THE SIPP MEASUREMENT QUALITY EXPERIMENT AND BEYOND:
BASIC RESULTS AND IMPLEMENTATION

THE SIPP MEASUREMENT QUALITY EXPERIMENT AND BEYOND: BASIC RESULTS AND IMPLEMENTATION

Kent H. Marquis[1]
U. S. Census Bureau

ABSTRACT

Initially, the Survey of Income and Program Participation (SIPP) learned that it had potential problems due to error in measuring program participation and related variables. Based on cognitive theory and research, new procedures were designed to reduce response errors and were tested using both an experimental design and an administrative record check. A key feature of the new procedures was getting households to use their personal income records. Results indicated that the new procedures had no important effect on under and overreporting participation in the income programs tested. The new procedures did achieve a substantial improvement in the reporting of amounts of income by the second interview in the panel. Experimental group households did use personal income records at astonishingly high rates and record use correlated with quality of reporting income amounts. The paper discusses why record use did not affect reporting of program participation while having important effects on reporting income amounts.

KEYWORDS

Measurement Error, Personal Records, Administrative Records, Cognitive Research, Interviewing procedures, Questionnaires, Quality Control, Personal Income, Program Participation

1.0 INTRODUCTION

Measurement error can be an important source of bias in estimates from surveys and censuses. Early studies of the Survey of Income and Program Participation (SIPP) showed important levels of measurement error. These prompted us to design new procedures to reduce response errors in reports of income and program participation by at least 25%.

We tested the new procedures using an experimental design and administrative record check so we could both estimate response errors and compare them to error levels from interviews conducted by conventional procedures.

The results of the test showed that the new approach was not more effective in reducing the underreporting of program participation or other income sources. However, it was able to reduce error in reported income amounts by the second interview in the panel. The new procedures got people to use their personal records but they cost substantially more to implement and failed to achieve satisfactory response rates.

This paper contains 4 more sections. The next section describes SIPP and discusses the research that led up to the current project. Section 3 covers the design of the new survey procedures intended to reduce response errors and the design of the evaluation test. The results section covers underreporting, overreporting and bias in reporting transitions in program participation and errors in reporting amounts of

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the U.S. Census Bureau.

income. It also compares the experimental and control treatments on costs, response rates and rates of using personal records. In the final section I discuss the results, their likely impact on future SIPP measurement, and offer some thoughts about using personal records in household economic surveys.

2.0 BACKGROUND

The Survey of Income and Program Participation is an important source of information about the economic situation of people and families in the United States. It is a longitudinal household survey, conducted by the Census Bureau, to measure both short and long term levels and changes of income and participation in government transfer programs.

Early research indicated there might be some measurement problems in the survey. Specifically, my colleague, Jeff Moore and I conducted a full-design record check study on the first 2 interviews of the 1984 SIPP panel (Marquis and Moore, 1990). We obtained administrative data for four entire states for up to eight income programs: Social Security income, federal Civil Service retirement income, AFDC, food stamps, veterans benefits, unemployment insurance, workers' compensation, and supplemental security income. We matched the SIPP interview data to the administrative records and calculated the amount of error SIPP respondents made in their survey reports.

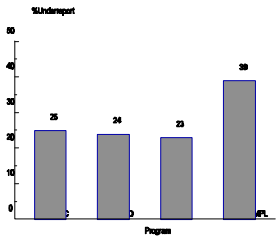

Figure 1: Earlier research on the 1984 panel showed substantial underreporting of program participation in SIPP.

Figure 1 (adapted from Marquis, Moore, and Bogen, 1993) contains results for the underreporting of participation in four government transfer programs. In three of the four programs, around 25% of the true months of program participation were not reported in the SIPP interview[2]. The fourth program, income from the unemployment insurance programs, had an underreport rate of about 40% for participation. These imply much higher rates of response bias than would be expected by looking at the survey methods literature on welfare reporting (e.g., Marquis, Marquis, and Polich, 1986).

There are several different kinds of constructive directions to take once such problems are known to exist. Two of the papers in this session today illustrate the basic approaches. Chris Bollinger and Martin David are concerned with using the response error information to correct estimates in policy models. The approach that my Census Bureau colleagues, Jeff Moore and Karen Bogen, and I took was to design new measurement procedures to reduce the occurrence of the response errors in the first place.

Our design was aided by the earlier record check research which also addressed causes of measurement errors. That research, for example, ruled out memory decay, proxy response, learning to underreport, and variation due to individual interviewers as major causes of error in program participation

[2] The 25% underreporting estimate for AFDC excluded households in the Pennsylvania sample because this group exhibited some unique problems discriminating between AFDC and general assistance programs.

reports. Those analyses did suggest that small amounts of error could be reduced by eliminating occasional cognitive confusion about program names (e.g., social security and supplemental security) and confusion about the recipient of the benefits. More recent cognitive research (Marquis, 1990) suggested that the overly simple strategies that respondents use to reconstruct their past income streams might be at the root of a larger segment of the measurement errors.

3.0 METHODS

This section begins with some guiding assumptions we made about respondents, questionnaires and interviewers, followed by a description of the new interviewing procedures that we created to reduce measurement error. It concludes with some highlights of the design we used to evaluate the new procedures.

3.1 Behavioral Assumptions Underlying the Experimental Procedure.

The experimental interviewing procedures were a radical departure from the techniques used by the conventional SIPP control procedures. To appreciate the differences, I'll begin by considering some basic assumptions that guided their design. These assumptions are based, in large part, on the results of cognitive interviews conducted by the Census Bureau (Marquis, 1990) and by Westat (Cantor, Brandt and Green, 1991).

3.1.1 The Respondent. We assumed that SIPP respondents had good intentions but were often unable to accurately recall all the relevant details of each income stream, such as the gross (vs. net) amount. In many cases, respondents never knew income details for other people in the household[3]. In other cases, details such as the name of the program or the "official" beneficiary were subject to comprehension mistakes. As a result, we assumed that our well-intentioned respondent would often use simplistic reconstruction strategies (based on general and faulty knowledge) as a substitute for detailed, accurate recall or as a substitute for using personal records. Furthermore, respondents who used simple strategies were assumed to be unaware or unwilling to acknowledge that such tactics are prone to error, hence unwilling to change them. Our new procedures were designed to preempt the respondent's use of simple heuristic reconstructions and, instead, to substitute the use of accurate, complete information from personal records.

We assumed that well-intentioned respondents would sometimes volunteer all the information they can think of, even if it were not relevant to the question being asked at the time. Such information could get "lost" by an interviewer who is conscientiously following the questionnaire script. Our procedures were designed to accommodate such volunteered information.

