THE SURVEY OF INCOME AND
PROGRAM PARTICIPATION

AN EXPERIMENT TO REDUCE MEASUREMENT ERROR IN THE SIPP: PRELIMINARY RESULTS

No. 197

K. H. Marquis, J. C. Moore & K. Bogen
Bureau of the Census

U. S. Department of Commerce BUREAU OF THE CENSUS

AN EXPERIMENT TO REDUCE MEASUREMENT ERROR IN THE SIPP: PRELIMINARY RESULTS[1]

Kent H. Marquis, Jeffrey C. Moore, Karen Bogen, U.S. Bureau of the Census
Kent Marquis, U.S. Bureau of the Census, Washington, D.C. 20233-9150

Key Words: Cognitive techniques, Administrative records, Personal records

A record check study[2] conducted on the 1984 Survey of Income and Program Participation (SIPP) in 4 states, showed underreporting rates[3] for 3 income programs at around 25% and even higher for the unemployment insurance program (see Figure 1). These errors not only distort estimates of means and percentages but also, as Bollinger and David (1993, 1994) have shown, distort estimates of relationships in important policy models that rely on SIPP data.

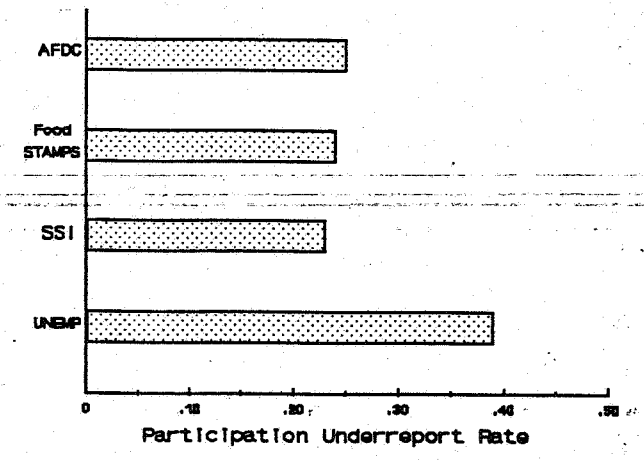

Figure 1. High Underreporting in SIPP (1984 Study).

This led us to design new interviewing procedures with the goal of reducing SIPP underreporting error by at least 25% for these major transfer income programs.

Let us begin with an overview of the various study procedures and then look at the results of an evaluation of the procedures. We shall tell you now that the evaluation didn't come out as we expected. So we'll discuss what seems to have gone wrong and where there may, indeed, have been some successes, namely in reducing error in reporting amounts of income.

Step one for this study was to create face-to-face interviewing procedures that would reduce underreporting errors substantially, by at least 25%. To do this we borrowed heavily from cognitive psychology and from the survey research literature. We highlight differences between the standard and new procedures next.

The standard strategy for reporting income sources and amounts is for the respondent to remember what (s)he can and use that to reconstruct a history for an income source. The new procedures, on the other hand, encouraged respondents to base their reporting on a complete set of personal income records.

In regular SIPP, each adult is interviewed separately. In the new procedures, adults were interviewed together in the first interview to help each other out.

Standard household interviewing uses structured questions and, in standard SIPP, we ask for sources of income first and later ask about income amounts. The new procedures took an unstructured approach, letting respondents decide in what order to report. And to correct any forgetting, we read a list of income sources to the respondents so they could recognize any that they forgot to mention.

Standard SIPP gets monthly totals for each source of income for each person, the new procedures required reporting each income payment separately, relying on the computer to aggregate.

Standard SIPP reminds respondents what income sources were reported in the last interview before getting new information. The new procedures got the new information first, then recalled the old reports and reconciled them. If there were inconsistencies, they were resolved and corrections made in either the current OR THE PAST questionnaire.

Standard SIPP restricts reports to the 4 full months before the current interview. The new procedures included those 4 months but also the days in the current month up to the day of the interview.

Finally, standard quality control involves reinterviewing a sample of completed cases, usually by telephone, to learn whether the interviewer really did conduct an interview. For the new procedures, we taped every interview, systematically coded what the interviewer did from a sample of tapes, and sent a summary report, based on the codes, to the interviewer each month.

After developing the new procedures, we conducted 2 pretests to spot difficulties and introduce refinements. We matched survey reports to administrative records in the second pretest. Figure 2 shows both the pretest underreport rates and repeats the rates from the 1984 study. The pretest error rates were considerably lower. So, encouraged by the pretest results, we set out to conduct a formal evaluation of the new procedures.

