THE SIPP COGNITIVE RESEARCH EVALUATION EXPERIMENT: BASIC RESULTS AND DOCUMENTATION

Jeffrey C. Moore, Kent H. Marquis, and Karen Bogen[1]

Center for Survey Methods Research
Statistical Research Division
U.S. Bureau of the Census

January 11, 1996

ABSTRACT

In response to known and suspected problems in the measurement of program participation and related variables, Census Bureau research staff developed new, experimental survey procedures, based on cognitive theory and research, to reduce response errors in the Survey of Income and Program Participation (SIPP). We implemented a "field laboratory" test of the new procedures, 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 as a substitute for faulty, minimum effort memory retrieval. Results of the SIPP Cognitive Research Evaluation Experiment indicate that the new procedures had no important effects on reducing either underreporting or overreporting errors in respondents' reports of participation in the income programs tested. However, by the second interview wave, the new procedures did produce substantial improvement in the reporting of income amounts. Experimental group households did use personal income records at astonishingly high rates; furthermore, record use correlated with the quality of income amount reporting.

This paper describes the basic features of the experimental procedures and their evolution; it presents the available evidence concerning the implementation of those procedures; it summarizes the key substantive findings of the experiment concerning program participation and program income reporting quality; and finally, it offers some possible reasons why record use did not affect reporting of program participation, but did have important effects on income amount reports.

---

[1] This paper reports the general results of research undertaken by Census Bureau staff. It borrows substantially from a recent summary paper prepared by the second author for the Census Bureau's Annual Research Conference (Marquis, 1995), and from several other earlier reports. The views expressed are the authors' and should not be interpreted as official positions of the U.S. Census Bureau.

THE SIPP COGNITIVE RESEARCH EVALUATION EXPERIMENT: BASIC RESULTS AND DOCUMENTATION

Jeffrey C. Moore, Kent H. Marquis, and Karen Bogen

1.0 INTRODUCTION

The Survey of Income and Program Participation (SIPP) 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.

Measurement error can be an important source of bias in estimates from surveys and censuses. Prior research from a full-design record check study on the first two interviews of the 1984 SIPP panel (Marquis and Moore, 1990) indicated some serious measurement problems in SIPP. We obtained administrative record data for four states for eight income programs: Social Security, federal Civil Service Retirement income, Aid to Families with Dependent Children (AFDC), Food Stamps, veterans benefits, unemployment insurance, workers' compensation, and Federal Supplemental Security Income (SSI). We matched the SIPP interview data to the administrative records and calculated the amount of error SIPP respondents made in their survey reports.

This research indicated much higher rates of measurement error than would be expected from a review of the survey methods literature on welfare reporting (e.g., see Marquis, Marquis, and Polich, 1986). For several key government transfer programs -- AFDC, Food Stamps, and SSI, for example -- approximately 25% of the true months of program participation were not reported in the SIPP interview. A fourth program, unemployment insurance income, had an even higher underreport rate -- almost 40% of true participation months went unreported (Marquis, Moore, and Bogen, 1993).

There are several different constructive paths to take once such problems are known to exist. One possibility is statistical correction -- use the response error information to adjust estimates derived from policy models (e.g., Bollinger and David, 1993; Bollinger and David, 1995). A second approach, and the one that is the focus of this report, is to design new survey measurement procedures to reduce the occurrence of the response errors in the first place.

For the design of our new measurement procedures we drew on the earlier record check research, which also addressed possible causes of measurement errors. That research allowed us to rule out several of the "usual suspect" causes of survey measurement error -- for example, memory decay, proxy response, learning to underreport, and variation due to individual interviewers -- as major causes of SIPP's program participation reporting errors

(Marquis and Moore, 1990). The record check analyses did suggest that small amounts of error could be reduced by eliminating occasional cognitive confusion about program names (e.g., confusion between Social Security and SSI) and confusion about the official recipient of the benefits. Other exploratory cognitive research (Marquis, 1990) suggested that the overly simple strategies that respondents tend to use to reconstruct their past income streams (often subtly encouraged by interviewers) might be at the root of a larger portion of the measurement errors.

2.0 THE EXPERIMENTAL MEASUREMENT METHODS

This section outlines some guiding assumptions we made about respondents, questionnaires, and interviewers, followed by a description of the new interviewing procedures that we created in an effort to reduce measurement errors.

2.1 Behavioral Assumptions Underlying the Experimental Procedures

The experimental interviewing procedures were a radical departure from the current, conventional SIPP interviewing techniques. Some of the basic assumptions that guided their design -- 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) -- are as follows:

2.1.1 The Respondent

We assumed that SIPP respondents have basically good intentions but are often simply unable to perform SIPP's primary task -- recalling accurately all the relevant details of each income stream, including especially the gross (versus net) amount. In many cases, respondents never know the income details they are called upon to report for other people in the household[2]. In other cases, details such as the name of the program or the "official" beneficiary are subject to comprehension mistakes. As a result, our well-intentioned respondents often use simplistic reconstruction strategies (based on general and error-prone knowledge) as a substitute for detailed, accurate recall or as a substitute for using personal records. Furthermore, we assumed that respondents who use these simple strategies are unaware or unwilling to acknowledge that such tactics are prone to error, and hence are not eager to change them. We designed our new procedures to preempt respondents' use of simple heuristic reconstructions and, instead, to substitute the use of accurate, complete information from personal records.

We assumed that well-intentioned respondents will often volunteer much useful and relevant information before it is specifically called for in the questionnaire script. Such information can easily get "lost" by an interviewer who conscientiously follows the

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

questionnaire sequence. We designed our procedures to accommodate important 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 records and be willing to save them for use in future interviews if they were asked to do so. We instituted a set of procedures, therefore, that clearly indicated, to both interviewers and respondents, our serious intention that respondents use available records to report their income.

2.1.2 The Questionnaire

The earlier research revealed additional minor problems among respondents in comprehending some questions and instructions. We assumed that most of these misunderstandings could be corrected by reorganizing the questionnaire into more logical, cohesive sections, and by making the objectives of each section explicit. Also, we assumed we could minimize interviewers' problems in following complex skip instructions by using such design formats only when absolutely necessary. For example, faced with a choice between a complex skip format and asking a question of a slightly larger universe of respondents than necessary, we often opted to abandon the skip, given that the added burden on respondents was minimal and that the skip could be recreated by computer edits. (For an illustration of problems with skip instructions in early SIPP, see Hill (1993).)

2.1.3 The Interviewer

We assumed that skilled interviewers can make any reasonable set of procedures "work" if they are taught the required skills and understand the priorities. SIPP standard practice is to reward interviewers for high response rates and high interview productivity. We assumed that these priorities often work against obtaining high quality responses if they encourage interviewers to "get the interview" at all costs, and to avoid any interaction with respondents that is even remotely challenging on the issue of response quality, because of the possible impact of such challenging behavior on future cooperation. We redesigned interviewer training to increase the focus on quality. We instituted a completely new system of monthly performance ratings that emphasized quality-oriented interviewing practices 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.

2.2 Design of the Experimental Procedures

Based on the above assumptions, we devised a set of experimental interviewing procedures that we felt would reduce substantially the underreporting of participation for selected income types. This section describes the primary features of the new procedures and

contrasts them with standard SIPP.

2.2.1 Basic Procedures

The experimental procedures placed the highest priority on acquiring accurate income responses, even if doing so might increase costs or decrease response rates. To support this priority we made changes to virtually every aspect of SIPP, including training, questionnaire design and organization, interviewing procedures, supervision, and data processing.

We added many new design features to encourage respondents' use of personal records. We revised the focus of interviewer training to emphasize skill in getting accurate responses over efficiency and response rates. We required self-response from all eligible adult sample persons (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 retrieve missing records, callbacks to interview a self-respondent rather than a proxy, callbacks to ensure 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 possible after the interviewer had completed a monthly assignment.

Standard SIPP procedures, in contrast, include performance feedback focused primarily on productivity, response rates, and questionnaire entry errors caught during a clerical edit of interviewers' completed work. The standard SIPP instrument is a completely scripted questionnaire with complex skip instructions. Interviewers' primary task is to ask all questions, as worded, in a prescribed order, for each eligible person in turn. Standard SIPP procedures recommend self response if the person is present when the interviewer calls at the household, but encourage use of proxies in order to complete a household in a single visit. Efficiency and high response rates are encouraged via training and monthly feedback to each interviewer. Quality control consists of a telephone reinterview of a sample of each interviewer's work, the primary purpose of which is to detect interview falsification ("curbstoning").

2.2.2 Personal Records

The experimental treatment emphasized the use of personal records for the reporting of income details; standard SIPP procedures do not. At the outset of a standard SIPP interview, the interviewer reads a statement which suggests that the respondent may want to consult available records if he cannot recall information from memory. No further mention of record use is required. For most income sources -- work-related income is an important exception -- interviewers indicate on each questionnaire whether or not the respondent used any records to report about the income source. This information is only sporadically analyzed, and only at

headquarters, not as part of any interviewer performance evaluation system.

For the experimental interview, personal record use was the keystone around which virtually all other procedures were designed to fit. For example, the decision to emphasize personal income records led to a parallel decision to change the questionnaire to require separate reporting of each individual income payment, as opposed to standard SIPP's request for the monthly total of all payments from each income source. This change meant that, for comparability with standard SIPP output, we had to revamp the computer processing system to produce monthly summaries of the individual payments in each income stream.

At the outset of the experimental interview, interviewers suggested to respondents that they use their records to report their income. We designed these statements to be as matter-of-fact as possible -- as if it were completely natural to request personal records for any official, important government survey seeking high quality income data. We trained interviewers to be comfortable with whatever "down time" elapsed while respondents sought out their records. For each income source reported, the interviewer asked what records accompanied those income payments, and, if the respondent hadn't already done so, asked the respondent to retrieve those records. If there had been records which were now no longer available, the interviewer explored whether the respondent could get replacement records from the income source, offering either to telephone or revisit the household when the missing information became available.

Near the end of the first interview the interviewer noted any reported income sources that were lacking a complete set of records. The interviewer 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. The interviewer gave the household a record keeping folder in which to save all future records for the next interview. Also at the close of the first interview, the interviewer asked permission to telephone the household and remind them to save their records.

2.2.3 Questionnaire

The experimental questionnaire was a radical departure from the standard SIPP instrument; instead of asking specific, scripted questions about each income source, it first used an unscripted, open-ended format. The basic approach was to explain the goals of the survey -- that is, that we wanted respondents to report all their income and, for maximum, "to-the-penny" accuracy, to report it using their records -- but to let the respondents dictate how to report their income, and when. We refer to this as the "free recall" section of the instrument[3]. One person could report all his or her income and then someone else could report, or the reports could be mixed together. Respondents could list all of their income sources first, and then the details of payment dates and amounts, they could alternate sources and details, or

[3] Appendix A is the Wave 1 experimental questionnaire; the free recall section of the interview appears on page 3.

they could do some of each. Basically, the questionnaire allowed respondents to report information in any order they chose. The interviewer had specific information objectives for each income source, and asked unscripted questions as necessary to meet those objectives. In contrast, the standard SIPP questionnaire focuses on one person at a time, and imposes a highly structured and scripted time to report income sources for each person, and a different time, equally structured and scripted, to report income amounts.

After free recall, the experimental interview employed a set of recognition lists to ensure complete reporting of all income sources[4]. This part of the interview structured the reporting task to the extent of requesting payment date and amount details as soon as a new income source was uncovered. As with the free recall, however, the interviewer knew what details were required, and used whatever unscripted questions were necessary to collect them.