We also assumed that high levels of reporting accuracy would require that respondents use their personal income records and that most respondents would be willing to use their personal records and be willing to save their records for use in future interviews if they were asked. So we instituted a set of procedures that clearly indicated we seriously wanted respondents to use their records to report income.

3.1.2 The Questionnaire. The earlier research revealed some additional minor problems that respondents had in comprehending the questions and instructions. We assumed that most of these

[3] For a recent discussion of strategies that proxy respondents use to report difficult-to-recall information about others, see Schwarz and Wellens (1994).

misunderstandings could be corrected by reorganizing the questionnaire into logical sections and making the objectives of each section clear. Also, we assumed we could minimize the problems that interviewers had with skip (branching) instructions by using formats that did not require the instructions[4].

3.1.3 The Interviewer. We assumed that skilled interviewers could make any set of procedures "work" if they were taught the required skills and understood the priorities. But SIPP standard practice is to reward interviewers for high response rates and high levels of efficiency in conducting interviews. We assumed that these priorities could work against obtaining high quality responses if they encouraged interviewers to "barge in, rush through the interview, and get out." We redesigned the interviewer training. To shift priorities, we instituted a completely new system of monthly performance ratings that emphasized high quality responses as well as response rates and efficiency. For the evaluation test, we hired inexperienced interviewers for the experimental treatment because we assumed experienced interviewers would find it difficult to shift their priorities.

3.2 The Experimental and Control Treatments

On the basis of the assumptions we devised an experimental interviewing procedure that we felt would reduce the underreporting of participation in the 4 selected programs by 25%.

3.2.1 Basic Procedures. The experimental procedure placed the highest priority on acquiring accurate income responses, making the extra effort to achieve quality even if it might increase costs or decrease response rates. To do this we ended up making changes to just about every aspect of SIPP, the training, the questionnaire, the interviewing procedures, the supervision, and the data processing.

We included a lot of features to get respondents to use their personal records. We completely revised interviewer training to emphasize skill in getting accurate responses, with less emphasis on efficiency and response rates. We required self-response from people 15 years and older whenever possible during the first interview. We insisted on a distraction-free interview setting. And we emphasized that the Census Bureau was more than willing to pay the cost of callbacks in order to meet these requirements, callbacks to get missing records, callbacks to get self-response, callbacks to get a distraction-free setting. Interviewers also got monthly feedback about how well they and their respondents were implementing the quality-oriented procedures. The feedback was based on tape recording all interviews, coding a sample of them, and summarizing the codes as soon as the interviewer had completed the monthly assignment.

The control procedure for the experiment was the Standard SIPP interview, conducted by interviewers who received standard SIPP training and who were given standard performance feedback about their costs, response rates, and recording mistakes caught during the clerical edit of the questionnaires. The control procedures involved following a scripted questionnaire and complex skip patterns to ask all questions in a prescribed order for one person at-a-time. Self response was encouraged if the person was present when the interviewer called at the household. Efficiency and high response rates were encouraged via training and monthly feedback to each interviewer. Quality control was achieved with a telephone reinterview of a sample of each interviewer's work.

3.2.2 Personal Records. The experimental treatment emphasized using personal records, the control treatment did not. At the outset of the standard SIPP control interview, the interviewer read a statement which suggested that the respondent may want to consult available records if he cannot recall information

[4] For an illustration of problems with skip instructions in early SIPP, see Hill (1993).

from memory. No further mention of record use was required. In the experimental interview, personal record use was the keystone around which everything else was made to fit.

By personal income records we mean written records of how much was paid, to whom, from whom, and when. Types of records include the stub or slip that comes with a paycheck, bank statements, brokerage account and mutual fund statements, personal ledgers, passbooks and program authorization letters.

Once we made the decision to emphasize personal income records, we decided to change the questionnaire to require reporting each individual income payment. Standard SIPP, on the other hand, requests monthly summaries of payments from each income source. This change meant that we had to completely revamp the computer processing system to produce smooth, monthly summaries of the individual payments in each income stream.

In the experimental treatment, the interviewer suggested that the respondents get their records at the outset of the interview. The request was made as if this were what every government survey expected of its respondents. The interviewer was trained to be comfortable with the so-called "wasted time" that elapsed while respondents sought out their records. For each income source reported, the interviewer asked what records came with each income payment and, if necessary, asked the respondent to retrieve those records. If the respondent had discarded the records, the interviewer explored the possibility of the respondent's getting replacement records from the income source, offering either to telephone or revisit the household to record the missing information. Near the end of the first interview the interviewer noted which income sources were reported without a complete set of records. The interviewer then instructed the respondent about how to save future records for that source or, if necessary, how to write down the key details of each payment for use in the next interview. At this point the interviewer gave the household an attractive record keeping folder to keep future records in. (In debriefing interviews, this folder was about the only thing most respondents said they liked about the interview besides the interviewer.) Also at the close of the first interview, the interviewer asked permission to telephone the household and remind them to save their records. Although most respondents granted the permission, interviewers seldom made the reminder calls. All other aspects of the record-use procedures were implemented well, according to the tape recordings, and, as we will see later, record use rates were very favorable.

3.2.3 Questionnaire. The main questionnaire change was to shift from asking specific questions about each income source to an open-ended format. Instead of asking about income from jobs and then income from AFDC and then questions about food stamps, etc. the experimental procedures told the respondents that we wanted them to report all their income and to report it using their records. The instructions used to prompt complete income reporting and use of personal records are in the Appendix. In standard SIPP there is a specific time to report income sources for each person and a different time to report income amounts. In the experimental treatment, it was up to the respondents to decide what to report and when and for whom. One person could report all his income and then someone else could report. The respondents could list all their income sources first and then get into the details of payment dates and amounts or they could do some of each. Basically the respondents were in control of this part of the interview and structured it however they felt most comfortable. The interviewer had a set of concrete information objectives for each income source reported and asked unscripted questions as necessary to meet the objectives.

After the free recall section was finished, the experimental interviewer used a set of recognition lists to make sure the respondents had thought about all income sources relevant to them. Many additional income sources were obtained via the list, primarily income from assets. During this part of the interview, the interviewer structured the reporting by requesting the detailed information about dates and amounts

as soon as a new income source was uncovered.