To evaluate the new procedures, we used an experimental design with households randomly assigned to one of two treatments: the experimental treatment which used the new procedures, or the control treatment which used standard SIPP interviewing. People aged 15 and

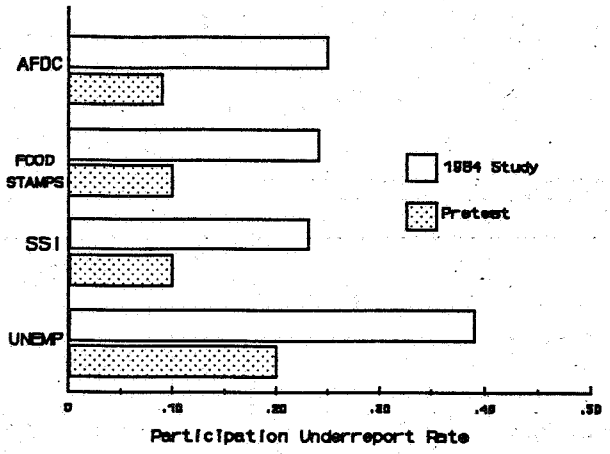

Figure 2. Less Underreporting in Pretest than 1984 Study.

over and living in Milwaukee, Wisconsin, were sampled from the administrative records of the 4 income programs you saw earlier. We interviewed at each sample address. If we found the expected sample person at the address, we kept that household in the sample for the second interview, otherwise we deleted it.

The sample was designed to yield interviews with about 700 households, 350 in each treatment at the end of wave 2. This is what it actually did.

We conducted two face-to-face interviews with each household, called wave 1 and wave 2.

We matched survey and administrative record information on social security number, address, and name for each sample person.

For the experimental treatment we converted the payment-by-payment income data to monthly summaries, smoothing out any artificial time gaps in the recipiency spell created by this aggregation process. We also removed any duplicate income reports made about the same payment but in 2 different interviews.

By matching survey reports to the administrative records we could estimate monthly underreporting of the income sources we checked.

Looking at Figure 3, there are 4 possible outcomes of matching a yes-no survey response to the administrative record. Either "a" or "d" occurs when both sources agree, "b" and "c" are the disagreements. Since we sampled from administrative records, we are interested only in outcomes a and c. The underreporting rate is the number of c-type disagreements over a + c, which is the number of yesses in the administrative record. We averaged over months and people to get the underreporting rates for each treatment.

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

Figure 3. Cross Classifications of Survey and Record Outcomes

Let's take a look at the underreport results for the evaluation study (Figure 4). Here we show the experimental and control group underreporting rates for each of the 4 income programs. The treatments achieved ABOUT THE SAME underreporting rates within each of the programs. Clearly we did not expect these results. We had expected the experimental group rates to be much lower than the control group rates. So what the heck went wrong?

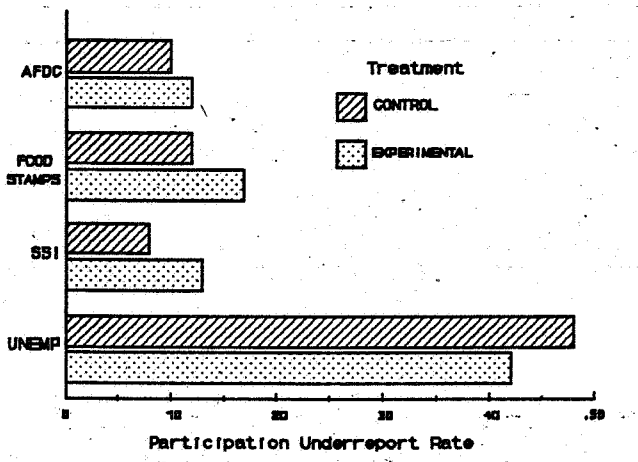

Figure 4. No Treatment Effect on Underreporting in the Experiment.

Let us do a little speculating, first about whether the treatments were implemented correctly and then about whether our ideas were way off base. One thing that may not have worked well was the interviewer staffing. Interviewers were not randomly assigned to treatments. The most experienced interviewers were assigned to the control treatment and they ended up doing most of the work. As would be expected their response rates were considerably higher than those of the experimental treatment.

Control group workloads were larger so interviewers were able to schedule their work more efficiently. In part, this may be why control group interviews cost less than experimental interviews.

On the other hand, the experimental procedures, themselves, may have aided in decreasing response rates and increasing costs. They required more contacts per household and more time to complete an interview (Bogen, Moore and Marquis, 1994).

But let us not forget that even if there were differences between interviewing staffs, these differences would not necessarily affect response errors. The thing we felt would have the greatest impact on response errors was the use of records by households to report their income.

Figure 5 shows that the record use rate was much greater in the experimental treatment than the control for both wave 1 and 2. We conclude that the features promoting personal record use were relatively well implemented in the experimental treatment. Although things may not have been perfect, we are not inclined to blame the unexpected response error results on poor implementation.

	Experimental	Control
Wave 1	49%	12
Wave 2	66	11

Figure 5. Household Record Use Percents

So we turn to a different set of possible reasons for failing to obtain our expected reduction in underreporting.

