 1 and 2 sample persons, reported SSN's and reports of not having an SSN were verified and, if necessary, corrected, by the Social Security Administration. Sater (1986) estimates that as a result of this operation the SIPP file contains a valid SSN for about 95 percent of SIPP sample persons who have one.

The wealth of other data - last name, first name, house number, street name, apartment designation, city, zip code, sex, and date of birth - is sufficient for high quality matching even in the absence of a unique identifier such as SSN. In addition, to aid us in evaluating the impact of any remaining match errors, the Census Bureau's matcher produces an ordinal measure of the goodness of the match/nonmatch of each survey observation to its appropriate administrative record counterpart.

4.2. The Census Bureau's Computerized Match Procedures

The Record Check Study uses computerized statistical matching procedures applying the theoretical work of Fellegi and Sunter (1969). These procedures were developed at the Census Bureau, primarily for purposes of census undercount estimation.

Computerized statistical matching is the process of examining two computer files and locating pairs of records - one from each file - that agree (not necessarily exactly) on some combination of variables. The process involves multiple discrete steps, but basically there are four: standardizing the common data fields in the two files which the matcher will examine to determine whether a pair of records is a match or not; sorting the two files into small subsets of records (or "blocks") which constitute a feasible number of pairs to be examined by the matcher; determining and quantifying the usefulness of each data field to be considered in the match for identifying true matched pairs; and implementing the computer algorithms which perform the actual record matching.

4.2.1. Standardization

We will process all data files in the Record Check Study - both the SIPP files and the administrative record files - through an address standardizer which standardizes the format of various components of an address (e.g., street name, type, and direction; city name; state abbreviation; etc.) and parses each component into a fixed data field. Several programs have been developed for this purpose; we currently use the ZIPSTAN standardizer developed at the Census Bureau, but may soon switch to a new generation product developed by our Geography Division.

In addition to the standardization procedures which apply to all data files, many of the files require modifications to individual data fields to ensure a common format across files for matching. Common examples of variables which pose problems of this type are sex (which can be represented by either an alpha ("m" or "f") or a numeric ("1" or "2") code); date of birth (which has many variants - e.g., "mm-dd-yy," or "cc-yy-mm-dd," or the Julian format); and name (which may be a single field or which may have separate fields for each component). Currently we prepare custom-made programs to carry out this type of standardization but a new version of the Census Bureau's Generalized Data Standardizer (GENSTAN) may soon take over this task.

4.2.2. Blocking

Blocking - establishing subsets of records for the matcher to examine in searching for matched pairs of records (e.g., Jaro (1985)) - is a necessary strategy when matching files with large numbers of records. Obviously, the probability of finding all true matches would be highest if, for each record on one file, the entire other file were searched for a match. However, for large files such unrestricted searches for matched records is simply not feasible. Blocking each file into subsets of records makes matching large files feasible, but at the cost of excluding some records from the search, thus increasing the likelihood that some true matches will be missed. Ideal blocking components, therefore, have sufficient variation to ensure the partitioning of the files into many (and therefore smaller) blocks, and are effective match discriminators - that is, nearly always agree in true match record pairs and nearly always disagree in true nonmatch record pairs. (The latter also implies that an ideal blocking component must be largely error-free on both files.)

The first of these criteria - sufficient variation - is easy to achieve; the second is more problematic. The primary blocking strategy for the SIPP Record Check Study employs the first three digits of the United States Postal Service's five-digit zipcode and a four-character SOUNDEX code derived from the sample person's/recipient's last name. The former is a sub-state geographic indicator which generally is recorded quite accurately according to Census Bureau matching experts. The latter is a widely-used algorithm for creating a standard length, standard format code from input character strings of varying lengths. The code is comprised of the first letter of the string (here, the last name), followed by a numeric code which is based on only certain letters in the remainder of the string. The advantage of such encoding for blocking purposes is that it minimizes blocking errors due to misspellings, although it cannot eliminate such errors entirely.