When respondents did not have personal records for an income source and felt that they could not obtain replacements, experimental interviewers were instructed to use special reconstruction strategies to improve recall. Basically, the strategies were to elicit whatever heuristic rules the respondent wanted to use to report the payment dates and amounts and then to probe for exceptions. For example, "when do you get your check if it is supposed to arrive on a holiday?" "Did you ever work any overtime?" or "Did you get a cost of living increase?" The control group interviewers did not use special reconstruction strategies.

The standard SIPP questionnaire in use during the experiment contained very complex skip patterns and some organizational problems. For the experimental questionnaires we did a lot to simplify procedures and to reorganize things to increase both the respondent's and the interviewer's understanding of what we wanted.

The experimental interview took a different approach to assuring the accuracy of reports about income near the time boundaries of adjacent interviews (the seam). First the experimental interview encouraged respondents to report any income they received in the interview month (between the first of the month and the day of the interview). We call this the overlap period because, in the next interview, respondents also reported income they got during this period. The reports in the second interview were "independent," however, because we did not remind the respondent of what had been reported previously. After all income had been reported in the second interview, the interviewer pulled out the reported income information from the first interview, matched it up with the income reported in the second interview, called the respondent's attention to any inconsistencies in the overlap period or to income reported in only one of the interviews, and obtained the respondent's information about what was most correct. Corrections were recorded and became part of the computer data base.

The Standard SIPP Control procedure was to remind the respondent of what was reported in the last interview and to record whether the report was correct or not. Data processing constraints did not permit information given in a prior interview to be corrected in the control group data base.

3.3 The Experimental and Data Collection Designs

To evaluate the two treatments, the experiment randomly assigned sampled households to the interviewing treatments and conducted up to two interviews with each household. The reference or recall period was the previous four months. To spread the work efficiently over time, we divided the sample into four rotation groups, and began interviewing the first group in September, 1993. All interviews were conducted face-to-face at the respondent's household.[5]

Two interviewing staffs were used, one for each treatment. No interviewer worked on both treatments. The experimental interviewing staff was much larger, less experienced, more racially diverse, and assisted by an inexperienced crew leader. Assignment sizes per interviewer were relatively small in most cases. One interviewer was terminated for fabrication and missing deadlines. The control interviewing staff was smaller, more experienced, white and guided by an experienced crew leader who ended up doing a substantial part of the work. There was some turnover but it was voluntary. We attribute the

[5] Start-up problems forced us not to conduct wave two interviews with the September rotation group. Instead we added a fifth rotation group to each treatment and interviewed it twice. Data from all groups are used in the analyses unless specified otherwise (e.g., when the effect of wave or time is estimated).

differential response rate and cost results, mentioned later, in part to differences between the interviewing staffs and in part to the intrinsic features of the experimental and control procedures. But since treatment and staffing characteristics were confounded, we cannot estimate how much each contributed.

The interviewers were supervised from the Kansas City Regional Office, several hundred miles from the experiment site. Later in the field period, crew leaders assisted in supervisory tasks.

3.4 The Sample and Record Check Evaluation Designs

Our evaluation goal was to reduce underreporting of participation in the major government transfer programs[6] by at least 25%. To be eligible to underreport, a person must have been participating in a program. To approach the objective efficiently, we drew samples of people whom we knew were participating in one of four programs at some time during the interview reference period. The four programs were Aid to Families with Dependent Children (AFDC), Food Stamps, Supplemental Security Income (SSI) and Unemployment Insurance (UNEM). To learn about wage and salary reporting errors, we also drew a small sample of people who worked for a large employer in the area (JOB).

We created an initial frame by randomly sampling people from each source who had a zip code within our interviewing area, the city limits of a moderately sized midwestern city within the jurisdiction of our Kansas City Regional Office. We sampled twice from each source, in June and September. This assured that some sampled people would have truly participated in one of the programs during their interview reference period. We unduplicated names across the income sources and the two samples. We eliminated ineligible selections such as people without a street address on the administrative record. For our final sample, we stratified our frame on program and zip code and set our final sampling fractions so as to yield about 700 wave 2 interviews after taking into account, for each stratum, the likely nonresponse rates and the probability of actually finding the sampled person at the address when we visited it.[7]

After interviewing was completed, the administrative record sources sent us participation and income amount information for each person originally selected from their records.[8] They also included the sample person's social security number and often included other information such as name. We used this auxiliary information to match the survey and administrative record reports. We refer to this group as the Sample Persons Group and use it to estimate participation underreports.

We submitted a second list of names to each welfare record source in order to construct a large enough group to make overreporting estimates. This list consisted of two groups:

1. all interviewed adults in the experiment who were not part of the original frame list provided by the agency and

---

[6] Social security is the most important government transfer program. Participation in it was seldom either underreported or overreported in the study of the 1984 SIPP. Hence, it was excluded from the experiment.

[7] Interviewers were not informed about the record check or the name of the sampled person. After the initial interview was completed, the names of the people in the household were sent to headquarters. Headquarters determined whether the sample person was listed in the household. If not we dropped the household from the wave 2 interview assignment lists.

[8] Getting record information for people not actually in our sample was one measure used to protect the privacy/confidentiality of people actually sampled. The record-providing agency did not know whether a given person on the frame list was actually in the sample or not.

2. all people from other agency frame lists whom we did not select for the final sample (foils).

We refer to the first group as Other People in Households of Sampled People.[9]

For each Sampled Person and each Other Person, we extracted their social security number as reported in the interview and sent it, along with their name, age and sex to our employees at the Social Security Administration for verification. If the number could not be verified, the people at the Social Security Administration either told us what the correct number was or told us that they could not verify the information.

We sent the social security number information for the Other People group to each agency. The agencies searched their records for the listed people, based on the social security numbers we provided. If a person received income from the agency during the one year period that included the interview reference periods, the agency sent us the income information.

We matched agency and survey information based on social security number and determined, on the basis of auxiliary information such as name, whether the match was correct. A small number of incorrect matches due to submitting multiple or incorrect social security numbers were eliminated. Then survey and record participation and income data were compared in order to make the response error estimates, primarily the participation overreporting calculation.[10]

[9] All sampled people were on both kinds of lists but not simultaneously for a program-specific estimate. For example, someone sampled from the SSI frame was on the Other Persons lists for AFDC, Food Stamps, and Unemployment Insurance.

[10] At this point in the data analysis, we do not claim that the Other People sample is representative of any particular population.