For situations in which respondents did not have personal records for an income source, and could not obtain replacements, we instructed experimental interviewers on the use of special reconstruction techniques to improve recall[5]. These techniques were designed to elicit respondents' simple strategies for reporting 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 work any overtime?" "Did you get a cost of living increase?") Standard SIPP interviewers are not trained on any special reconstruction strategies, nor do they have any guides for using such strategies on site in the field.

The accuracy of reports about income receipt near the reference period boundary between adjacent interviews (the "seam") has long been of concern to SIPP (Moore and Kasprzyk, 1984; Burkhead and Coder, 1985). Although the exact mechanisms are not fully understood, the "seam bias" problem appears as an overabundance of income source changes at the seam relative to pairs of months within a single interview's reference period. Standard SIPP procedures try to minimize spurious change by using dependent interviewing procedures -- reminding respondents of income sources reported in the last interview and asking whether receipt continued in the current reference period. The standard SIPP questionnaire also permits recording that the prior report of receipt was incorrect, but processing constraints do not permit changing prior interview data.

The experimental procedures took a much more exacting approach to assuring the accuracy of income changes at the seam. First, we extended the reference period to the day of the interview. We refer to this last partial month of the reference period, between the first of the interview month and the day of the interview, as the "overlap" period, because it is also included in the reference period of the next interview. In Wave 1, experimental treatment

---

[4] See pages 4-8 of Appendix A.

[5] We devoted substantial training time to these special reconstruction techniques. In addition, they were outlined in the experimental questionnaire for interviewers' reference; see, for example, page 2 of Appendix A.

respondents reported income they received during the overlap period just as they did for the "standard" reference period, which was the preceding four calendar months.

Unlike Wave 2 interviews in standard SIPP, the Wave 2 experimental interview reports were initially independent of the Wave 1 reports; to avoid spurious consistency, we did not remind respondents what sources had been reported previously. After completing the Wave 2 free recall and recognition sections of the interview, when all income had supposedly been reported, the interviewer retrieved the Wave 1 income "worksheets'[6] and matched them up with the income sources reported in the second interview[7]. The interviewer pointed out to the respondent all income sources that had not been reported in both interviews, and either verified that this was correct or recorded the necessary details for the missing source. For income sources reported in both interviews, the interviewer examined all payment activity in the overlap period and, with the respondent, resolved any inconsistencies. These corrections -- even those that affected Wave 1 information -- became part of the data base.

3.0 EVALUATION EXPERIMENT PRETESTS

SIPP redesign budget and especially schedule constraints placed real limits on the research design for testing the new procedures. Ideally, we would have implemented a series of small-scale experiments to develop, test, and refine the major components of the new procedures individually; instead, we had to take an unquestionably "kitchen sink" approach, including all of the new procedures as a single package. Our developmental research program consisted of an initial informal field pilot test and two more formal small-scale field pretests prior to the Evaluation Experiment. The purpose of these tests was to assess the feasibility of the experimental interviewing procedures, to assist their further refinement, and to test some key features of our record-based evaluation of them. This section briefly describes the design and results of the pilot test and field pretests.

3.1 Pilot Test

In the spring of 1991, Westat, Inc., under contract with the Census Bureau, administered an abbreviated prototype of the experimental interviewing procedures to a small convenience sample of households in the Washington DC area. We limited the pilot test interview to only the collection of basic household roster information and income sources and amounts. The pilot test served as an initial feasibility assessment of some of the more radical of our new procedures, with which we and the Census Bureau field organization had little or no prior experience -- e.g., group interviews, tape recording, free recall of income sources, and especially the use of records. The basic issue was whether respondents would accept

---

[6] Appendix C contains the Wave 1 and Wave 2 "worksheets" used to record all relevant information about each reported income source.

[7] These "last wave review" procedures are on pages 10 and 11 of the Wave 2 questionnaire; see Appendix B.

these procedures, and, if not, how (or whether) they could be modified for greater acceptance.

The pilot test results were surprisingly positive. The contractor found no evidence to suggest that any of the basic features of the new procedures met with undue resistance from respondents, or were otherwise in need of important modifications. The pilot test yielded valuable insights into potential improvements to the details of how some procedures were implemented, but the "large picture" we drew from it was an endorsement to proceed with more rigorous testing. (See Cantor, 1991, for a detailed description of the pilot test and its findings.)

3.2 Pretests 1 and 2

Following the success of the initial pilot test, we made some modest refinements to the experimental procedures and shifted to more formal and controlled pretesting activities. We carried out two field pretests at the Evaluation Experiment site, the first focusing again on operational matters, and the second more on the details of how to implement the record check component of the evaluation. (See Moore, Bogen, and Marquis (1993) for more detailed descriptions of the pretest studies and their results.)

3.2.1 Pretest 1 Design

For the first pretest we drew a sample of 130 randomly-selected addresses. Relying on census income data, we used a sampling scheme which overrepresented poor areas in our selected site. This produced pretest field circumstances comparable to what we expected to be the case in the Evaluation Experiment, when our sample would consist primarily of poor households receiving benefits from one of several means-tested income transfer programs.

The Kansas City Regional Office hired the five-person Pretest 1 interviewing staff, which consisted entirely of people with limited interviewing experience, all of which was with the decennial census. The interviewers took a one-week training course before Pretest 1, covering basic SIPP concepts as well as the particulars of implementing the experimental interviewing procedures. Before Wave 2 we offered a short refresher training session, primarily covering new procedures specific to Wave 2. During the pretest field period we held frequent debriefings with the interviewers and their supervisors, to learn more immediately about how our procedures were working from the field perspective, and in some cases to make minor modifications "on the fly."

Pretest 1 interviewers completed 92 Wave 1 interviews in August and September, 1991 using a four-month (plus the interview month "overlap period") reference period. They returned to Wave 1 interviewed households in October and November and completed 74 Wave 2 interviews. Unlike Wave 1, the Wave 2 interview used an abbreviated, two-month

reference period[8].

The primary purpose of Pretest 1 was to continue to assess the feasibility of, and refine as necessary, the experimental survey procedures and instruments. We especially wanted to determine, under more controlled and realistic conditions, whether respondents would accept -- and interviewers would be able to administer -- the unscripted, "free recall" portion of the interview, and whether our procedures and forms were effective guides to stimulating this sort of information exchange and capturing the data such an exchange produces. A related question was respondents' acceptance of -- and interviewers' administration of -- the income source recognition lists, after already having gone through the reporting of income by free recall methods. We also wanted additional information about other important issues: respondents' willingness to be tape recorded, and our ability to use the tapes in an effective performance monitoring system; respondents' ability and willingness to find and use their income records; group interview logistics; and many other more minor procedural and instrument changes that we had put in place following the pilot test.

3.2.2 Pretest 2 Design

The second pretest followed immediately on the heels of Pretest 1. The primary difference between the two tests was that the sample for Pretest 2 consisted of 130 addresses of individuals drawn from the official record system of one of five income sources: AFDC, Food Stamps, Unemployment Insurance, SSI, or earnings from a specific area employer. In all other important respects, the design of Pretest 2 was the same as Pretest 1. It employed the same interviewers and essentially the same procedures, and therefore did not require any additional formal training program. It used the same basic interviewing design: two months of Wave 1 interviews (in December 1991 and January 1992) with a four-month reference period, and two months of Wave 2 interviews (in February and March 1992) with an abbreviated two-month reference period. Interviewers completed 88 Wave 1 interviews, and 79 Wave 2 interviews with households which had completed Wave 1.

The primary purpose of Pretest 2 was to test procedures for sampling from and matching to administrative record files, and to gain experience with data entry, database management, and data analysis. We also continued to monitor the experimental procedures and instruments.

3.2.3 Pretest 1 and 2 Results

The results of Pretests 1 and 2 were in general quite positive. On virtually all direct indicators, the new procedures worked well. About 75% of the 333 completed Pretest 1 and 2 interviews were successfully tape recorded, and virtually none of the taping failures resulted

[8] We shortened the Wave 2 reference period in both pretests in order to be able to complete the research program in time to meet survey redesign schedule deadlines.

from respondent reluctance. Over 90% of the 168 eligible adult respondents in Pretest 1 self-responded; three-fourths of the 143 who lived with at least one other eligible-to-be-interviewed adult participated in a group interview. Record use levels far exceeded expectations -- for example, respondents used at least one record to substantiate payment date and amount information for over 70% of all income sources reported, compared to about 20% in standard SIPP (Singh, 1991; Singh, 1992). Respondents' record use increased in Wave 2. Data quality, as indicated by a reduced seam bias and, in Pretest 2, reduced underreporting errors, also appeared to be improved[9].

Not all indicators were positive, however. Our first attempts at implementing a performance quality monitoring system were clearly flawed -- feedback was too slow, for example, and was perceived by interviewers to be focused on negative feedback almost exclusively. Although the tests did not use an experimental design, and thus did not offer any direct means of comparison, the combined pretest household response rates, 73% in Wave 1 and 87% in Wave 2, were substantially lower than those typically achieved by standard SIPP. Per case costs were perhaps 50% higher. We looked for direct evidence implicating the new procedures as the cause of the nonresponse and cost increases; what we found, in interviewers' descriptions of their noninterviews and in their reports of all of their visits to sample households, suggested a small contribution of the new procedures to these negative outcomes, not nearly sufficient to fully explain the differences.

In Moore, Bogen, and Marquis (1993), we summarize the pretest results as follows:

> "[I]ndications from small-scale pretests are that the new procedures have the potential to substantially reduce some of the survey's important measurement problems. At the same time, the operational difficulties encountered in the pretests -- high nonresponse and high costs -- clearly put at risk the notion that they are a viable option for national, production implementation." [p. 39]

4.0 THE EVALUATION EXPERIMENT DESIGN

The pretest results suggested that, while important operational questions remained, the experimental procedures were moving in the right direction with regard to improving the quality of key SIPP measurements. We designed the SIPP Cognitive Research Evaluation Experiment to provide clear, statistical evidence of the data quality effects of the new procedures. However, because of the operational questions, we defined this test as a necessary, but by no means sufficient, step on the path toward implementation of the new procedures in a production SIPP.

4.1 The Sample and Record Check Evaluation Designs

[9] Small sample sizes and important design differences (the abbreviated Wave 2 reference period, for example, in the case of the seam bias results) make the pretest data quality comparisons to standard SIPP somewhat questionable.

Our objective in designing the new procedures was to substantially improve the quality of SIPP measurement -- specifically, to reduce underreporting of participation in selected major government transfer programs by 25%. Only true program participants can underreport, and true program participants are fairly rare in the general population. Therefore, to approach our objective efficiently, we drew samples of people who we knew were participants in one of four programs at some time during reference period of our Wave 1 interview. The four programs were Aid to Families with Dependent Children (AFDC), Food Stamps (FOOD), Supplemental Security Income (SSI), and Unemployment Insurance (UNEM)[10]. Because income from earnings comprises such a major portion of total income, we also drew a small sample of people who worked for a large employer in the area (JOB), in order to learn about wage and salary reporting errors.

4.1.1 Sample Persons

We designed the Evaluation Experiment to be able to detect a 25% difference between standard SIPP procedures and the experimental procedures. According to the sampling experts we consulted, this required approximately 350 completed Wave 2 interviews per treatment -- 75 from each of the four programs and 50 from the employer -- for a total of 700. The administrative agencies for each of the four transfer programs created our initial sample frame by randomly sampling approximately 600 cases from their active case files of people whose residential ZIP code was within our interviewing area, the city limits of a moderately sized midwestern city. Each agency drew two samples of approximately 300 cases each. The first sample drew from active June cases, the second from cases active in September. This timing ensured that sampled people would have at least one month of true participation in the program during the Wave 1 reference period.