Although the samples are different and we used slightly different definitions of the underreport rates, notice (in Figure 6) that the control group rates are considerably lower than the rates from the 1984 study. We expected them to be the same. Perhaps there is something unique about Milwaukee or the Milwaukee interviewers, or perhaps underreporting error has been reduced over the intervening years in SIPP, by changes in procedures and increasing experience of the interviewing staff. We cannot say for sure, but one reason that we didn't get experimental effects could be that much of the underreporting we sought to reduce had already been taken care of.

On the other hand, underreporting of unemployment insurance payments was consistently high over all of the studies and treatments and serves to remind us that some of the most severe response error problems remain to be solved.

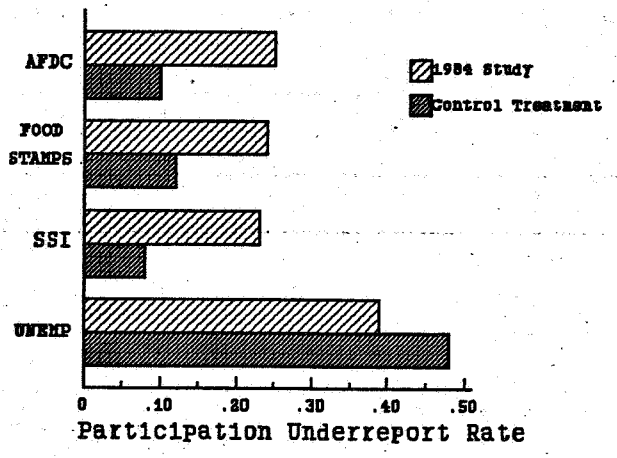

Figure 6. Much Underreporting Already Reduced.

So, while we are on the topic of wrong-headed ideas, let us admit to our assumption that, because we used an income recognition list, people would report their income sources correctly. It turns out, however, that most of the monthly underreport errors stem from underreporting an entire source of income, rather than from underreporting some of the benefit payments. In Figure 7 we've partitioned each underreport rate bar into two components: on the left is the underreportting due to not reporting an entire income source and, on the right is underreporting due to other causes. Notice that the left component is always bigger than the right component.

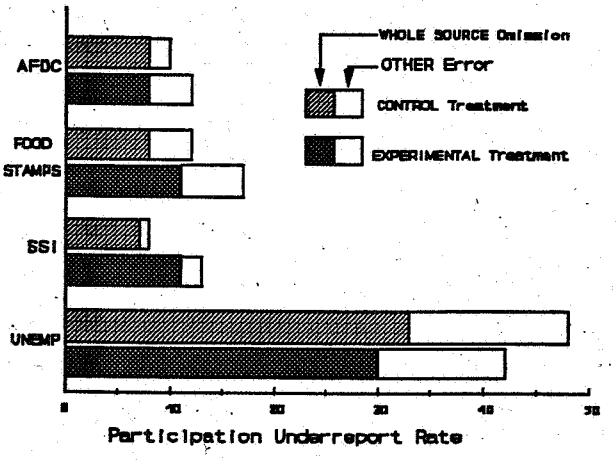

Figure 7. Most Underreports due to Whole Source Omission.

This leads us to conclude that we should have paid more attention to whole source underreporting when designing our new procedures, rather than just trying to improve the reporting of the income details.

Already people are speculating about how to reduce the underreporting of entire income sources. If you think the cause is cognitive, then you suggest better recognition cues, and ways of focusing the respondent's attention on them.

If the underreporting is intentional or motivated, we need to rethink parts of the new procedures. For example, by requiring family group interviews, individuals don't have the privacy to report income that they don't want to discuss with the rest of the family. So private telephone or face-to-face interviews might help, although clearly they would cost more.

Also, confidentiality concerns need attention, especially in cases where people may be getting benefits they aren't eligible for or other income that violates the law. Perhaps we should not tape record interviews and find some other way to monitor interviewer behavior.[4] And we need to find ways to be more persuasive about our ability to maintain absolute confidentiality for anything the respondent reports.

We'd like to end with good news based on an exploratory analysis of reporting income AMOUNTS.

If the respondent and administrative record agreed that there was a payment in a particular month, we said the reported amount was correct if it was within 5% of the truth. For AFDC and food stamps (top of Figure 8), it appears that accuracy gets relatively better over time for the experimental treatment. These interactions are statistically significant in a repeated measures analysis of variance using sample people included in both waves.

And the trend continues for the last two programs, SSI and unemployment insurance (bottom of Figure 8). The SSI interaction is statistically significant; the unemployment interaction, based on fairly small sample sizes, is not.

This is what we really want in a panel survey, for respondents to learn what is expected and to get better at doing it over time.

In conclusion, let us review the main points.