Because the success of the match is so sensitive to the blocking scheme, the study will use at least two and possibly three separate blocking strategies - each employing totally unrelated blocking components - for each pair of files to be matched. This will minimize the likelihood that a true match pair will escape detection as a result of blocking. These subsequent blocking arrangements will not be uniform for all matches (because of variations in the availability of some data fields or because of known problems with quality) but are likely to include some combination of sex, month of birth, day of birth, SOUNDEX code for city or street name, or partial SSN.

4.2.3. Data field match weights

With some variation, the data fields used in the matching of the SIPP and administrative record files will include house number, street name, apartment number, city, zip code, SSN, sex, date of birth, last name, and first name. Intuitively, these fields are not equivalent when it comes to determining whether a particular pair is a match or not - agreement on sex is not as indicative of a true match as is agreement on SSN, for example. Fellegi and Sunter (1969) include, in their presentation of a general theory of record linkage, discussions of weight calculations reflecting different data fields' differing discriminating powers and how these weights feed into optimal decision rules. The Census Bureau's Record Linkage Research Staff has developed programs using Newton's method for non-linear systems (see Luenberger (1984)) to solve the Fellegi-Sunter equations, and these programs are being used in the SIPP Record Check Study to compute final match weights.

4.2.4. The computer matcher

The Census Bureau is developing a computer matcher (CENMATCH) operating on IBM personal computers, on an IBM 4361 mainframe, and on other hardware, which executes the procedures of Fellegi-Sunter on a user-defined set of data fields on files sorted (blocked) according to user specifications. The user enters the initial match weights for each field, defines the type of agree/disagree comparison for each field (whether the fields must be exactly comparable in order for the matcher to treat them as agreeing, or whether only approximate comparability is necessary), identifies missing value entries and specifies how they are to be treated (included or ignored in the calculation for a composite match weight), and sets the composite weight cutoff values for matched pairs an nonmatched pairs. The user generates the appropriate COBOL program codes to conduct a match according to these specifications through GENLINK, the Census Bureau's Record Linkage Program Generator (LaPlant (1987)).

In simple terms, the matcher: (1) searches each data file for comparable blocks of records - that is, records which agree exactly on the designated blocking components; (2) counts the number of records in found blocks to ensure that neither file's block size exceeds the preset maximum; (3) computes composite weight for all possible pairs of records in the block; (4) assigns each record in the smaller block to a paired record in the larger block according to a formula which maximizes the total composite weight for all pairs in the block; (5) applies the Fellegi-Sunter decision procedure to determine whether a pair is a match, a nonmatch, or requires further review; and (6) produces a "pointer" file map to the skipped records (i.e., records in a block on one file that is not matched with a corresponding block in the other file) and the paired records (matched /review /unmatched) in each file.

5. ANALYSIS

Our goals for the record check study are to estimate selected measurement error parameters for our samples of people, content, and times, and to assess how these errors relate both to each other and to variables that reflect survey design features. Our general plan is to use the matched data to estimate for each dichotomous participation variable:

- the response bias (using the survey-minus-record difference score);
- predictors of the response bias (using logistic or probit regression techniques or possibly LISREL techniques based upon matrices containing polyserial and tetrachoric coefficients of association (Jöreskog and Sörbom (1984));
- the response error variance (e.g., derived from regression residuals);
- the conditions or groups associated with very large and very small response error variances; and
- the kinds and amounts of confusion among transfer programs that contribute to the response errors (using covariance structure analysis procedures such as LISREL).