4.0 RESULTS

This section contains results of the comparisons of the two interviewing treatments on several outcome variables. The results do not show important treatment differences for the underreporting of program participation. Most such underreports were due to never mentioning the source at all. Generally, both treatments produced about the same rates of overreport errors also. There were differences, however, in the pattern of errors for the measures of transitions in program participation status, especially for transitions on the seam between interview reference periods. The experimental procedure did get more accurate reporting of benefit amounts in the second interview of the panel. But the experimental procedure cost considerably more to use and it achieved much lower response rates. Those people who did cooperate in the experimental procedure used their personal records at much greater rates than people in the control treatment and the record use, for experimentals, was clearly associated with higher quality reporting over time. Each of these topics is discussed more fully next.

4.1 Program Participation Underreporting

The goal for the experimental procedures was to reduce the underreporting of months of program participation by at least 25%. The people sampled directly from administrative records were used to make the participation underreporting estimates. To be eligible to underreport a month of participation, the administrative record must have indicated that the person was participating in the program that month. If the participation was reported in the survey, it was considered a correct response. If the participation was not reported, it was considered an underreport.

Survey Report	Administrative Record Value	
	Yes	No
Yes	a	b
No	c	d
	a + c	

Figure 2: Cell C represents survey underreports in this cross-classification of administrative record and survey reports

This can be seen in Figure 2. An underreport is indicated by the letter c, where the administrative record indicates participation but the participation is not reported in the survey. The underreport rate is c divided by a + c. These were averaged over months and people to obtain the underreport rates or percentages used in the analyses.

4.1.1 Basic Results. While we expected a 25% difference between the experimental and control groups, the actual results (Figure 3) show essentially no difference in participation underreporting between the treatments. For example, there was 12% underreporting of AFDC participation by the experimental group and 10% underreporting by the control group.

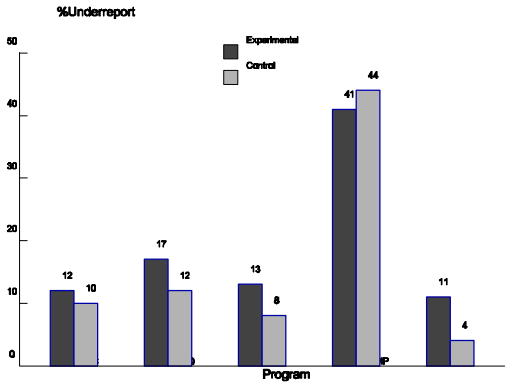

Figure 3: The experimental treatment did not reduce the underreporting of program participation

For each of the programs, the treatment difference was not statistically significant, indicating that the experimental and control groups made about the same levels of participation underreporting errors. Not only did the experimental procedures not reduce participation underreporting errors, the trend is for that procedure to perhaps generate slightly more error.

For unemployment, both the experimental and control procedures produced astonishingly high levels of underreporting: over 40% error. This is in the same range as we found in the study of the 1984 SIPP. It will be important to keep this result in mind when designing new procedures to mitigate response errors.

This is the first time we have estimated job underreporting errors for SIPP and the news appears to be good. The underreporting percents for "participation" were relatively low. They are also consistent with the low gross error and net bias rates in the published methodological literature on wage and salary reporting (e.g., Marquis, Marquis and Polich, 1986).

We will look at the basic underreporting results a couple of other ways for insight into why the expectations were not confirmed.

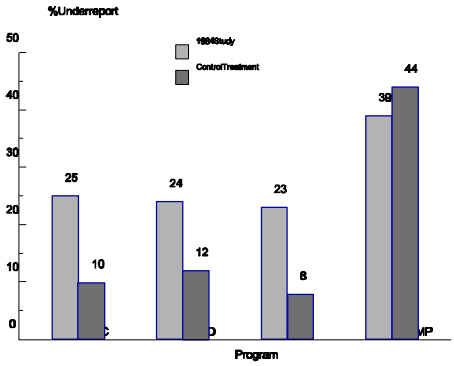

Figure 4: The historically high underreporting of program participation had been reduced already.

4.1.2 Underreporting Already Low. In Figure 4, it is clear that the control group underreport rates are substantially below those obtained by similar, standard procedures in the 1984 study, for 3 of the 4 programs. We had expected the rates to be the same for both studies. Yet the error levels appear to have been reduced considerably without using the new experimental procedures. In other words, much of the error we sought to eliminate by using new procedures had already been eliminated for other reasons.

One reason may be that the populations in the two studies were different. The 1984 study was based on cross section samples in 4 states. The current study used mostly households from one inner-city that we sampled from administrative records. But knowing these population differences, I suspect few would have predicted the direction of the underreport difference, let alone its size.

Either the control interviewing staff was exceptional or, more likely, the implementation of SIPP procedures has improved over the years and SIPP interviewers have become much better at conducting the standard interview. As a result, much of the response error we sought to reduce by the experimental procedures had already been reduced.

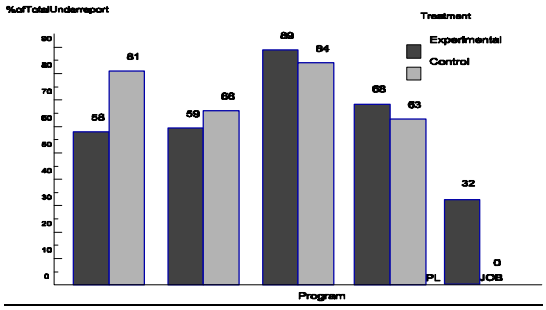

Figure 5: Most underreporting was due to not reporting the entire income source.

4.1.3 Most Underreports Due To Omitting Entire Source. Another concern was whether respondents were making occasional errors by underreporting part of the income stream or whether it was the whole set of income payments that was underreported. So I looked at how much of the observed participation underreporting stemmed from failing to report an entire income source and how much for other reasons. As shown in Figure 5, for both treatments and 4 of the 5 income sources, most of the underreporting occurred because the respondent never mentioned the income source. "Never mentioned" means that this source of income was never mentioned by anyone in either interview for any household member, including not ever overreporting it and not ever correctly reporting it.

In retrospect, it now seems clear that getting people to use their personal income records really is not going to improve their ability to remember an entire income source nor is it likely to increase their willingness to report income that they have decided not to report.

4.2 Overreporting Errors

The other kind of error a respondent can make is to overreport participation in an income program. If the administrative record indicated that the respondent did not participate in the program in a given month, but the respondent reported participation, an overreporting error resulted. This is cell b in Figure 2. The overreport rate comes from dividing a person's monthly b frequency by b-plus-d, the frequency of true nonparticipation according to our inferences from the administrative records. These rates are averaged over months and people for the next set of results.