Although we designed some really fantastic, new procedures, our experimental evaluation showed no treatment effects on participation underreporting.

We feel that the new procedures probably received a "fair test" because a main feature of the experimental treatment, increased use of personal records, was implemented adequately.

We note that much of the underreporting we set out to reduce was not present in the control group. Perhaps

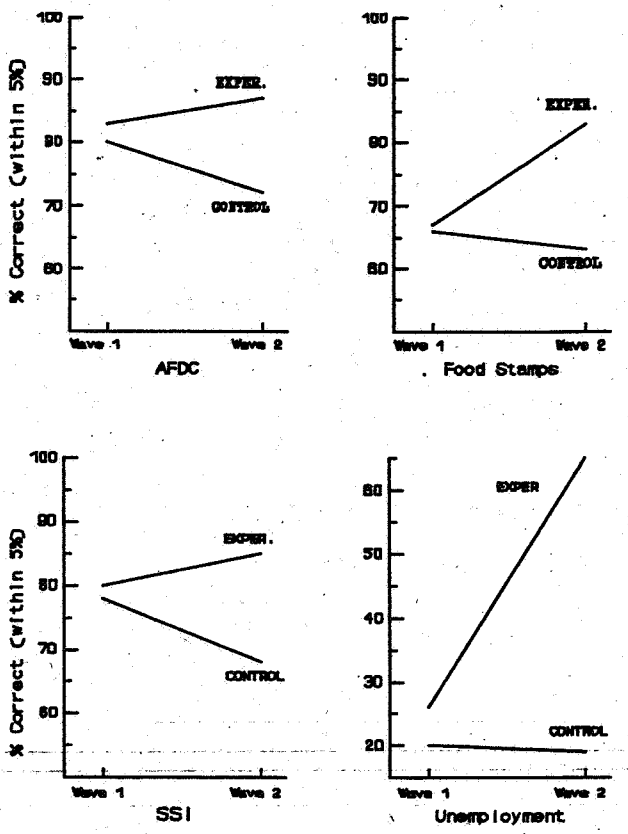

Figure 8. The correct reporting of AMOUNTS got relatively better over time in the Experimental Treatment.

Milwaukee was atypical or perhaps much of the error we found in the early SIPP has been dissipated by improved procedures, better interviewer training, and so forth. In any case, much of the error that we had targeted was not there to be reduced.

But before getting too hopeful, recall that underreporting of unemployment insurance has been and remains drastically high. It did not improve at all.

The underreporting of whole sources of income seems to underlie the majority of the monthly underreporting of participation. If the underlying causes are motivational, we need to focus additional design attention on privacy and confidentiality issues.

The new procedures do show some promise for better accuracy of reporting amounts of income. We hope to have some good news to report next year in this area.

REFERENCES

Bogen, Karen, Jeffrey Moore, and Kent Marquis (1994), "Can We Get Respondents to Use Their Personal Income Records?," presented at the 1994 Conference of the American Association of Public Opinion Research.

Bollinger, Christopher and Martin David (1993), "Modeling Food Stamp Participation in the Presence of Reporting Errors," Proceedings of the 1993 Annual Research Conference, Bureau of the Census.

Bollinger, Christopher and Martin David (1994), "Modeling Food Stamp Participation in the Presence of Reporting Errors," 1993 Proceedings of the Social Statistics Section, American Statistical Association.

Marquis, Kent and Jeffrey Moore (1990), "Measurement Errors in SIPP Program Reports," Proceedings of the 1990 Annual Research Conference, Bureau of the Census.

Marquis, Kent, Jeffrey Moore, and Karen Bogen (1994) "Effects of a Cognitive Interviewing Approach on Response Quality in a Pretest for the SIPP," 1993 Proceedings Section on Survey Research Methods, American Statistical Association.

Sanchez, Maria E. (1993), "Enhancing Compliance with Record-Keeping in a Household Survey," Agency for Health Care Policy and Research. Paper presented at the 1993 Conference of the American Association of Public Opinion Research.

Singh, Rajendra P. (1992) "Wave 2 Results of the Record Check Study," Unpublished memorandum for the SIPP Research and Evaluation Steering Committee, U.S. Census Bureau.

Notes

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

[2] For general results, see Marquis and Moore (1990). For underreporting rates, see Marquis, Moore, and Bogen (1994).

[3] The underreport rate for a program is U/T, where T is the number of participation months recorded for all people in the survey from administration records, R is the number of participation months reported for all people in the survey, and U is T-R.

[4] Since tape recording may have large, positive effects on data quality in most households, to abandon it one would need to show 3 things: (1) that it has negative effects in selected households (e.g., it causes ineligible recipients to withhold reporting their recipiency), (2) That those effects can be mitigated by not taping and (3) that the quality losses for the few households outweigh the quality gains for most households.