Procedures for the employer sample were somewhat different. The employer provided us with a single data file consisting of all current employees as of June 28, 1992. However, the vast majority of the 6,215 cases on this file were out of scope for our purposes. By the time we eliminated duplicates, cases without a geographically precise home address, addresses outside of the city limits of our test site, certain employee categories, and "employees" with a 0 percent work schedule, the total sample frame was reduced to 695 cases. From this file we drew an initial random sample of 183 employees.

This initial sample frame -- approximately 600 cases from each of four programs -- plus the 183 employer cases, was substantially larger than the 700 interviews we actually needed for the experiment for several reasons. Based on our pretest experience, we included cushions to accommodate the likelihood of nonresponse (both Wave 1 nonresponse and Wave 2 attrition), and the probability of actually finding the sampled person at the address indicated on the administrative record. Confidentiality was a consideration as well; the additional initial

[10] Social Security is the most important government transfer program, in terms of both number of participants and dollars, and thus might be considered a high priority program for study. However, since our earlier research found that it is seldom either underreported or overreported in SIPP (Marquis and Moore, 1990), we did not include it in the experiment.

sample also served to prevent disclosure of the actual final sample to the source agencies.

We eliminated ineligible selections, such as those without a street address on the administrative record. We unduplicated names across income sources, and within each program source across the two half samples. For our final sample, we stratified our frame on program and ZIP code, and selected cases for a Wave 1 interview as shown in Table 1.

Record Source	Initial Sample Frame	Final Sample for Wave 1 Interview		Desired Number of Completed Wave 2 Interviews	
		Exper.	Control	Exper.	Control
AFDC	596	208	207	75	75
FOOD	595	171	174	75	75
SSI	595	196	192	75	75
UNEM	710	146	143	75	75
JOB	695	89	94	50	50
TOTAL		810	810	350	350

<u>Table 1</u> The large initial sample frame protected the confidentiality of the final sample; the final sample included additional cases as a cushion against nonresponse and other sample attrition.

Subsequently, the administrative record sources sent us participation and income amount information for the relevant time period covered by the two interviews of the Evaluation Experiment (see below) for each person originally selected from their records[11]. They included the sample person's social security number (SSN) and name, and often included other identifying information such as date of birth. We used this information to ensure that we matched the survey and administrative record reports correctly. This group of sample cases -- sampled from records and later matched to record information -- is our primary analysis group of interest. We refer to them as the Sample Persons and use them to estimate participation underreports.

[11] We insisted on getting record information for the entire initial frame, which contained a large number of "foil" cases not actually included in our final sample, in order to protect the confidentiality of the final sample. We informed the agencies of this general strategy and its rationale, but did not reveal the extent of our subsampling of the initial frame.

4.1.2 Extra Persons

A second group of people, those who were not themselves Sample Persons but who were interviewed in a household containing a Sample Person, served as the foundation of our analysis group for estimating participation overreporting errors. We refer to this group as Extra Persons. The Extra Persons group for each program also included Sample Persons from other programs -- for example, an SSI Sample Person was an Extra Person for AFDC, Food Stamps, and Unemployment Insurance. We submitted all Extra Persons' SSNs to each source agency, along with a large number of "foil" cases (all original frame people from other agency lists whom we did not select for the final sample) to maintain the confidentiality of the sample. Table 2 shows the number of Extra Persons by program source and treatment.

Program Source	Number of Extra Persons	
	Exper.	Control
AFDC	1047	1263
FOOD	1017	1288
SSI*	1896	2079
UNEM	1202	1301

NOTE: Children can receive SSI benefits, so the Extra Persons group for SSI includes all household members, regardless of age; all other programs include only adults (15+) in the Extra Persons group.

Figure 1 Program agencies also checked their records for participation information for all "Extra Persons."

The success of the overreport analysis in particular, because it involved sending "new" cases to each agency, depended to a great extent on supplying accurate SSNs to the agencies for matching purposes[12]. We took extra pains, therefore, to ensure the completeness and validity of SSNs for the entire Extra Persons group. We sent all Extra Persons' survey-reported SSN, name, age, and sex information to Census Bureau staff at the Social Security Administration (SSA) for verification. If the SSN was missing, the SSA people provided it. If the reported SSN could not be verified, the SSA people provided the correct number. In only a very few cases was a missing number not found, or a not-verified number left without a verified replacement.

We sent the SSN information for the Extra Persons group (and foils) to each agency. The agencies used the SSNs to search their records for program participants. The agency sent us all relevant income information and a small set of person identifiers for all file "hits" -- people who had received income from the agency during the one year period that included the interview reference periods. We examined all such matched cases and determined whether the match was correct. We eliminated a small number of incorrect matches which resulted from submitting multiple or incorrect SSNs. In a later section we use the final set of matched cases from the Extra Persons group to estimate participation overreporting errors.

[12] The access key to the employer's wage and salary file was not SSN, but rather a unique numerical identifier which we did not attempt to collect during the interview. Therefore, it was only possible to carry out the overreport analysis for the transfer programs, and not for the wage and salary reports.

4.2 The Experimental and Data Collection Designs

We randomly assigned sampled addresses to one of the interviewing treatments and conducted one or two interviews with each household. For each interview, the reference or recall period was the previous four calendar months (including, as noted above, and for the experimental treatment only, that portion of the interview month up to the day of the interview). To even out the interviewing workload we followed the standard SIPP procedure in both treatments, dividing the sample into four rotation groups, one of which was interviewed each month for four months. We began interviewing the first rotation group in September, 1992[13]. All interviews were conducted by personal visit at the respondent's residence. Interviewers were blind to the record check aspects of the experiment, nor did they know the name of the sample persons we expected to find at the addresses assigned to them. Figure 1 shows the rotation group design and data collection schedule for the Evaluation Experiment.

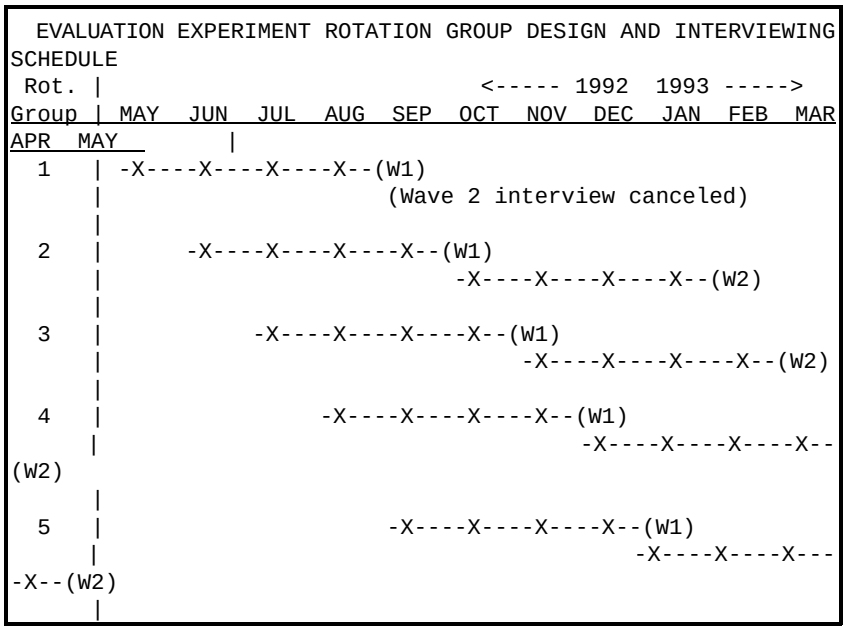

Figure 1 The Evaluation Experiment used the standard SIPP rotation group design, with interviews spaced at 4 month intervals.

We used two separate interviewing staffs. No interviewer worked on both treatments. Although not by design -- our intent was quite the opposite, in fact -- the interviewing staffs differed on a number of character-istics. The experimental treatment used a considerably larger staff than the control treatment -- 15 Wave 1 inter-viewers and 10 in Wave 2, versus 9 and 6 for the con-trol condition. Compared to the standard treatment, the experimental interviewing staff was also less experienced, more racially diverse, and assisted by a less experienced and less productive crew leader. Experimental treatment assignment sizes per interviewer were generally small relative to the control treatment. One experimental treatment interviewer was terminated for fabrication and missing deadlines; the only turnover in the control treatment was voluntary. Later in this paper we attribute the differential response rate and cost results in part to differences between the interviewing staffs and in part

[13] Start-up problems led us to abandon Wave 2 interviews with the September rotation group. Instead, we added a fifth (January) rotation group to each treatment and interviewed it twice. In the analyses we use data from all groups unless otherwise specified (e.g., when we estimate the effects of wave or time).

to the intrinsic features of the experimental and control procedures. Since the experiment confounded treatment and staff characteristics, however, we cannot estimate the separate contributions of each factor.

The interviewers were supervised from the Kansas City Regional Office, several hundred miles from the experiment site. Later in the field period, when it became apparent that the inexperienced experimental treatment interviewers especially needed more direct and immediate supervision, the RO assigned local crew leaders to assist with supervisory tasks in the field.

5.0 EVALUATION EXPERIMENT IMPLEMENTATION RESULTS

This section summarizes operational and monitoring data which address the implementation of the experimental procedures in the Evaluation Experiment.

5.1 Procedure Outcomes

5.1.1 Locating Sample Persons

We learned from Pretest 2 (see above) that the sample person whom we identified through the source administrative record system was often not to be found in the roster of persons living at the address indicated in the records. Among the many beneficial lessons of Pretest 2, none was of more practical importance for purposes of planning the Evaluation Experiment than this. We adjusted for this anticipated sample attrition by increasing our initial sample sizes, to ensure a sufficient number of final cases for our primary underreporting analysis. Table 3 shows the percent "yield" of target sample persons from interviewed Wave 1 households, by record source, for each experimental treatment separately and for the Evaluation Experiment as a whole. For most of our sources, the sample loss due to failure to find the sample person at the address listed in the records was between 20 and 30 percent.

Record Source	Final Sample for Wave 1 Interview		Wave 1 Interviews Completed		Wave 1 Sample Persons Found at Sample Addresses		Total % Found
	Exper.	Control	Exper.	Control	Exper.	Control	
AFDC	208	207	158	186	120	130	73%
FOOD	171	174	134	146	101	102	73%
SSI	196	192	143	170	106	130	75%
UNEM	146	143	108	131	87	101	79%
JOB	89	94	68	78	59	65	85%

Table 3 Household rosters at interviewed sample addresses often failed to include the target sample person.

We would expect some address inaccuracies due to mobility, and perhaps the mobility of a program participant sample might be greater than the population at large. However, we doubt that mobility is the cause of more than a small proportion of the observed attrition, especially given the "freshness" of the sample. In addition, since the addresses we used to try to locate sample persons are the addresses used by the agencies to mail benefit checks to the recipients, one would expect those recipients to be highly motivated to keep their addresses current in the agency's records. Our interest in this phenomenon was largely practical, and once we adjusted plans for the size of our final sample we did not make any attempt to explore its causes. We document these results here primarily because of their potentially biasing impact on our program (and job) participant samples, although with what impact on our estimates it is difficult to know. These results may also serve as a caution to those who would use administrative record data such as these as a substitute for more direct enumeration -- for example, in the decennial census.

5.1.2 Self-Response and Group Interviews

As described above, primarily in the interest of increasing the use of personal records, the experimental procedures strongly emphasized self-response in the Wave 1 interview. In contrast, standard SIPP procedures allow liberal use of proxies in cases where self-response is not immediately available. In Wave 1 we also encouraged a group interview arrangement, whereby respondents could assist each other in completely recalling all sources of income (and which would also provide implicit license for subsequent proxy reporting); under standard SIPP procedures interviewers only conduct (or, are only supposed to conduct) individual interviews. In Wave 2 the experimental procedures focused on trained respondents, self or proxy, using records, and so de-emphasized both self-response and group interviews.