Overreport rates are usually low relative to underreport rates. The low rates can be deceptive, however, because their effects depend partly on the incidence or prevalence of what is being overreported. For example, a one percent overreport rate among a sample of 2000 people, 99% of whom are true nonparticipators, results in 20 errors. A ten percent underreport rate among the one percent of true

participators results in only 2 errors. Even though the underreport rate is 10 times the overreport rate, the resulting net bias results in a substantial overestimate of participation level.

Survey evaluators have often overlooked the possibility of overreporting errors, either because of their low rate or perhaps because their underlying theories of memory decay deal only with underreporting. Recently in psychology there have been several tragic demonstrations of the capacity of people to sincerely believe and report doing things that they never did. Among the more well known is Ofshe's implanting of a false memory of child abuse into someone who later confessed to committing that exact crime (cited in Loftus, 1993). Beth Loftus has also published several demonstrations of implanting innocuous false memories in ordinary people who come to believe them and report elaborate reconstructions of events related to the false memories. In terms of survey methodology, a very convincing demonstration of dominant overreporting errors is given in an excellent study of the national Fishing and Hunting Survey (Chu, et al., 1992).

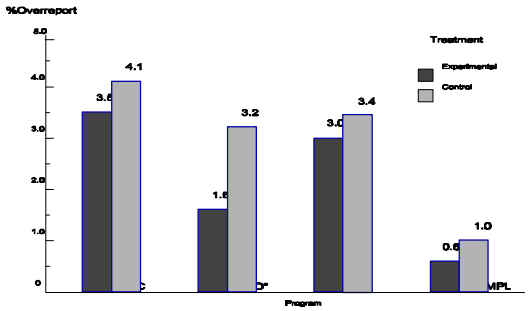

Figure 6: The experimental treatment had little or no effect on the overreporting of program participation.

For this research, overreports were estimated using the "other people" in sampled households. To overreport, it was necessary not to have participated in the target program for the month being analyzed ("No" according to administrative records in Figure 2). The "other people" sample was specifically designed to meet this requirement. This group was not representative of the entire population of people who truly did not participate so point estimates should not be generalized that way[11]. But the group provided a basis for a useful comparison of the effects of the experimental and control treatments. We learned what effects the experimental treatment had on a group very much more likely to overreport because they lived in households that did receive government transfer payments of at least one kind.

[11] The U.S. population has a much higher incidence of middle class households who do not participate in the kinds of government income transfer programs studied here.

Finally, for practical reasons we did not design our sample to be able to estimate overreports of jobs at the large employer in the study. The "job" income source does not appear in the results graphs for overreports.

The overreport results, in Figure 6, show very little effect of the treatments on overreporting of program participation within the special group. There is a trend for the experimental treatment to get fewer overreports of participation in each of the four programs. The difference is statistically significant ($p < .05$)[12] only for the Food Stamps (FOOD) program. The main conclusion to be reached from the empirical results is that there is little or no effect of the new interviewing procedures on reducing overreporting of program participation in this special group.

4.3 Transition Reporting Bias

There is considerable interest in the errors in reporting transitions in participation or, as some say, the beginning and ending of "spells" on programs. In the past, even without record check research, we knew that there were more transitions detected on the seam between interviews relative to changes taking place between comparable time periods within the interview reference periods (e.g., Moore and Kasprzyk, 1984).

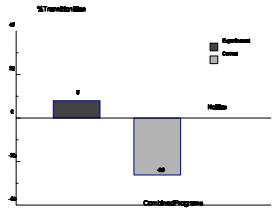

Figure 7: Treatments have different transition reporting biases.

A participation transition is inferred from the reports of program participation in adjoining months. Any change in a person's status (e.g., from receiving to not receiving benefits or from not receiving to receiving benefits) is considered a transition. To provide an early look at the transition data, I have taken some short-cuts in the analysis. Instead of calculating underreporting and overreporting error scores for each person, I have merely counted the number of transitions reported for each program for all eligible sample people in the program and subtracted them from the count of the number of transitions for those people in the administrative records. (I divided the difference by the true number and I call the result, multiplied by 100, the percent transition bias). Because program participation transitions are relatively rare events, I have combined the individual results for both sample groups and for the five income sources. Although producing more easily interpretable results, this approach allows errors to be offset across months for the same person and across people in the same treatment group; by treating all changes alike, it ignores the distinction between starting and ending participation. These results are preliminary and they may be influenced by special data processing procedures that were used for the experimental treatment.[13]

---

[12] 2-tailed, t-test, ignoring any effects of intrahousehold clustering due to more than one adult interviewee per household. None of the tests in the paper takes account of geographic clustering of sample households because there was no deliberate clustering in the sample design.

[13] For programs that issue monthly checks, the payment-by-payment reporting method often resulted in observing two payments in one month and no payments in the next month. In the experimental treatment, respondents were not explicitly asked if they received a payment in each of the 4 months of the reference

The combined data in Figure 7 suggest that the experimental treatment obtained about the right number of transitions in program participation (just 8% too many) and that the control treatment produced too few (detecting about 25% fewer changes than appear in the administrative records).

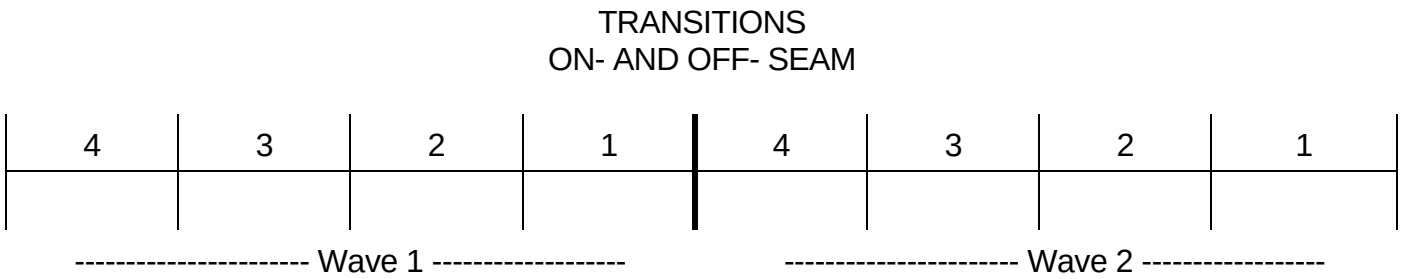

Figure 8: A transition "on the seam" (between interviews) happens between month 1 of the wave 1 interview and month 4 of the wave 2 interview.

Prior research has indicated that relatively more transitions are measured on the seam between interviews than between months measured within the same interview (e.g., Burkhead and Coder, 1985). Other research (the 1984 study) suggests that too many transitions are measured on the seam and too few elsewhere.