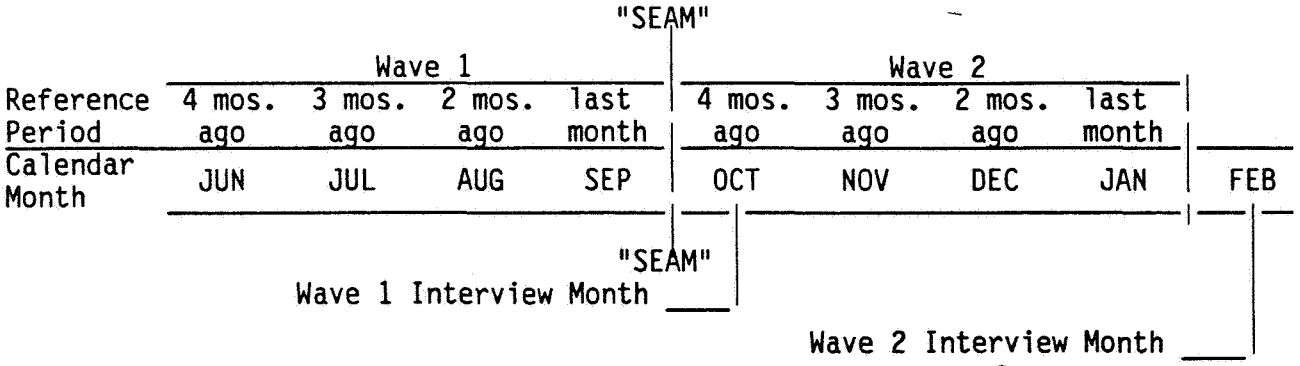
We plan to estimate the same parameters for reports of the amounts of money received from each transfer program but we have not yet selected our basic estimation approach.

The measurement error issues to be addressed fall into one of two categories: issues which apply to all time periods and issues that require comparing errors across time periods. In the former category are estimates of the amounts of response errors for self

and proxy respondents or contributed by interviewers. In the latter category are the errors arising from panel surveys with familiar labels such as telescoping, time-in-sample bias, memory decay, rotation group bias, etc. - those implying that measurement errors will differ across time periods when everything else is held constant. To this list we add what Hill (1987) has referred to as the "seam" bias in longitudinal surveys, which we discuss below.

To appreciate the applied questions we wish to address about the different time periods, consider Figure 2, which presents the interview and reference month calendar for one rotation group of SIPP respondents.

Figure 2:
SIPP Survey Time Periods for Rotation Group 1



The figure shows two interviews. The first takes place in early October and asks about what happened in September (last month), August (two months ago), July (three months ago), and June (four months ago). Similarly, the second interview taken four months later, asks about January, December, November, and October. We refer to the transition between September and October as the "seam" because it is between the reference periods covered by the two interviews.

To investigate the internal telescoping hypothesis (which asserts that events are not forgotten, just remembered as having happened closer to the present time), we will be testing whether the response bias for the early months of the reference period (June and July in Wave 1 and October and November in Wave 2) is negative and the response bias for later months (August and September or December and January) is positive, and that the two biases sum to zero.

We plan to test the bounded interview hypothesis, which says that events from the remote past are reported as happening within an unbounded reference period (June through September), but that this will not happen in reference periods bounded by a previous interview (here, October through January).

To examine the hypothesis about memory decay (that the probability of forgetting an event increases with the passage of time), we will test whether the response bias is more negative for the early months of each reference period than for later months.

The time-in-sample and rotation group hypotheses suggest that response errors will be greater in the second interview than the first, after correcting for any seasonal effects. We plan to examine this and, if we find it to be true, test some of the ideas in the literature about why it may be true. Are the sample elements that survive from the first to the second interview different, as Stasny and Fienberg (1985) suggest, or does the quality of the survivors' reporting deteriorate as the Neter and Waksberg (1966) conditioning hypothesis might predict?

We don't know yet the extent to which SIPP is experiencing these more traditional problems of longitudinal surveys. One problem for which there is evidence, however, concerns the estimation of month-to-month changes in program participation (Burkhead and Coder (1985)). Specifically, more changes in program participation take place at the "seam" between interviews (between September and October in Figure 2) than between the months covered by any one interview (e.g., between June and July or July and August or August and September). The Census Bureau has not published monthly program participation transition estimates from SIPP yet because the estimates show a pattern that appears to be affected heavily by measurement error. Moore and Kasprzyk (1984) and Hill (1987) have speculated about what kinds of response, nonresponse, or procedural errors might be producing the pattern and which set of transition estimates is more accurate. By addressing the problem with administrative data, we hope to come much closer to a definitive explanation about the role of response and nonresponse errors in producing the observed pattern.