Table 4 shows the extent to which interviewed respondents provided self-response information, by experimental treatment and interview wave. We show results for two different sets of respondents: all respondents regardless of household composition, and respondents in households with two or more interviewed adults, the only households where an option other than individual self-response exists. For the experimental treatment, Table 4 shows separate results for the two possible kinds of self-response -- either in an individual interview or as part of a group interview. Table 4 suggests that experimental treatment interviewers increased the level of self-response compared to the control treatment by about 10 percentage points[14].

[14] We do not have complete confidence in the recording of these data, especially for the experimental treatment. The missing data casts some doubt on the experimental treatment estimates, as does the fact that monitors sometimes found their assessment of response status to be at variance with interviewers' reports. Because of these limitations, we do not attempt any statistical assessment of treatment differences.

Analysis Group	Wave	Number of Interviewed Persons		Response Type by Treatment, as % of Interviewed Persons				
				Experimental			Control	
		Exp.	Cont.	Self-Group	Self-Indiv	Proxy	Self	Proxy
All HH's	1	1134[1]	1365	44	38	18	72	28
	2	653[2]	795[3]	38	39	24	67	33
All 2+ Adult HH's	1	892[4]	1051	54	24	23	65	35
	2	515[5]	623[3]	46	24	30	58	42

NOTE: Tallies of "Interviewed Persons" exclude those for whom "Response Type" is missing; superscripted cell entries exclude the following numbers of missing cases: 1) 79; 2) 70; 3) 1; 4) 75; and 5) 59.

Table 4 The experimental treatment increased the frequency of self-response among interviewed respondents.

Table 4 also provides evidence of the extent to which experimental treatment interviewers implemented the new group interview procedures. About half of all interviewed persons in households with two or more interviewed adults participated in a group interview -- 54% in Wave 1, and 46% in Wave 2. Interviewers conducted a group interview in 58% of eligible Wave 1 households, 54% in Wave 2 (data not shown). As noted, standard SIPP procedures, at least officially, recognize only individual interviewing. Anecdotally, group (or group-like) interviews are not unheard of, but there are no procedures for recording when they occur, and thus there is no control treatment estimate against which to compare the experimental group interview results.

Whether these results indicate a "successful" implementation of the experimental procedures is a somewhat subjective judgment. The self-response differences, even if they proved to be statistically significant, are certainly less substantial than we anticipated given the differential emphasis on this feature in the two treatments, and only about half of experimental treatment respondents eligible to do so participated in a group interview. Perfect, 100% success for self-response and group interviews are not realistic goals, but we suspect that higher rates are achievable; in fact, they were achieved in our pretests (Bogen, Moore, and Marquis, 1992).

5.1.3 Free Recall of Income Sources

We adopted the "free recall" procedures for reporting income in the experimental interview primarily to allow respondents a straightforward and immediate way to report highly salient income sources. Observational evidence (e.g., Marquis, 1990) suggested that SIPP's highly structured approach often serves as a barrier to respondents who are eager to comply with what they perceive to be the main survey task -- reporting their income. Free recall attempted to avoid the unnecessarily slow and painstaking extraction of income information from respondents, while allowing the necessary flexibility to capture the important information from a wide array of income types.

Table 5 shows, for income sources reported by at least 10 people, the sources that respondents reported almost exclusively in the free recall section, and those they reported primarily elsewhere. It suggests that, with one notable exception, this procedure did elicit the most common and highly salient income sources. Pensions (including Social Security), income for the support of indigent families with children, and job income all emerged almost exclusively during free recall. Interestingly, Wave 1 free recall usually did not elicit Food Stamps reports, an important source of income for a large proportion of our sample. The free recall introduction refers to "income, pay, and other money," which apparently is not an effective recall cue for this non-cash income source. (In Wave 2 free recall did elicit a majority of Food Stamps reports, although the proportion reported in other sections of the interview was still substantially higher than for other comparably important sources.) Respondents generally reported rare and irregular income sources, and asset income, which for most people supplies only a small, largely unnoticed income stream, only under direct questioning, in the recognition section.

It does appear, however, that free recall was a learned behavior for respondents. The observed free recall rates for four of the six high frequency sources, and all ten of the low frequency sources, were higher in Wave 2 than in Wave 1. According to a simple sign test (Snedecor and Cochran, 1967), the wave effect across the 16 income types in Table 5 is highly significant ($p<.01$). (Because the data in Table 5 represent all respondents, and because the Wave 1 and Wave 2 results derive from different sets of respondents, we cannot rule out the possibility of sample differences confounding the wave effect. However, a separate analysis limited to only households interviewed in both waves (not shown) yields very similar results.)

5.1.4 Record Use A central feature -- the central feature -- of the experimental procedures was the emphasis on the use of personal records to assist the accurate reporting of income sources and details. Table 6 summarizes record use in the experiment, under varying definitions, by experimental treatment and interview wave.

For All Income Sources (n≥10), Percent Reported During Free Recall					
A. Income Sources Reported Almost Exclusively in Free Recall			B. Income Sources Reported Mostly Outside of Free Recall		
Source	Wave 1	Wave 2	Source	Wave 1	Wave 2
Social Security (n1=211 n2=118)	94%	93%	Food Stamps (n1=324 n2=193)	37%	64%
Federal SSI (n1=182 n2=122)	96%	96%	Money from Relatives (n1=19 n2=14)	16%	29%
State SSI (n1=66 n2=28)	88%	96%	Lump Sums (n1=54 n2=32)	24%	41%
AFDC (n1=260 n2=150)	93%	96%	Incidental Earnings (n1=71 n2=33)	37%	79%
Pensions (n1=27 n2=15)	81%	93%	Energy Assistance (n1=34 n2=25)	9%	20%
Job Income (n1=563 n2=345)	92%	93%	Savings Acct Interest (n1=215 n2=176)	22%	41%
			Money Market Acct Interest (n1=20 n2=16)	35%	81%
			CD Interest (n1=61 n2=54)	48%	50%
			Checking Acct Interest (n1=75 n2=56)	23%	38%
			Stock Dividends (n1=59 n2=57)	34%	54%

Table 5 Free recall procedures in the experimental treatment elicited income reports for the most common and salient income types.

The household-level and income source-level estimates of record use for the experimental treatment are somewhat uncertain because those procedures captured record use both for each income source as a whole (which is comparable to the control treatment procedures) and for each individual payment received from each source. Basing record use estimates on these two methods leads to occasional minor discrepancies, as indicated by the range of values in Table 6. By any measure, however, the results demonstrate clear success

for the implementation of the experimental procedures. There was some use of records in approximately 70% of Wave 1 experimental households, significantly higher than the 25% rate among control households. The use of records in experimental households increased in Wave 2; in the control treatment it remained stable. Similar results are evident in an analysis of individual income sources[15]. And, at the finest level of analysis, records corroborated 39% of the individual payments reported in Wave 1 and 63% of those reported in Wave 2. (No comparable data are available for the control treatment, since standard SIPP collects only monthly aggregate amounts, not individual payments.) Even despite some implementation shortcomings[16], the experimental treatment results far exceed common expectations about the level of record use that is possible in a household income survey.

Wave	Level of Analysis	Record Use Rates by Treatment			
		Exper.		Control	
		n	%	n	%
1	households	611	71-74%	711	25%
	income sources	2343	49-51%	3004	12%
	payments	12,384	39%	--	--
2	households	366	84-87%	404	22%
	income sources	1481	69-70%	1716	11%
	payments	7749	63%	--	--

NOTE: Table entries exclude cases for which record use information was missing.

Table 3 The experimental treatment obtained extremely high rates of record use in the Wave 1 interview, which increased even more in Wave 2.

[15] The data shown in Table 6 are based on interviewers' reports. Validation (of a sort) is possible for monitored experimental treatment cases. Although monitoring usually confirmed experimental treatment interviewers' reports, record use estimates based on monitoring are generally slightly lower than those shown. The record use results are presented in more detail in Bogen, Moore, and Marquis, 1994, which also includes a summary of the statistical tests associated with the effects described here. In addition to the results across all income sources, Marquis (1995) demonstrates that the overall patterns also hold for the specific programs of interest to the Evaluation Experiment.

[16] For example, although almost all respondents granted permission for the interviewer to call back before the Wave 2 interview as a reminder to save records, interviewers seldom made the reminder calls.

5.1.5 Tape Recording

As described in Section 2, above, another new feature of the experimental procedures was an interviewer performance evaluation and feedback system which attempted to shift to a much greater emphasis on quality interviewing behaviors, and a reduced emphasis on traditional indicators such as response rates and productivity. This system involved tape recording interviews in the field, monitoring a sample of the taped interviews on a set of pre-specified, objective behavioral measures[17], and feeding back the monitoring results to interviewers in a timely manner.

The implementation of this feedback system was far from perfect, especially at the beginning of the field period. The logistics of controlling incoming tapes, assigning and completing monitoring work, and delivering clear and timely feedback were difficult to bring under control. We were certainly not aided by the fact that we were trying to impose a fundamental, systemic change within the context of a relatively small, short-term study in one geographic area managed by an overburdened supervisory staff with many competing responsibilities.

Interviewers saw some benefits of the monitoring system -- more than one interviewer told us that the monitoring form provided a crystal-clear indication of what was of primary importance in the new procedures. However, we were never fully successful in selling monitoring to interviewers as a positive feature of the procedures designed to assist them in improving their performance. They saw it instead as burdensome and overly critical.

In one important respect, however, the actual tape recording of the experimental treatment interviews, the monitoring system was highly successful. The procedures called for the taping of every completed interview. We did not retain data from the field on the taping outcome for every completed case; instead we use the sample of cases selected for monitoring to produce a reasonable estimate. Using Wave 1 data only, and based on the monitored sample, experimental treatment interviewers tape recorded 92% of their completed interviews[18].

---

[17] The monitoring forms for the Wave 1 and Wave 2 experimental treatment interviews are included in Appendix D.

[18] This figure overestimates the taping success rate for all completed experimental treatment interviews to an unknown but undoubtedly small extent, because it does not include initial refusal cases re-assigned to supervisory field staff for conversion. Anecdotal evidence suggests that those attempting refusal conversion were rarely successful at taping converted interviews -- often because they simply abandoned any attempt to do so. The monitoring/performance evaluation/feedback system focused on the regular field staff and did not include cases completed by supervisory interviewers.

5.2 Behavior in the Experimental Interview -- Evidence from Monitoring

The performance evaluation system for the experimental treatment consisted of regular feedback to interviewers of the results of a systematic monitoring by KCRO staff of a sample of their tape-recorded interviews. This system produced monitoring data for 189 Wave 1 and 131 Wave 2 interviews, data which also shed some light on how interviewers administered the experimental interview.

5.2.1 Introducing the Free Recall Section

Table 7 summarizes interviewers' Wave 1 and Wave 2 behaviors, as assessed by the monitor, during their introduction to and explanation of the free recall procedures. In general, interviewers carried out this task according to their instructions. In fact, in Wave 1 fully 79% of the monitored interviews received a "perfect" score on all nine items. Only three of the Wave 1 items failed to achieve a 90% compliance rate, although it is the case that two of these items (stating the importance of accu-racy and the need for the use of records for every income source) represent perhaps the most central components of the experimental procedures. Even these "failures," however, were still successfully administered in over 85% of monitored Wave 1 interviews.