Figure 8 depicts what we mean by on and off the seam. Changes within the interview reference period (between months 4 and 3, 3 and 2 or 2 and 1) are off-seam. Changes between the interview reference periods (from month 1 of the wave 1 interview to month 4 of the wave 2 interview) are on-seam.

The pattern of errors for on- and off- seam transitions for each treatment is in Figure 9. The control group results are consistent with what we found in the 1984 study: there is net overreporting of transitions on the seam and net underreporting of change within the interview

PERCENT BIAS IN NUMBER OF TRANSITION REPORTS

	On-Seam		Off-Seam	
TREATMENT	True N	% Bias	True N	% Bias
Experimental	18	**106**%	147	- **3**%
Control	21	**40**	169	-**40**

Figure 9: For both groups, change is overreported on the seam. For the control group, change is underreported off the seam.

period. We inferred that someone participated in a program if he got a payment that month. But if the normal receipt date fell on a weekend or holiday, many programs mailed or distributed their checks early. And many experimental group respondents reported these altered receipt dates correctly. To avoid creating artificial transitions, we applied a computer algorithm to smooth out these spurious transitions in both survey and administrative record data for the experimental treatment.

The second unique processing feature arose because of the overlap in interview reference periods. In the experimental group, the first and second interviews covered a common time period or overlap period which was at the end of the first interview's reference period and the beginning of the second interview's period. Interviewers were supposed to spot any duplicate income reports for this period and correct the data so only one report remained. Occasionally an interviewer failed to do this so we programmed the computer to detect and correct these oversights. Both the computer smoothing and deleting duplicates were unique to the experimental treatment and may have affected our transition estimates in unknown ways.

reference periods. The experimental group results for the on-seam bias are a surprise to us. The data suggest that there were too many changes occurring on the seam, even though about the right number were reported for the other times. (We expected the no-bias result for the off-seam changes but, because we took elaborate procedural precautions, we expected little or no on-seam bias either. The large, positive bias was not expected. Small n's may be misleading us here.) This is very much "work in progress." We will be following up on these results in the future.

4.4 Error In Reported Income Amounts

Next we will examine the effects of the experimental procedures on errors in reporting amounts of income. These results are fairly encouraging since they suggest that the experimental procedures eventually cause an important improvement in reporting quality.

This analysis also uses the group of Original Sample People. To be eligible for inclusion in this analysis, the respondent and the administrative record must have agreed that the sample person participated in the program for the given month. Then I compared the reported amount to the amount in the records for each month, averaging over months and people for each program. To test for time effects, I included only original sample people who were interviewed (about) in both wave 1 and wave 2.

For these analyses, the reported amount is considered correct if it is within 5% of the recorded amount. This avoids having to deal with transformations of the distributions and making decisions about the inclusion of outliers. It also lets us include don't know answers, as incorrect, instead of having to impute for them and then difference the imputed and true values. Finally, it eliminates the effects of a small difference between the treatments due to the experimental amounts being recorded in dollars and cents while the control amounts are recorded in whole dollars only.

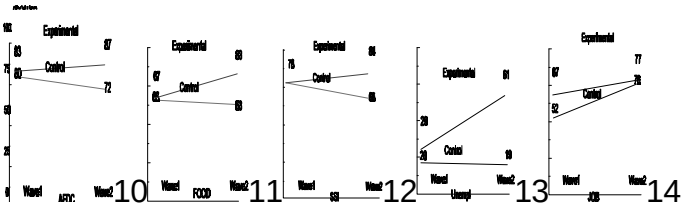

Figure 10: The experimental treatment usually produced better reporting of income amounts by wave 2.

Using the 5 parts of Figure 10, we can look at the effect of treatment and time (Wave 1 and Wave 2 interviews) on the percent of amounts correctly reported. The basic result is that, over time, reporting is better in the experimental group than in the control. For most of the 5 programs the treatments achieved about the same percentage of correct amount reports in the first interview. However, by the end of the second interview, the experimental treatment was usually producing substantially better reports than the control.

For the first three income sources (AFDC, Food Stamps and Supplemental Security Income), the patterns were similar: both treatments got about the same percent correct in the wave 1 interview, but the experimental treatment got a higher percentage of correct amount reports in the wave 2 interview. For each of these income sources, the treatment-by-wave interaction was significant ($p \leq .05$) in a repeated measures analysis of variance using people who correctly reported their participation in both waves.

For unemployment, even though we had a very small number of cases, the trends were similar to the long-term welfare programs: the treatment groups had similar error levels in the first interview and there was better reporting from the experimental group in the second interview. Only the main treatment effect was statistically significant, however.

Finally, the percent correct amount reports for earned income from a job are in the last panel of Figure 10. Although there appears to be a small difference between the treatments, especially in wave 1, neither the treatment main effect nor its interaction with wave was significant statistically. The good news is that respondents in both treatments improve their reporting over time and the main effect of wave was significant in the repeated measures ANOVA.

I find these results encouraging. I am certainly pleased that income amounts reporting seemed to improve because of the experimental procedures, even if the effects were somewhat delayed in time. This is what we wish to happen in panel studies, for our respondents to learn what we want them to do and to get better at doing it over time. Right now, I can only wonder what the results would have looked like if we had run the experiment for another few waves. Would reporting accuracy have continued to improve over time in the experimental group? Would accuracy have declined in the control group?

4.5 Response Rates and Costs

While the experimental procedures got people to use personal records and appear to have increased the quality of income amounts reporting in the second interview, these gains may have come at too high a price in terms of response rates and costs per interview. We will look at costs and response rates next.

But keep in mind that the interviewers using the experimental procedures were largely inexperienced and were often given inefficiently small assignment sizes. Some of the control group interviewing staff also worked on regular SIPP during the experiment and it was difficult to get accurate costs for each activity separately. While the experimental procedures themselves probably drove costs up and response rates down, the confounding of experimental treatment and interviewer characteristics makes it hard to tell how much each contributed.

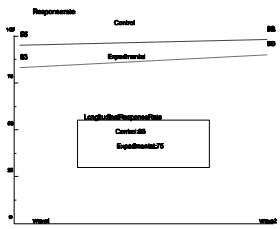

Figure 11: Response rates were considerably lower in the experimental group.

The household response rates for wave 1 and wave 2 for the control and experimental treatments are in Figure 11. The response rate is the number of completed household interviews divided by the number of eligible households in the sample during that wave. Reasons for classifying an eligible household as not responding include repeated calls with no one at home, a final refusal, or all occupants were temporarily away during the interviewing period (e.g., on vacation). Control group response rates were higher than experimental group response rates in both waves. The regional office also reported that they were higher than for regular SIPP in the same area during the same time period.