Related, possibly, to the seam bias issue is the better-understood phenomenon that measurement error variance tends to inflate estimates of gross change or underestimate stability. Recent literature (e.g., Fuller (1986)) suggests several possible approaches to the problem. We plan to begin the empirical exploration of the measurement error effects on the transition estimates to learn whether, for example, we can base corrections for the response errors on estimates from reinterviews.

Finally, we have hinted previously at the problems that may arise in getting unbiased estimates of the errors if the records also contain errors. We plan, with the use of reinterview measures (that identify the estimate of Var e) to estimate the record error variance (Var u). However, we have no plans to relax the assumption that the records are unbiased.

6. PRELIMINARY FINDINGS

To illustrate our approach, let us look at the "seam" issue with some test data we are using to get experience with data processing procedures. Recall that the seam problem is that monthly survey reports about program participation status produce more frequent status changes between months covered by separate interviews than between other months (covered by the same interview).

Some initial questions about the survey data that administrative record information would help answer include these:

1. Are there too many transitions reported at the seam?
2. Are there too few transitions reported for other months?
3. Do the different sources report the same number of changes over the whole time period but distribute them differently?

Next we will show what we call "aggregate comparison" data relevant to these questions noting, however, that the data come from a convenience sample and do not necessarily represent any population of interest. Also note that there are a small number of cases by Government survey standards. For these reasons we will stick to descriptive statistics.

Aggregate comparisons do not involve case-by-case matching of survey and record data; in this example, however, we use exactly the same sample of 1,536 people for both the survey and record values. This eliminates differences in coverage definitions that often plague this method.

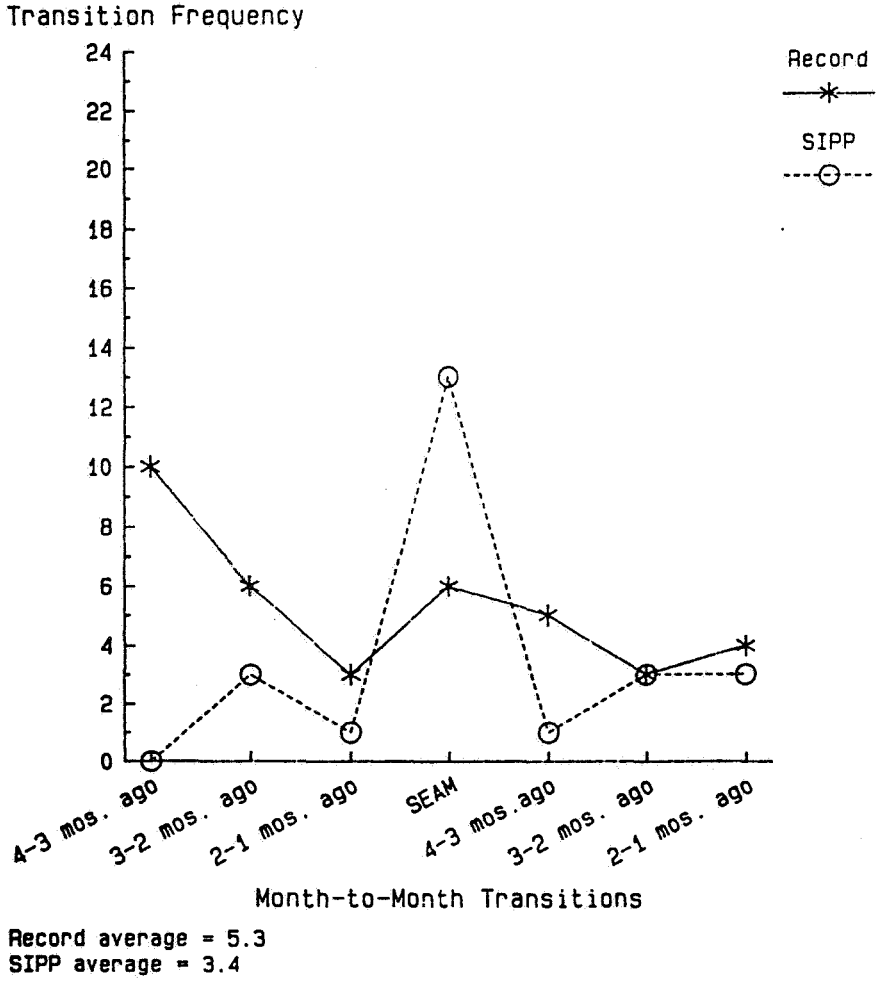
Assuming that the record data are correct, the AFDC graph (Figure 3) suggests:

1. Too many transitions are inferred at the seam from the survey;
2. Too few transitions are inferred for the other months; and
3. Too few transitions overall are reported in the survey, a net underreporting problem as well as a time-placement problem.

Turning to the Food Stamp graph (Figure 4), we see similar but not identical trends:

1. There are still too many transitions inferred at the seam;
2. But whatever underreporting bias there is in the other months does not seem severe; and
3. Both survey and record contain about the same number of total transitions, suggesting just a time-placement problem and not a net bias phenomenon.

FIGURE 3:
AFDC Transitions as Reported in SIPP and in Records

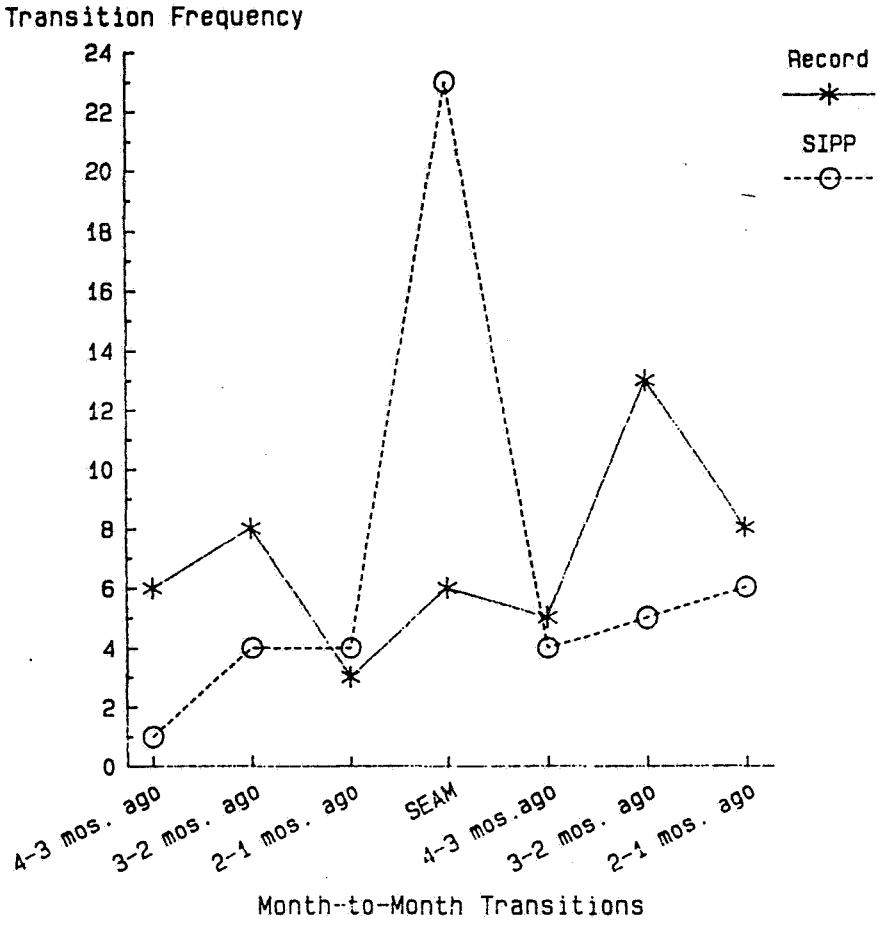


Record average = 5.3
SIPP average = 3.4

There are many more tests to be done and many hypotheses to explore before we start to draw conclusions about the nature of the measurement errors and their probable

causes. We feel that the administrative record data will allow us to make important advances toward understanding the sizes and forms of these survey errors and perhaps suggest their causes.