Monitoring of the Free Recall Introduction/Explanation	% "Yes" by Wave	
Did the FR correctly...	1	2
...state purpose?	99%	99%
...state section goal?	98%	99%
...name people?	90%	76%
...show/mention a worksheet?	99%	88%
...describe information needed?	94%	84%
...describe reference period?	89%	95%
...state that accuracy is important?	87%	70%
...mention record use for each income source?	87%	--
...show/mention the calendar?	94%	82%
...make appropriate arrangements (if R didn't have records ready)?	--	62%
all items: % "yes" to all 9 items	79%	48%
8 common items: % "yes" to all 8 items	80%	50%
% "yes" to 6 or more items	92%	82%
Mean # of "yes" items	7.5	6.9

Table 4 Experimental treatment interviewers were generally highly successful in their introduction of the "free recall" procedures, especially in Wave 1.

Although compliance with Wave 2 free recall procedures remained quite high, Table 7 offers some evidence of slip-page compared to Wave 1[19]. Here only three items exceeded a 90% compliance rate, and only 48% of monitored Wave 2 interviews received a perfect score. It is again noteworthy that the two lowest scores in Wave 2 were on the assess-ments of interviewers' delivery of the "accuracy is important" message, and on their actions in response to respondent's failure to have records ready for the Wave 2 interview. The drop in Wave 2 performance is clear in a comparison which combines the eight common items on the Wave 1 and Wave 2 monitoring forms, although even the lower quality Wave 2 interviews still performed appropriately on an average of about seven of the eight items common to the two waves.

In sum, these data suggest that interviewers in large measure introduced the free recall section to respondents as we intended, especially in Wave 1. However, the results also suggest some reluctance on the part of the interviewers to fully commit to confronting respondents with the most important information about the new procedures -- that accuracy was paramount, and that personal records were to be used to report income details.

5.2.2 Administering the Recognition Section

The experimental interview included a series of "recognition" tasks in order to ensure the complete reporting of all income sources. This part of the interview followed more standard, scripted interviewing procedures. Interviewers were to read the introduction to the section and each of several income category descriptions or "stems" (e.g.,"Did (you/anyone) get any money because (they/you) were unable to work, such as from ..."); read all income source examples ("items") grouped under each stem (e.g., "unemployment compensation, workers' compensation, temporary sickness benefits, black lung benefits, veterans' compensation, government disability pension, Social Security disability, or any other kind of disability payments?"); acknowledge any prior report of income from a source noted in the recognition lists; and, for all reported sources, probe for "any other" income of the type already reported. The monitor also rated interviewers' pace in reading the recognition lists, and also assessed their performance on the final task in this section of the interview --

[19] Because of the substantial attrition after Wave 1, differences between interviewers' Wave 1 and Wave 2 behaviors are confounded with possible sample differences. In this case, however, a separate analysis of Wave 1 monitoring results restricted to cases subsequently interviewed in Wave 2 yields results virtually identical to those shown in Table 7 (data not shown).

identifying any adult with no reported sources of income, pointing out that fact to the respondent(s), and probing for possible missed income sources for that person.

Table 8 summarizes the monitoring results for this section of the experi-mental interview. Clearly, interview-ers had little difficulty administering the recognition section according to instructions. On all four of the basic tasks, and in both waves, the average compliance level was in the mid- to upper-90% range. Perfect scores were the norm; in the worst instance, "only" 80% of monitored Wave 1 interviews achieved a perfect score with regard to reading all of the recognition section stems. Interviewers did not rush through the recognition list task, according to the monitor. However, they did fail to probe for possible missed income sources in about one-third or more of the cases where it would have been appropriate to do so.

Monitoring of the Recognition Section	Wave	
Did the FR correctly...	1	2
...read stem?		
average % "yes"	96%	97%
% "yes" to all	80%	86%
...read items?		
average % "yes"	94%	97%
% "yes" to all	85%	95%
...acknowledge prior information?		
average % "yes"	94%	98%
% "yes" to all	90%	96%
...ask "Any other?"		
average % "yes"	97%	100%
% "yes" to all	97%	100%
Reading pace: "About right"	85%	92%
("Very fast")	(5%)	(2%)
Did the interviewer identify "no income" adults and probe for missed sources? (% "yes" (n))	68% (47)	58% (31)

Table 5 Experimental treatment interviewers administered the "recognition" procedures well.

Two specific concerns about the admin-istration of this section failed to materialize. First, we had wondered whether interviewers' concerns about repeatedly "badgering" respondents with the recognition lists would cause their performance to deteriorate after Wave 1. There is, however, no evidence in Table 8 of a decline in Wave 2 performance -- the general trend, in fact, is in the opposite direction[20]. The second concern was about the impact of interview type -- group versus individual -- on carrying out the recognition procedures. Would individual interviews suffer by comparison with group interviews, due to interviewers' fatigue at having to repeat the section multiple times in the same household? This concern, too, was apparently groundless; the rates of compliance for the various aspects of administration of the recognition section were virtually identical for individual interviews conducted in multi-adult households and for group interviews (data not

[20] To control for possible attrition effects on the Wave 1-Wave 2 comparison, we conducted a separate analysis of Wave 1 results restricted to cases subsequently interviewed in Wave 2. This analysis brings the Wave 1 results somewhat more in line with Wave 2, but there is still no suggestion that interviewers carried out the recognition procedures less effectively in Wave 2 than in Wave 1 (data not shown).

shown).

5.2.3 Wave 1 End-of-Interview Procedures

Interviewers had several important tasks to accomplish at the end of the Wave 1 interview with regard to respondents' income records. First, where there were any missing records in the current interview, interviewers needed to review whatever arrangements they had made concerning the retrieval of those records. They also needed to review procedures for maintaining records for the next interview, for both presumably continuing income already reported in Wave 1, and for any new income that might enter the picture in the future. Finally, they needed to obtain agreement from all respondents who were not current record-keepers to keep records for the next interview and to accept a between-wave telephone call from the interviewer reminding them to do so.

Table 9 summarizes the monitoring results for these behaviors. According to the monitor, interviewers actually performed quite well with regard to reviewing arrangements for retrieving missing records; where there were such arrangements to be reviewed, they did so in about 9-out-of-10 cases. On the other end-of-interview behaviors, however, interviewers fared less well. Only about two-thirds reviewed the source-specific record keeping instruc-tions for Wave 1 income sources; only about half reviewed the general proce-dures for keeping records associated with any new income sources; and the majority failed to press respondents to start to keep records, and to accept a reminder call to assist them in main-taining this new behavior pattern. (Our data are sketchy at best, but, as noted earlier, we have very little evidence that interviewers actually made any reminder telephone calls -- if they did so, it was certainly at a rate far below even the 40% level implied by the monitoring results.)

Monitoring of the Wave 1 End-of-Interview Procedures Did the FR explicitly...	% "Yes" (n)
...review callback arrangements for missing records?	89% (65)
...review source-specific record keeping instructions?	65% (155)
...give record keeping instructions for future (new) income?	55% (173)
...request agreement to keep records for Wave 2 and to accept a reminder phone call?	42% (140)

Table 6 Experimental treatment interviewers were less successful in their administration of the Wave 1 end-of-interview record-related procedures.

5.2.4 Wave 2 "Last Wave Review" Procedures

As described above in Section 2.2.3, we instituted special procedures in Wave 2 to try to reduce spurious change in patterns of income receipt at the interview "seam." After completing the standard income reporting sections of the Wave 2 interview, interviewers matched the Wave 1 and Wave 2 income reports for the household and pointed out to the

respondent all "Wave 1 only" and "Wave 2 only" sources -- sources not reported in both interviews. The interviewers either verified that this was correct or recorded the necessary details for the missing source. For "both waves" income sources, the interviewer examined the overlap period reports in each wave and resolved any inconsistencies with the respondent.

Table 10 summarizes the monitoring results for this section of the Wave 2 interview. Surprisingly, despite the complexity of these procedures, and the substantial paper-shuffling they required, interviewers appear to have carried them out quite effectively, according to the monitor. The basic process of sorting worksheets into three categories -- "Wave 1 Only," "Wave 2 Only," and "Both Waves" -- was vir-tually never a problem in a monitored interview. Where probes or repair actions were required of the inter-viewers, the monitor judged them to have been correctly accomplished about 80% of the time. The vast majority of monitored interviews (82%) revealed no problem on any of the necessary "Last Wave Review" actions required of the interviewer. In Section 6.3.2 of this report we summarize findings which suggest that the experimental procedures did not yield the expected reduction in seam bias estimates. Although we are unsure why the proce-dures failed to reduce the seam bias, the monitoring results suggest that the failure was not due to interviewers' problems in implementing a too-difficult set of field procedures.

Monitoring of the Wave 2 "Last Wave Review" Procedures Did the FR...	% (n)
...have any problems classifying worksheets (Wave 1 Only, Wave 2 Only, Both Waves)? *(% "no")*	98% (130)
<u>For "Wave 1/2 Only" worksheets:</u> ...ask about receipt in the "other" wave? *(% "yes")*	84% (77)
...have any problems taking corrective action (if corrections were needed)? *(% "no")*	81% (62)
<u>For "Both Waves" worksheets:</u> ...did the FR have any problems identifying or resolving overlap period discrepancies (if there were any to resolve)? *(% "no")*	77% (65)
Percentage of Wave 2 interviews with no problems on any "Last Wave Review" monitoring item:	82% (130)

<u>Table 7</u> Experimental treatment interviewers were surprisingly successful in their administration of the complicated "Last Wave Review" procedures.

5.3 Field Outcomes

From an operational standpoint, the Evaluation Experiment confirmed the positive results of the pretests in demonstrating that the experimental procedures succeeded in many important respects -- for example, interviewers increased the frequency of self-response and persuaded many people to respond in a group setting; and respondents used personal records to a much greater extent than many believed possible. Unfortunately, the experiment also confirmed the primary negative outcomes. As in the pretests, nonresponse rates and costs per

interview were very high. In this case the experimental design permits a direct comparison to standard SIPP procedures.

5.3.1 Response Rates

Table 11 shows household response rates, by interview wave, for the control and experimental treatments. We calculate the household response rate as the number of completed household interviews divided by the number of eligible households in the sample during that wave. (Because the experiment did not include special followup procedures for difficult-to-locate mover households, we exclude Type D (mover) nonrespondents from the base of eligible households.) Control group response rates were higher than experimental group response rates in both waves; according to the Kansas City Regional Office, the control group rates were also slightly higher than for regular production SIPP sample cases in the same area during the same time period.

Wave	Household Response Rates	
	Exper.	Control
Wave 1	82% (n=749)	94% (n=753)
Wave 2	90% (n=410)	98% (n=418)
(longitudinal)	73%	92%

NOTE: Numbers of cases (in parentheses) indicate the number of eligible households out of all addresses assigned for interview.

Table 8 Compared to the control treatment, the experimental treatment experienced much higher rates of nonresponse.

The "longitudinal" response rate, the product of the Wave 1 and Wave 2 rates, estimates the proportion of eligible Wave 1 households interviewed in both waves. (Because of the deliberate sample reductions for Wave 2, the actual longitudinal rate is difficult to calculate precisely.) According to this estimate, the control group lost 8 percent of its Wave-1-eligible households to original nonresponse and subsequent attrition; the experimental group lost 27%. Even in the absence of agreed-upon standards, a 27% loss after only two interview waves is clearly unacceptably high for a production survey.

5.3.2 Costs

The experimental procedures also cost more to conduct than the standard control procedures. The cost-per-case data in Table 12 include hourly pay to interviewers (for both interviewing time and travel time) and reimbursement for automobile mileage associated with completing their interviewing assignments. Costs per assigned case were at least twice as high in the experimental treatment as in

Wave	Per-Case Costs	
	Exper.	Control
Wave 1	$51	$24
Wave 2	$49	$18

Table 9 Costs in the experimental treatment were much higher than in the control treatment.

the standard SIPP control treatment.