The longitudinal response rate reflects the proportion of eligible households interviewed through wave 2. The control group lost 7 percent of its eligible households to original nonresponse and attrition. The experimental group lost 25%. I feel that the 25% loss would be unacceptably high in a production survey. Had we gone on to further develop and test the experimental procedures, we planned to inaugurate several new features to achieve acceptable response rates. These included using already experienced interviewers.

The experimental procedure also cost more to conduct than the standard control procedure. The cost-per-case data in Figure 12 reflect hourly pay to interviewers for both interviewing time and travel time. They also include reimbursement for automobile mileage associated with completing their interviewing assignments.

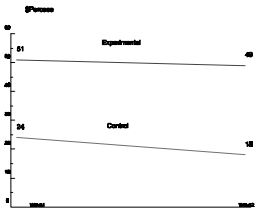

Figure 12: Cost per case was much higher for the experimental group.

Cost per assigned case was at least twice as much in the experimental treatment compared to the control (Bogen, et al. 1993). Had we gone on with this research, the next project would have used more experienced interviewers, larger assignment sizes per interviewer, and minimized the possibilities for charges going to the wrong project. Nevertheless, we would still expect the experimental procedures to continue to cost somewhat more for several reasons: 1) the interviewer must make additional callbacks to get missing information, to arrange initial family-group interviews, to obtain initial self-response and an interview setting conducive to good reporting; 2) interviews took longer since respondents needed to retrieve personal records and because they report income payment-by-payment rather than in monthly chunks.

4.6 Personal Record Use

Fundamental to the design of the experimental procedures were the assumptions that we could persuade respondents to use their personal records to report their income and that they could learn to do this well enough to improve the reporting of their income. This section examines those assumptions. The results will show that we were successful in getting experimental group households to use their records and that such record use was beneficial.

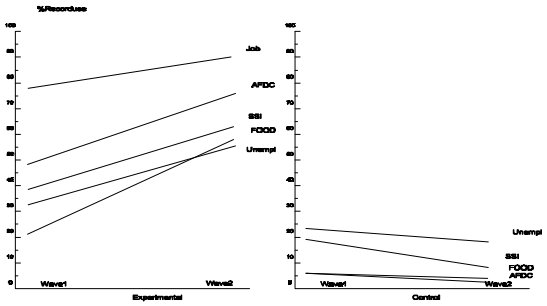

Figure 13: Personal record use rates were generally higher in the experimental than in the control group for all programs. Experimental group rates increased even further in the second interview.

4.6.1 Personal Record Use Rates. How often were personal records used to report income? Record use was defined as using at least one record to assist in the reporting of details of an income source. Record use was measured differently for each treatment. For interviews with experimental households, each time an income payment receipt was reported, the interviewer checked whether a personal record was used to substantiate the reported date and amount. For the control group, interviewers followed standard SIPP procedures by marking, on a separate page, which income sources were reported by each person and whether at least one record was used for each reported source.

We may use Figure 14 to look at the program-specific record use rates. These results are for people who reported receiving income from the source in both waves. The graphs show that people in the experimental group used records at much higher rates than control group people for each of the programs we have been analyzing.[14] Prior to this study, almost nobody thought that it was possible to get

[14] The selected programs for which this may not have been true are: JOB--because we did not measure record use for job income reporting in the control group and UNEMPLOYMENT--because, due to the small numbers of cases, the treatment differences are not statistically significant.

Rates of using personal records in the national 1991 SIPP panel, according to internal memoranda (Singh, 1991 and 1992), were:

Program	Wave 1	Wave 2
AFDC	11	9
FOOD	13	12
SSI	27	22
UNEMPLOY	20	19

Except for Unemployment, these national rates tend to be slightly higher than those achieved by the control

record use rates at these levels in a household sample survey about a sensitive topic. The graphs also show that record use performance improved considerably in the second interview for the experimental group but not for the regular SIPP controls (because of small n's, the treatment-by wave effect is not statistically significant for unemployment).

4.6.2 Effects on Amounts Reporting

Now it is time to ask the "So What?" question: If there was more record use in the experimental treatment, did it have any of the desired effects on reporting quality? I will try to answer that question next.

Because treatment and record use are highly correlated and because the definitions differ for the two treatments, I will look just at the effects of record use within the experimental treatment. For this analysis record use is still a yes-no variable: if the respondent used a record in either wave to report the income from the given source, the value is "yes" on the record-use variable. The dependent variable is, again, the percent of income amounts reported correctly in the survey. The amount is correct if it is within 5% of the correct value according to administrative records.

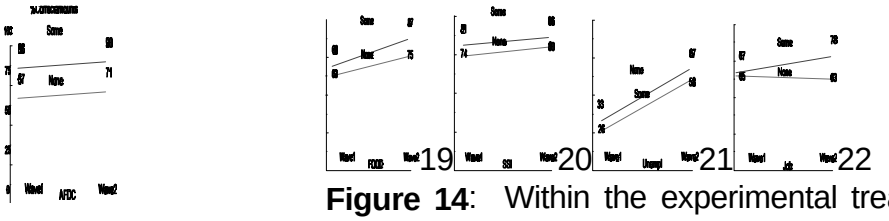

Figure 14: Within the experimental treatment, using some records (vs. none) usually produced better reporting of amounts in either wave.

Looking at the results for the three long-term welfare programs, the trend was clearly for higher percent correct scores to accompany some record use. In the first three comparisons the "some" line is higher than the "none" line, indicating that some record use got higher percents of correct reporting of income amounts.The record use effects were statistically significant for AFDC and Food Stamps. There was also a trend for people to improve their reporting over time, regardless of whether they used records. Since these estimates were not based on large numbers of cases, these effects were not always statistically significant. But the trends were for record use and time to be beneficial.

The remaining two income sources, unemployment insurance and job income show mixed pictures. For unemployment, the trend was actually for the non-record-users to be a little better than the record users. But the n's were small and neither the record-use nor the time effect was statistically significant.

For job income there appears to be the familiar interaction between treatment and time: things got better as the experimental panel continued, relative to the control panel. However, none of these apparent effects was statistically significant.

To recap, record use rates were high in the experimental treatment and low in the control treatment. The rates got higher over time in the experimental group. Some record use was better than no record use in

---

group in wave 1 of the experiment. Like the control group, the national rates also seem to get a little lower in Wave 2.

the experimental treatment. This was true for most income sources although the effect there was not always strong.