FIGURE 4:
Food Stamps Transitions as Reported in SIPP and in Records



Record average = 7.0
SIPP average = 6.7

ACKNOWLEDGEMENTS

The SIPP Record Check Study has already benefited greatly from the efforts of many people. While we cannot list here all who deserve recognition, we do gratefully acknowledge the particular contributions of: Jeannette Robinson, for preparing the multitude of administrative record files for matching; Bill LaPlant, for sharing his considerable expertise regarding the Census Bureau matcher and attendant software; Chris Dyke, for his tireless efforts to assist in making the matcher work on a new computer system; and Dan Kasprzyk, for his constant and patient support of this entire endeavor.

REFERENCES

Bailar, B. (1968). "Recent Research in Reinterview Procedures," Journal of the American Statistical Association, Vol. 63, 41-63.

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, Washington, DC.

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

Feather, J. (1972). *A Response Record Discrepancy Study.* University of Saskatchewan, Saskatoon.

Fellegi, I., and Sunter, A. (1969). "A Theory for Record Linkage," *Journal of the American Statistical Association*, Vol. 64, 1183-1210.

Fuller, W., and Tin, C.C. (1986). "Response Error Models for Changes in Multinomial Variables," *Proceedings of the Second Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 425-441.

Hill, D. (1987). "Response Errors Around the Seam: Analysis of Change in a Panel with Overlapping Reference Periods." Presented at the Annual Meetings of the American Statistical Association, San Francisco, CA, August 13.

Jaro, M. (1985). "Current Record Linkage Research." Presentation to the Census Advisory Committee of the American Statistical Association, U.S. Bureau of the Census, April 25, 1985.

Jöreskog, K., and Sörbom, D. (1984). *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood, Instrumental Variables, and Least Squares Methods*, Mooresville, Indiana: Scientific Software, Inc.

Koons, D. (1973). "Quality Control and Measurement of Nonsampling Error in the Health Interview Survey," *Vital and Health Statistics*, Series 2, No. 54, U.S. Public Health Service, Washington, DC.

LaPlant, W. (1987). "Maintenance Manual for the Generalized Record Linkage Program Generator (GENLINK) SRD Program Generator System." Statistical Research Division Internal Working Paper, Washington, DC: U.S. Bureau of the Census.

Luenberger, D. (1984). *Linear and Nonlinear Programming.* Reading, MA: Addison Wesley.

Marquis, K. (1986). "Discussion of 'Correlates of Reinterview Inconsistency in the Current Population Survey'." *Proceedings of the Second Annual Research Conference*, U.S. U.S. Bureau of the Census, Washington, DC, 235-240.

Marquis, K. (1978). *Record Check Validity of Survey Responses: A Reassessment of Bias in Reports of Hospitalizations.* The Rand Corporation, Santa Monica, CA. R-2319-HEW.

Marquis, K., Marquis, S., and Polich, M. (1986). "Response Bias and Reliability in Sensitive Topic Surveys," *Journal of the American Statistical Association*, Vol. 381-389.

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

Nelson, D., Mcmillen, D., and Kasprzyk, D. (1985). "An Overview of the Survey of Income and Program Participation, Update 1." *SIPP Working Paper Series*, No. 8401, Washington, DC: U.S. Bureau of the Census.

Neter, J., Maynes, S., and Ramanathan, R. (1965). "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, Vol. 60, 1005-1027.

Neter, J., and Waksberg, J. (1966). "A Study of Response Errors in Expenditures Data from Household Interviews," *Journal of the American Statistical Association*, Vol. 59, 18-55.

Sater, D. (1986). "SSN Response Rates and Results of SSN Validation/Improvement Operation." U.S. Bureau of the Census memorandum for R. Herriot, March 11, 1986.

Stasny, E., and Fienberg S. (1985). "Some Stochastic Models for Estimating Gross Flows in the Presence of Nonrandom Nonresponse," *Proceedings of the Conference on Gross Flows in the Labor Force Statistics*, Department of Commerce and Department of Labor, Washington, DC, 25-39.