5.3.3 What Caused the High Nonresponse and High Costs?

We cannot with certainty pinpoint the exact causes of the nonresponse and cost problems that the experimental procedures have consistently experienced. To some extent, no doubt, the procedures themselves are at fault. Certainly, one cost factor is that the experimental interviews themselves took longer than the control interviews -- about 1½ hours per household in Wave 1, versus 1 hour for the control -- since respondents needed to retrieve personal records and also had to report their exact, to-the-penny income, individual-payment-by-individual-payment, rather than in monthly aggregates. More important, perhaps, was that in order to improve measurement quality, many of our procedures deliberately deemphasized high productivity -- for example, the insistence on an interview setting conducive to high quality reporting (no distractions, self-response, group interviews, complete records, etc.) almost necessarily resulted in additional callbacks to the household and thus additional time.

Wave	Household Interview Status	Mean Number of Personal Visits Per Household			
		Visits to Try to Establish INITIAL Contact with the Household		Visits to Try to Complete a Case AFTER Initial Household Contact	
		Exper.	Control	Exper.	Control
1	all interviewed households	2.2 (n=595)	1.8 (n=689)	1.5	0.7
	all non-interviewed households	3.4 (n=187)	2.6 (n=94)	3.0	1.1
2	all interviewed households	0.9 (n=365)	0.6 (n=375)	1.5	0.6
	all non-interviewed households	1.4 (n=47)	1.7 (n=13)	3.0	1.5

NOTE: Numbers of cases (in parentheses) indicate the total number for which the record of visits information was not missing. The missing data rate exceeded 10% only for the Wave 2 "non-interviewed households" cells, where about 20% of experimental treatment cases and over 40% of the control cases were missing these data.

Table 10 Both before household contact, and especially after initial contact, experimental treatment interviewers made many more visits to sample households to try to complete their assigned cases.

Interviewers in both treatments kept a detailed record of all of their visits and calls to addresses in their sample assignments. Table 13 summarizes the personal visit results (i.e., ignoring telephone calls) derived from these records. Any effects of the experimental procedures on the number of contacts required to complete a case would only come into play after contact was made with the household. Table 13 shows the expected results; experimental treatment field staff made at least twice as many "after initial contact" visits to try to complete an interview as did those in the control treatment.

Regional Office supervisory staff hypothesized that the procedures might have contributed to higher nonresponse in two ways. First, we occasionally gave up "bird-in-the-hand" interviews by not simply getting whatever information was possible from whoever was available at the initial contact with an eligible respondent -- with the inevitable result that the potentially higher quality "two-in-the-bush" interview didn't always materialize. Each additional callback carries a certain risk of nonresponse; by increasing the number of visits some households required we inevitably reduced survey participation. The second factor was "lack of negotiating room." Our unwillingness to allow interviewers to compromise on the basic, quality-oriented procedures left them with very little to bargain with in trying to convince initially reluctant respondents to participate. Unlike standard SIPP, they couldn't agree to do the interview on the front stoop, or rush through it, or short circuit the procedures designed to ensure accurate income reporting.

There is undoubtedly some truth in these ideas. But it is also the case that the interviewers using the experimental procedures were very inexperienced, and Field Division data (Beach, 1991) suggest that response rates for experienced interviewers are typically higher than the rates achieved by inexperienced interviewers, and that this difference increases in more difficult to interview, highly urbanized areas. The experimental staff differed from the control treatment staff in other ways as well, as summarized earlier. In addition, experimental treatment interviewers were often given inefficiently small assignment sizes. These differences in interviewer characteristics and assignment sizes may help explain the fact that, as shown in Table 13, under most conditions, experimental interviewers also tended to make more personal visits to sample addresses before making successful contact with an eligible potential respondent. Inexperience may have resulted in more unproductive calls at non-optimal times; small assignments may have reduced the pressure to work with maximum efficiency. So, while the experimental procedures themselves probably drove costs up and response rates down, the confounding of experimental treatment and interviewer characteristics makes it impossible to determine exactly how much each contributed to the operational difficulties.

6.0 EVALUATION EXPERIMENT SUBSTANTIVE RESULTS

The primary purpose of the Evaluation Experiment was to provide defensible statistical evidence of the data quality effects of the experimental interviewing procedures. This section summarizes the comparisons of the two interviewing treatments on several substantive outcome variables having to do with key SIPP measurement issues -- most

importantly, participation underreporting and overreporting, program participation transitions, especially at the interview seam, and income amounts reporting. In general, the results of the experiment are disappointing. They do not show important treatment differences in the underreporting of program participation. Generally, both treatments produced about the same rates of overreport errors as well. We do find differences in the pattern of errors regarding transitions in program participation status, although these differences send mixed signals as to whether the experimental procedures produced markedly better performance than the control treatment. The experimental procedures did produce more accurate reporting of income amounts in the second interview of the panel, and we associate this effect with the increased use of records in the experimental interview. However, we judge these minimal positive benefits to measurement quality to be considerably outweighed by the negative operational aspects of the experimental procedures -- their greatly increased costs and reduced response rates.

6.1 Program Participation Underreporting

The primary goal for the experimental procedures was to substantially reduce the underreporting of program participation in SIPP -- by at least 25%. We used the Sample Persons analysis group (see section 4.1.1), the people sampled directly from administrative records, to make the participation underreporting estimates. For that group of people, we examined all months of "true" participation, according to the administrative records. We considered all survey reports of participation in true participation months to be correct reports, and all failures to report participation in true participation months to be underreports. We averaged underreports over months and people to obtain the underreport rates or percentages used in the analyses.

6.1.1 Underreporting Differences in the Experiment

We expected a large reduction in underreporting errors under the experimental interviewing procedures; in fact, the actual results show essentially no difference in participation underreporting between the Evaluation Experiment treatments. Using the AFDC results as an example, on average, experimental treatment respondents failed to report 12% of their true months of participa-tion; the control group did not report 10% of their true participation months. None of the differences in Table 14 is statistically significant, indicating that the experimental

Program/ Income Source	Average Participation Underreporting Percentage (Both Waves)			
	Exper.		Control	
	n	%	n	%
AFDC	186	12%	194	10%
FOOD	214	17%	219	12%
SSI	109	13%	127	8%
UNEM	68	41%	85	44%
JOB	64	11%	66	4%

NOTE: N indicates the number of Sample Persons with true participation in any month of either the Wave 1 or the Wave 2 reference period.

Table 11 Compared to the control treatment, the experimental treatment did not reduce underreporting of program participation.

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 in the opposite direction.

These results offer one important piece of good news regarding the reporting of job "participation." Of course, we have a small sample of cases from only one employer, but these limited results suggest that job underreporting is only a minor problem in SIPP. 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) from surveys other than SIPP.

6.1.2 Underreporting Already Low?

Table 15 shows that, for three of the four pro-grams for which there are comparable data, control group underreport rates were substantially below those obtained by stan-dard SIPP interviewing procedures in the 1984 SIPP Record Check Study. We had expected the rates to be about the same. One possible explanation for the difference is that SIPP's error levels have been reduced in the intervening years -- perhaps much of the error we sought to eliminate by using new procedures had already been eliminated by other events or for other reasons.

Program/ Income Source	Average Participation Underreporting Percentage (Collapsed Across Waves)	
	1984 SIPP Record Check Study	Evaluation Experiment Control Treatment
AFDC	25%	10%
FOOD	24%	12%
SSI	23%	8%
UNEM	39%	44%
JOB	N/A	4%

Table 12 The control treatment results show substantially reduced levels of participation underreporting compared to the 1984 SIPP Record Check Study.

The populations in the two studies were very different, however, making a direct com-parison quite difficult. The 1984 study was based on cross section samples in four states. The current study used a sample drawn from administrative records in one largely inner-city urban area. One might suspect that the effects of such sample dif-ferences would lead to elevated estimates of underreporting in the Evaluation Experiment, opposite to what we observe in Table 15. But any assumptions of this type must be tempered by acknowledgement of the major attrition from the sample which resulted from the failure to locate many of the target sample persons at the addresses provided by the source agencies. We can only conjecture about its effects, but it is possible, for example, that this sample loss differentially eliminated dishonest reporters, leaving a higher proportion of people with less tendency to underreport. Certainly the interviewers who conducted the 1984 SIPP Panel interviews differed from those who conducted the control treatment interviews; another possible explanation is that the control treatment interviewing staff was exceptional along some key dimensions associated with the accuracy of respondents' reports. In the end,

the causes of the reduction in underreporting levels from what we observed earlier remain unclear; what is clear is that our experimental procedures were directed at reducing a much smaller problem than we had anticipated.

6.1.3 Most Underreports Result from Omitting the Entire Source

What is the nature of respondents' underreporting errors? Do they tend to make occasional errors by underreporting some months of income from an otherwise reported income source, or do they underreport the whole source? Table 16 addresses these questions. It shows the proportion of observed under-reported months of participation that stems from failing to report the income source at all. As shown in Table 16, for both treatments, and for all of the four income transfer programs included in the experiment, most underreporting occurred because the respondent never mentioned the income source.

Program/ Income Source	Percent of Participation Underreporting Attributable to Failure to Report the Income Source at All			
	Exper.		Control	
	n	%	n	%
AFDC	108	58%	94	81%
FOOD	176	59%	137	66%
SSI	72	89%	75	84%
UNEM	53	68%	83	63%
JOB	31	32%	15	0%

NOTE: N indicates the number of Sample Persons who underreported true participation in any month of either the Wave 1 or the Wave 2 reference period.

Table 13 Most underreporting of participation was due to failing to report the source at all.

This finding is particularly important with regard to the primary feature of our experi-mental procedures -- getting respondents to use their income records. If underreporting is due to the failure of the entire source of income to surface during the interview, then getting people to use their personal income records is not the right solution. Records are of use after a source has been reported; they are not going to improve respondents' ability to remember income sources that they have forgotten about, nor are they likely to increase respondents' willingness to report income sources that they have decided not to report.

6.1.4 Why is UNEM Underreported So Frequently?

As Table 14 makes quite clear, the UNEM program underreporting rate is an obvious outlier, exceeding the rates for the other sources by a factor of 3 or 4 or more. (A similar difference, though less extreme, can be seen in the earlier SIPP Record Check results in Table 15.) There appear to be additional forces suppressing the reporting of UNEM receipt beyond those that affect the other programs. One possible candidate is the reluctance to reveal UNEM benefits received illegally if, for example, the recipient also receives income from a job.

If this were a factor contributing to the high UNEM underreporting rates, then we might expect to find a higher level of reported receipt of job income among UNEM underreporters than among those who correctly report receipt of UNEM benefits. The data, however, do not support this notion. Table 17 shows, by wave, but collapsed across the two Evaluation Experiment treatments, the reporting of any income from a job or business during the wave among UNEM Sample Persons who were whole-source underreporters and among those who correctly reported at least some of their UNEM receipt. Whole-source underreporters were not overly likely to report job income; in fact, the trend (n.s. for Wave 1; $p<.01$ for Wave 2) is in the opposite direction.

Wave	Percent With Job Income			
	UNEM Whole-Source Underreporters		UNEM Reporters	
	n	%	n	%
1	57	75%	110	82%
2	10	60%	46	96%

Table 14 UNEM whole-source underreporters were not more likely to have job income than UNEM recipients who reported their UNEM receipt.