5.0 DISCUSSION

What did we learn from this study and how will it be carried forward in future SIPP measurement?

The main thing we learned was that the experimental procedures are not the ultimate answer to reducing response error. As a result we have abandoned the multi-million dollar program to refine and implement the new procedures and have substituted a new redesign program with a very different focus. A program of incremental changes in procedures and questionnaires will continue as SIPP is redesigned. And it is my understanding that respondents will be encouraged to use their personal records, but not as strongly as we did in the experiment.

There are still key design questions left unresolved for SIPP. How can we achieve better reporting of program income sources and program participation transitions?

For underreporting whole income sources, remedies really depend on what is causing the underreporting. If the causes are cognitive, such as forgetting or confusion about the program name, then better name recognition cues could help. Along these lines, it is probably better to ask specific questions about each income source of interest than to use the lengthy and uninteresting recognition lists. It is my understanding that the redesigned SIPP will continue to use and improve upon the specific questioning approach.

On the other hand, if the whole source underreports are intentional or motivated, different remedies need to be tried, solutions that address the privacy and confidentiality concerns that prompt intentional underreporting. We need to accommodate people who really do not want to discuss their income with other family members by conducting private interviews. And we need to find ways to be more persuasive about our ability to maintain absolute confidentiality for anything the respondent reports.

Now, let me end my speculative discussion of results with some thoughts about record use. We succeeded beyond our wildest expectations in getting households to use their personal records. But record use did not reduce error in reporting program participation, most of which stems from never reporting the program as an income source anywhere in either interview. So, I want to acknowledge that record use probably can't help someone remember a forgotten income source nor can it motivate someone to report income that he doesn't want to report.

Record use did increase the accuracy of reporting income amounts, especially in the second interview. This makes more sense. Once you acknowledge the existence of an income stream, using personal records can have favorable effects, especially after gaining some experience in interpreting the records.

So at this early stage of methods research on using personal records, I speculate that record use is not the magic bullet we sought to reduce all response error. High rates of household record use can be obtained but perhaps at too high a price with respect to response rates and costs. Record use can help improve the accuracy of reporting details but is of little help in getting people to report the income source itself.

I think it would be useful for survey methodologists to continue research on using personal records but to recognize that record use is probably a very specialized tool, useful in tackling a limited set of response

errors.

ACKNOWLEDGEMENTS

I would like to thank Jeffrey C. Moore, who was co-principal investigator throughout the entire project, and Karen Bogen, Nola Krasko, Richard Taegel, Elaine Fansler, Peter Wobus and Lorraine Randall, who had key roles throughout the research.

REFERENCES

Bogen, K., N. Krasko, J. Moore and K. Marquis (1993), "Preliminary Field results of an Alternative Measurement Design for the Survey of Income and Program Participation," Proceedings of the Section on Survey Research Methods, American Statistical Association, Vol. II, pp.1027-1031.

Burkhead, D. and J. Coder (1985), "Gross Changes in Income Recipiency from the Survey of Income and Program Participation," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 351-356.

Chu, A. et al. (1992), "Measuring the Recall Error in Self-Reported Fishing and Hunting Activities," Journal of Official Statistics, Vol. 8, pp. 19-39.

Cantor, D., S. Brandt and J. Green (1991), "Results of First Wave of SIPP Interviews," Memorandum for Chet Bowie, Feb. 21, 1991.

Hill, D. (1993), "Response and Sequencing Errors in Surveys: A Discrete Contagious Regression Analysis," Journal of the American Statistical Association, Vol. 88, pp. 775-781.

Loftus, E. (1992), "The Reality of Repressed Memories," American Psychologist, Vol. 48, pp. 518-532.

Marquis, K. (1990), "Report of the SIPP Cognitive Interviewing Project," Internal report to R. Singh, Chair, SIPP Research and Evaluation Committee, August 22, 1990).

Marquis, K., M. S. Marquis and J. M. Polich (1986), "Response Bias and Reliability in Sensitive Topic Surveys," Journal of the American Statistical Association, Vol. 81, pp.381-389.

Marquis, K., and J. Moore (1990), "Measurement Errors in SIPP Program Reports," Proceedings of the 1990 Annual Research Conference, Census Bureau, Washington DC, pp. 721-745.

Marquis, K., J. Moore and K. Bogen (1993), "Effects of a Cognitive Interviewing Approach on Response Quality in a Pretest for the SIPP," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 318-323.

Moore, J., and D. Kasprzyk (1984), "Month-to-Month Recipiency Turnover in the ISDP," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 210-215.

Schwarz, N., and T. Wellens (1994), Cognitive Dynamics of Self and Proxy Responding: The Diverging Perspectives of Actors and Observers, Final Report submitted to the U.S. Bureau of the Census for Joint

Statistical Agreement 91-3, November 1994.

Singh, R. (1991), "SIPP 91: Wave 1 Results of the Record Check Study," Internal memorandum to the SIPP Research and Evaluation Steering Committee, December 19, 1991.

Singh, R. (1992), "SIPP 91: Wave 2 Results of the Record Check Study," Internal memorandum to the SIPP Research and Evaluation Steering Committee, June 15, 1992.

APPENDIX

For the experimental treatment, to encourage using personal records and accurate, complete income reporting, the initial statement that the interviewer read to the household in the first interview was:

"This survey is about the economic situation of people in the United States. Our goal is to get a complete, accurate list of all the income, pay, and other money that you (both/all) received during the last 4 months.

If more than 1 adult:
"This includes income for (*read names to confirm all adults covered in current interview, including persons for whom proxy information is being given).*

(Show Payment-by-Payment Worksheet).
"Here is the worksheet I'll be using. I need to list who received the money, where it came from, the date it was paid, and the amount before deductions. Accuracy is very important so please use all the records you have for each kind of income.

(Hand calendar to each adult present.) **"Here is a calendar. The time period is the past 4 months** *(name the months)* **and up to today,** *(read month and date).*

"What pay and other money did (you/each of you) get since _____ 1st?

Note: Instructions to the interviewer are in italics. Statement wording is in bold type. The interviewer chose an appropriate term from the phrases in bold parentheses. In place of the blank in the last sentence, the interviewer inserted the name of the month that started the reference period.

THE SIPP MEASUREMENT QUALITY EXPERIMENT AND BEYOND: BASIC RESULTS AND IMPLEMENTATION

by

Kent H. Marquis
U.S. Census Bureau

Paper prepared for presentation at the
Census Bureau 1995 Annual Research Conference
March 19-23, 1995
Arlington, VA