Thus, there is little in these results to suggest that UNEM recipients are particularly prone to this form of dissembling, leading to markedly higher underreporting rates than the other programs. Perhaps the underreporting difference is traceable more to objective differences between UNEM and the other programs in the nature of the benefits paid. UNEM ben-efits are typically paid weekly; the other programs pay monthly. UNEM is designed to be a very short-term program, and the average "spell" length for UNEM receipt is in fact much shorter than for the other pro-grams (Shea, 1995). (Indeed, the transitory nature of UNEM receipt caused some prob-lems for the analysis of UNEM underreport-ing in the experiment, since the great majority of UNEM Sample Persons actually received benefits in very few reference period months, especially in Wave 2.) Pre-vious research by Vaughan and colleagues (Klein and Vaughan, 1980; Goodreau, Oberheu, and Vaughan, 1984) has suggested that brief spells of AFDC receipt are the most likely to be omitted in survey reports. It is reasonable to speculate that perhaps the same spell-length-based mechanism that affects the quality of reporting within a single program also applies across programs, affecting respondents' overall propensity to underreport a particular income source.

6.1.5 Did the Experimental Procedures Exacerbate Income Source Underreporting?

Observers' reports occasionally suggested that the group interview component of the experimental procedures might be a cause of missed income sources, whether through interviewer inadvertence (e.g., failing to record all income sources named by group interview respondents during a particularly lively free recall session) or respondent reluctance to discuss

certain income sources in the presence of others. We examine this question in two different ways, but find no evidence in the data to support this concern.

Table 18 addresses the possible negative impact of group interviews by comparing the average number of income sources reported by experimental treatment self-respondents in multi-adult households interviewed individually versus those interviewed in a group. These data offer little evidence that group interviews inhibited the complete reporting of income sources. A simple analysis of variance suggests only a significant main effect for interview wave, but no significant effects of type of interview on the number of income sources reported.

Wave	Interview Type	# of People	# of Income Sources	Avg. Sources per Person
1	Self-Indiv	227	486	2.14
	Self-Group	479	1038	2.16
2	Self-Indiv	143	377	2.63
	Self-Group	236	589	2.49

NOTE: This analysis is restricted to respondents with non-missing interview type in multi-adult households.

Table 15 The average number of income sources recalled in group interviews did not differ from the average in individual self-response interviews.

Table 19 addresses the issue of the possible negative effects of experi-mental treatment group interviews from a different perspective, by examining the whole source under-reporting phenomenon for the specific record-checked sources included in the Evaluation Experiment. Once again, there is little in these results to suggest that the group interview sup-pressed the reporting of sources of income, resulting in an increase in whole source underreporting. Analysis of variance tests show no significant effects of interview type for any program.

6.2 Overreporting Errors

The other kind of error respondents can make is to overreport participa-tion in an income program. Survey evaluators have often overlooked overreporting errors, perhaps because of their low incidence, or perhaps because models of memory decay generally deal only with under-reporting. Nevertheless, overreports can be an important component of the quality of survey estimates. They may also present a difficult dilemma for survey designers, since the design remedies for overreporting problems may be of a very different nature than those which attempt to ameliorate underreporting.

Overreport rates are usually low relative to underreport rates. The low rates can be deceptive, however, because their effects on the bias in a survey estimate depend partly on the incidence or prevalence of what is being overreported. Consider the

Program/ Income Source	Wave	Rate of Whole Source Underreporting by Interview Type (Individual vs. Group)			
		Individual Interview		Group Interview	
		n	Percent	n	Percent
AFDC	1	38	13.2%	56	5.4%
	2	29	10.3%	28	10.7%
FOOD	1	42	11.9%	58	17.2%
	2	29	3.5%	29	17.2%
SSI	1	16	25.0%	33	9.1%
	2	11	9.1%	19	5.3%
UNEM	1	14	35.7%	25	40.0%
	2	9	11.1%	8	0%
JOB	1	14	0%	29	6.9%
	2	11	0%	23	0%

NOTE: This analysis is restricted to interviewed experimental treatment Sample Persons, with true program/job participation and non-missing interview type, in multi-adult households.

Table 16 Whole-source underreporting in the experimental treatment was not related to the type of self-response interview.

Survey Report	Administrative Record Value		Total
	Yes	No	
Yes	1363	90	1453
No	76	6021	6097
Total	1439	6111	7550

Table 17: **A small overreport rate can overwhelm a substantially larger underreport rate, resulting in a positive net bias in a survey estimate.**

example presented in Table 20, which uses one typical month of OASDI ("social security") data obtained in the SIPP Record Check study (Marquis and Moore, 1990; see Appendix Table 1). A sample of 7550 people consists largely of true non-participators (n=6111). Ninety true non-participators overreport, for an overreport error rate of 1.5%. The underreport rate among the 1439 true participators is 5.3%, as a result of only 76 underreport errors. Even though the underreport rate is 3½ times the overreport rate, the resulting survey estimate of participation, 19.2% ([1363+90]/7550), is slightly positively biased relative to the true participation rate.

6.2.1 Overreporting Differences in the Experiment

For this analysis we used the Extra Persons analysis group (see section 4.1.2), the interviewed people who entered the sample by virtue of their residence with a person sampled from records, and also including for each income source, as described above, Sample Persons from all of the other sources. (The Extra Persons group is not representative of the general population of program non-participants, and we make no claims that the estimates derived from this group are generalizable. The U.S. population has a much higher incidence of non-participant, middle class households, none of whose members participate in the kinds of government income transfer programs studied here.) Here we examined all months of "true" non-participation -- months identified by the absence of any indication in the administrative records that the respondent had participated in the program in that month. We considered all survey reports of participation in true non-participation months to be overreporting errors, and we calculated the overreport error rate by dividing the overreported months by the total months of true non-participation. For the overreport analysis that follows we averaged these rates over months and people.

The results, in Table 21, show very little effect of the different interview proce-dures on overreporting of program parti-cipation within the Extra Persons group. (As noted earlier, the design of the Eval-uation Experiment did not permit esti-mating overreports of jobs at the participating employer.) There is a modest trend for the experimental treat-ment to get fewer overreports of partici-pation in general; this difference is statistically significant ($p<.05$; 2-tailed, t-test, ignoring any effects of intra-household clustering due to multiple adult respondents per household) only for the Food Stamps (FOOD) program. In the main, however, there appears to have been little or no effect of the new

Program/ Income Source	Average Participation Overreporting Percentage (Both Waves)			
	Exper.		Control	
	n	%	n	%
AFDC	910	3.5%	1057	4.1%
FOOD	849	1.6%	956	3.2%
SSI	979	3.0%	1136	3.4%
UNEM	1088	0.6%	1253	1.0%

NOTE: N indicates the number of Extra Persons with true non-participation in any month of either the Wave 1 or the Wave 2 reference period.

Table 18 The experimental treatment had little or no effect on program participation overreporting.

interviewing procedures on reducing overreporting of program participation.

6.2.2 Overreporting and the "Free Recall" of Income Sources

Several factors motivated our decision to include "free recall" procedures for reporting income in the experimental interview -- for example, we wanted to allow a straightforward and immediate vehicle for reporting high salience sources that do not need to be painstakingly extracted from respondents, and we wanted an extremely flexible system so that respondents could find ways within the general outlines of the reporting task to report their own income in a manner consistent with their own preferences. We also suspected, however, that such freely-recalled sources, while they might not constitute a complete set of all of a respondent's income sources, might be less subject to overreporting errors than sources reported in response to specific cues. Marquis, Marshall, and Oskamp (1972) show, for example, in a legal interrogation setting, that reports obtained via spontaneous narrative testimony are more accurate -- but less complete -- than reports obtained in response to specific, detailed questioning. More recently, Cohen and Java (1995) offer similar results in a study of medical history reporting.

We examined overreporting errors among experimental treatment Extra Persons who had reported participation in one of the four programs of interest, according to whether they reported the income source during free recall or in some other section of the interview. The analysis (not shown) suffers from small n's -- only FOOD was reported with any substantial frequency outside of the free recall section. Nevertheless, there is no indication that income sources reported during free recall were any less subject to overreporting errors than sources reported elsewhere.

6.3 Participation Change Reporting Bias

There is considerable interest in the accuracy of reports of program participation changes -- the beginning and ending of participation "spells." Past investigations, even without the assistance of administrative records, have shown that respondents' reports generally yield higher estimates of participation transitions at the seam between interviews than between months within a single interview's reference period (e.g., Moore and Kasprzyk, 1984; Burkhead and Coder, 1985). Marquis and Moore (1990) suggest that the seam bias is a net result of too many transitions measured at the seam and too few measured elsewhere.

6.3.1 Constructing "Transition Bias" Estimates

In neither the standard nor the experimental SIPP interview do respondents actually report participation transitions; analysts must infer them from participation reports in adjoining months. For this analysis we considered any change in status, regardless of direction (e.g., from receiving to not receiving benefits, or from not receiving to receiving), to be a transition. Our estimate of "transition bias" is admittedly a rough one -- for each respondent we counted the number of survey-derived transitions in each of the five target income sources and subtracted the survey count from the actual number of transitions as

shown in the administrative records. Summing these differences, dividing the sum by the true number of transitions, and multiplying the result by 100, yields a percent transition bias. We combine the results for both Sample Persons and Extra Persons, and for the five income sources, because program participation transitions are rare events[21].

We note that the results to follow may be influenced by special data processing procedures required for the experimental treatment for programs that issue monthly checks. Due to minor vagaries in check-mailing schedules -- for example, if the normal receipt date fell on a weekend or holiday, many programs mailed or distributed their checks early -- the payment-by-payment income reporting method for such programs often resulted in observing two payments in one month and no payments in the next month. Since we asked for exact dates, many experimental treatment respondents reported these irregular receipt dates correctly. We did not ask respondents in the experimental treatment whether they participated in a program in each of the four months of the reference period; instead, we inferred participation from receipt of income. 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 "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, however, 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.

[21] Although producing more easily interpretable results, this approach allows errors to offset each other 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.

6.3.2 Transition Bias Comparisons

The results are summarized in Table 22. The combined data in the bottom row of Table 22 suggest that the experimental treatment produced about the right total number of transitions in program participation (just 8% more than the true number shown in the records) and that the control treatment produced too few (about 26% less than the true number).

The measurement errors for off- and on-seam transitions for each treatment show quite different patterns. The control group results are consistent with what Marquis and Moore (1990) found in the earlier SIPP Record Check Study: a net underreporting of off-seam transitions (within an interview's reference period), and a net overreporting of on-seam transitions. The experimental group results suggest an even greater positive bias for changes on the seam, but virtually no bias for changes at other times. The large positive seam bias result is a surprise because we took elaborate procedural precautions to avoid spurious change at the seam; the small n's for the on-seam results may be misleading us here.

6.4 Errors In Reported Income Amounts

This section examines the effects of the experiment on errors in reporting amounts of income. First we summarize the basic differences between the experimental and control treatments in the quality of respondents' income amount reports, and then we examine the specific role of the use of personal income records. These results are more encouraging since they suggest that the experimental procedures eventually cause an important improvement in reporting quality.

6.4.1 Treatment Differences in the Quality of Income Reports

This analysis also uses the Sample Persons analysis group, with some exclusions. First, the respondent must report -- and the administrative record must agree -- that he or she participated in the program for the given month. Second, to test for time effects, the person

Type of Transition (On/Off Seam)	Number of Participation Transitions and Percent Transition Bias					
	Experimental Treatment			Control Treatment		
	Survey	Record	**Bias**	Survey	Record	**Bias**
Off Seam	142	147	**-3%**	110	169	**-40%**
On Seam	37	18	**+106%**	30	21	**+43%**
TOTAL	179	165	**+8%**	140	190	**-26%**

Table 19 The experimental and control treatments show different patterns of participation transition bias.

must have been interviewed (or interviewed about) in both Wave 1 and Wave 2. For this subset, we compare the reported amount to the amount in the records for each month, averaging over months and people for each program. We consider the reported amount to be correct if it is within 5% of the recorded amount, and we consider "don't know" answers as incorrect.

Table 23 shows, for each of the target income sources, the effects of treatment and interview wave on the proportion of income amounts reported correctly. The basic result is that, over time, reporting improved in the experimental group. In general, the two treatments achieved about the same percentage of correct amount reports in Wave 1. However, by the end of Wave 2 the experimental treatment was usually producing substantially better reports than the control. For the first three income sources, AFDC, FOOD, and SSI, the patterns are remarkably similar: treatment differences at Wave 1 are minimal, but by Wave 2 the experimental treatment elicits a higher percentage of correct reports. For each of these income sources, the treatment-by-wave interaction is significant ($p \leq .05$) in a repeated measures analysis of variance using people who correctly reported their participation in both waves. The UNEM results suggest a similar pattern, although only the main treatment effect is statistically significant.

The percent correct amount reports for JOB income show a different pattern. 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 is statistically significant. Respondents in both treatments improved their reporting over time ($p \leq .05$ for the main effect of wave in the repeated measures ANOVA).

6.4.2 Effects of Record Use on Amounts Reporting

Program/ Income Source	Wave	Number of Matched Survey/Record Cases Interviewed in Both Waves		Percentage of Correct (±5%) Income Amount Reports	
		Exper.	Control	Exper.	Control
AFDC	1 2	114	115	83% 87%	80% 72%
FOOD	1 2	123	130	67% 83%	66% 63%
SSI	1 2	67	77	78% 84%	78% 68%
UNEM	1 2	9	8	29% 61%	20% 19%
JOB	1 2	46	46	67% 77%	52% 76%

NOTE: Entries in the "Number of Matched..." column indicate the number of Sample Persons who were interviewed (or interviewed about) in both waves and who had one or more months in which both survey and record agreed that there was receipt of program/job income.

Table 20 The experimental treatment usually produced better reporting of income amounts by Wave 2.

Earlier we presented results which suggest that respondents in the experimental treatment used personal records far more often than control treatment respondents. Did the increased record use account for the improved reporting of income amounts? Because treatment and record use are highly correlated, this analysis looks just at the effects of record use within the experimental treatment. We treat record use as a simple yes-no variable: the value is "yes" if the respondent used any record in either wave to report the income from the given source. The dependent variable is, again, the percent of income amounts reported correctly in the survey, and the amount is correct if it is within 5% of the true value according to administrative records.

The results are summarized in Table 24. For the three long-term welfare programs, AFDC, FOOD, and SSI, the trend is clearly for some record use to be associated with greater reporting accuracy in both interview waves. For these sources, entries in the "yes" column are consistently higher than those in the "no" column. (The main effect for record use is statistically significant only for AFDC ($p<.05$) and FOOD ($p<.10$).) These data also support the general trend, noted earlier, for experimental treatment respondents to improve their reporting over time; this appears to be the case regardless of whether they used records. (These estimates are based on small numbers of cases; the main effect for wave is statistically significant only for FOOD ($p<.01$) and SSI ($p<.10$).) The remaining two income sources, UNEM and JOB, show mixed pictures. For UNEM, the observed trend is actually for the non-record-users to be slightly better reporters than the record users, although the extremely small n's make any such comparison highly suspect (neither the record use effect nor the apparently much larger effect of interview wave is statistically significant). JOB income appears to show a pattern parallel to the interaction between treatment and wave: reports improved with record use and the difference increased over time. However, neither this interaction nor either of the individual main effects is statistically significant.

Program/ Income Source	Wave	Average Percentage of Correct (±5%) Income Amount Reports (Experimental Treatment Only) by Respondents' Use of Records			
		Used Records		Did NOT Use Records	
		n	%	n	%
AFDC	1 2	94	86% 90%	20	67% 71%
FOOD	1 2	80	69% 87%	43	63% 75%
SSI	1 2	44	81% 86%	23	74% 80%
UNEM	1 2	6	26% 58%	3	33% 67%
JOB	1 2	42	67% 78%	4	65% 63%

NOTE: N indicates the number of Sample Person true partici-pants with a matched survey report, and non-missing record use information, and who were interviewed in both waves.

Table 21 Within the experimental treatment, using records usually produced better reporting of income amounts.

7.0 DISCUSSION

Notwithstanding their apparent benefits to income amount report-ing, the experimental procedures are clearly not the answer to reducing SIPP's most critical response error issues. They failed to produce improved reporting of program income sources and pro-gram participation transitions, and even had they done so their asso-ciated nonresponse and cost problems may well have proved intractable. So then, the question remains: How can we

achieve bet-ter quality SIPP reports?

We find that most underreporting is due to the failure to report entire income sources, as opposed to fail-ure to report all months of partici-pation in an otherwise-reported program. Remedies for the tendency to underreport entire income sources 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. On the other hand, whole source underreports that are intentional or motivated need different remedies that address the privacy and confidentiality concerns that prompt intentional underreporting. We may need to accommodate people who really do not want to discuss their income with other family members by ensuring that the option always exists to conduct truly private interviews. We may also need to find ways to be more persuasive about our ability to maintain absolute confidentiality for anything the respondent reports.

The apparent failure of the experimental procedures to produce better estimates of participation transitions at the seam between SIPP interviews remains a mystery. We designed the experimental treatment interviewing procedures to flag for verification all cases in which a respondent reported an income source in one wave but not the other, and in which anything about his or her participation during the overlap period differed between the two waves' reports. The procedures for these checks and verifications were cumbersome and difficult, especially with a paper-and-pencil mode of survey administration, although our monitoring data suggest that, in general, interviewers carried them out quite effectively.

The major success story for the experimental procedures is record use. We succeeded far beyond expectations in getting respondents to use their personal records. Record use increased the accuracy of reporting income amounts, especially in the second interview. It makes sense that once a respondent acknowledged the existence of an income stream, using personal records had favorable effects on reporting income details, especially after gaining some experience in interpreting the records. However, record use is clearly not the panacea for reducing all response error in SIPP. What it failed to do was to reduce error in reporting program participation, most of which stems from never reporting the program at all as an income source. This, too, makes sense. Record use can only help after an income source has been reported. It cannot help someone remember a forgotten income source, nor can it motivate someone to report income that he or she does not want to report.

ACKNOWLEDGEMENTS

Many people contributed in important ways to this research project. In particular, we acknowledge Nola Krasko and Richard Taegel for their roles in implementing the experiment in the field, and Elaine Fansler, Peter Wobus, and Lorraine Randall, who played key roles throughout the research.

REFERENCES

Beach, M. E. (1991), personal communication.

Bogen, K., J. Moore, and K. Marquis (1994), "Can We Get Respondents to Use Their Personal Income Records?" Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 1252-1257.

Bollinger, C. and M. David (1993), "Modelling Food Stamp Participation in the Presence of Reporting Errors," Proceedings of the 1993 Annual Research Conference, U.S. Bureau of the Census, Washington DC, pp. 289-312.

Bollinger, C. and M. David (1995), "Sample Attrition and Response Error: Do Two Wrongs Make a Right?" Proceedings of the 1995 Annual Research Conference, U.S. Bureau of the Census, Washington DC, (forthcoming).

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.

Cantor, D. (1991), "Results of SIPP Interviews Used for Preparation for Milwaukee Pre-Test," Memorandum for Kent Marquis, Karen Bogen, and Jeff Moore, September 5, 1991.

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

Cohen, G. and R. Java (1995), "Memory for Medical History: Accuracy of Recall," Applied Cognitive Psychology, Vol. 9, pp. 273-288.

Goodreau, K., H. Oberheu, and D. Vaughan (1984), "An Assessment of the Quality of Survey Reports of Income from the Aid to Families with Dependent Children (AFDC) Program," Journal of Business and Economic Statistics, Vol. 2, pp. 179-186.

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.

Klein, B. and D. Vaughan (1980), "Validity of AFDC Reporting Among List Frame Recipients," Chapter 11 in J. Olsen (ed.), Reports from the Site Research Test, U.S. Department of Health and Human Services, ASPE/ISDP/SIPP, Washington, DC.

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. (1995), "The SIPP Measurement Quality Experiment and Beyond: Basic Results and Implementation," Proceedings of the 1995 Annual Research Conference, U.S. Bureau of the Census, Washington DC, (forthcoming).

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

Marquis, K., J. Marshall, and S. Oskamp (1972), "Testimony Validity as a Function of Question Form, Atmosphere, and Item Difficulty," Journal of Applied Social Psychology, Vol. 2, pp. 167-186.

Marquis, K., and J. Moore (1990), "Measurement Errors in SIPP Program Reports," Proceedings of the 1990 Annual Research Conference, U.S. Bureau of the Census, 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., K. Bogen, and K. Marquis (1993), "A 'Cognitive' Interviewing Approach for the Survey of Income and Program Participation: Development of Procedures and Initial Test Results," Proceedings of Statistics Canada Symposium 92, "Design and Analysis of Longitudinal Surveys," pp. 31-40.

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.

Shea, M. (1995), "Dynamics of Economic Well-Being: Program Participation, 1990 to 1992," Current Population Reports, P70-41, U.S. Bureau of the Census, Washington, DC.

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.

Snedecor, G. and W. Cochran (1967), Statistical Methods, Ames, IA, The Iowa State

University Press, sixth edition.

Appendix A

SIPP Cognitive Research Evaluation Experiment:

Experimental Treatment Wave 1 Questionnaire

Form SIPP-11100 (X) G U.S. Department of Commerce Bureau of the Census SURVEY OF INCOME AND PROGRAM PARTICIPATION - CR Wave 1 Questionnaire	
1. Final Interivew Status: 1-[] Complete for all persons 15+ 2-[] Partial Household Compilation (enter person numbers in each category) > Complete interviews: _ _ _ _ _ _ _ _ _ _ _ _ > Type Z (refusal): _ _ _ _ _ _ _ _ _ > Type Z (other): _ _ _ _ _ _ _ _ _	
2. Pre-Interview Transcription Time: (Wave 2 only)	
4. Total HH Visits/Contacts: _ _ (in person) _ _ (phone)	
6. Total Interview Time for this HH (sum from Record of Visits Card): _ _ _ (hrs) (mins)	
8. 1-[] "Telephone Hold" Case _	
10. Office Operations: a._ b._ c._ d._ e._	

Appendix B

SIPP Cognitive Research Evaluation Experiment:

Experimental Treatment Wave 2 Questionnaire

Appendix C

SIPP Cognitive Research Evaluation Experiment:

Experimental Treatment Income "Worksheets" (Waves 1 and 2)

Appendix D

SIPP Cognitive Research Evaluation Experiment:

Experimental Treatment Monitoring Forms (Waves 1 and 2)

Appendix E

SIPP Cognitive Research Evaluation Experiment:

Experimental Treatment Miscellaneous Other Field Materials

- Introductory letters (Wave 1 and Wave 2)
- Employment calendars (Wave 1 and Wave 2)
- Record of Visits card
- Income Without Receipts form

Appendices available upon request:

Appendix A - Experimental Treatment Wave 1 Questionnaire

Appendix B - Experimental Treatment Wave 2 Questionnaire

Appendix C - Experimental Treatment Income "Worksheets" (Waves 1 and 2)

Appendix D - Experimental Treatment Monitoring Forms (Waves 1 and 2)

Appendix E - Experimental Treatment Miscellaneous Other Field Materials
- Introductory letters (Wave 1 and Wave 2)
- Employment calendars (Wave 1 and Wave 2)
- Record of Visits card
- Income Without Receipts form