SIPP RECALL LENGTH DECISION: AN ALTERNATIVE TO EXPERIMENTATION

Hertz Huang, Kevin Cooper*, Gary Shapiro, Kathleen Short, U.S. Bureau of the Census
Hertz Huang, U.S. Bureau of the Census, Washington, D.C. 20233

I. INTRODUCTION

The Survey of Income and Program Participation (SIPP) is a nationwide survey designed to provide more accurate and comprehensive information than has previously been available about the income and program participation of persons and households in the United States.

Currently, SIPP interviews sample households at 4-month intervals, with corresponding 4-month recall, for a total of about 8 interviews. There was a serious proposal in 1992 to change to 6-month intervals with 6-month recall. This change would allow us a substantial increase in sample size, and the possibility of interviewing sample for a longer calendar period. The obvious potential disadvantage to such a change would be lower quality data as a result of higher response variance and response bias due to the longer recall period.

The usual Census Bureau reaction to this type of proposal is to conduct a field test. For this situation, an experiment would directly compare 4-month and 6-month recall. Since, a definitive experiment would require a large sample size and therefore be very expensive. For this proposal, we deviated from our norm and considered whether we could reach a conclusion without conducting an experiment. A Census Bureau workgroup attacked the problem and succeeded in making a strong recommendation about recall length without an experiment. The workgroup consisted of Hertz Huang (chair), Rajendra Singh, Kevin Cooper, Enrique Lamas, Kathleen Short, Martina Shea, Louisa Miller, Patrick Benton, Zelda McBride and Steve Willette. Gary Shapiro, I-Li Lu and Jeff Moore also made significant contributions to the work.

This paper reports on the basic approach of our workgroup. We hope that readers will find this paper of value to them when they face sample design decisions for which an experiment may not be feasible (or may even be impossible).

There were two major aspects of the workgroup's approach. One was a simulation that compared "truth", 4-month recall, and 6-month recall for the distribution of the amount of time(spell length) receiving food stamps, receiving aid to families with dependent children (AFDC) or participating in any programs. Section III describes the general approach of this simulation. The second major aspect was a comparison of mean square errors for various characteristics and sample sizes under alternative bias assumptions (based on our simulations). Section IV discusses this aspect. The mean square errors showed that 4-month recall was clearly superior to 6-month recall. Past studies on recall length were also useful in making a decision. The information on this is summarized in the next section. Also, decision making tools suggested by Neustadt and May (1986) helped and are discussed in Section V.

II. PREVIOUS RECALL-LENGTH STUDIES

Several previous recall-length studies were helpful to us in making our decision. These studies found that some statistics were significantly affected by recall length. Several studies concluded that 6-month recall led to greater bias than 3-month recall. The most useful of these studies was the one for the Income Survey Development Program which is discussed here.

The Department of Health, Education and Welfare initiated the Income Survey Development Program (ISDP) to develop concepts and contents and to examine and resolve technical and operational issues prior to adopting a final design for the proposed SIPP. The ISDP Site Research Test (SRT), was conducted in Fall 1977 and Spring 1978 as the first phase of the ISDP. The sample of 5,500 adult respondents was drawn from AFDC administrative records, Supplemental Security Income (SSI) administrative records, and a household area frame. The purpose of the SRT was to test variations in questionnaire detail and in recall lengths. Four collection methods were tested: two treatments referred to the level of detail in the questionnaire (short versus detailed form); and two treatments referred to the length of recall (3 months versus 6 months). Two interviews were conducted for the 3-month recall and one was conducted for the 6-month recall covering the same time period.

The SRT concluded that: 1) A longer recall period will result in significantly fewer persons reporting income, especially for earlier periods, that is, periods furthest from the interview date. This general

* Kevin Cooper is now at the University of Michigan.

result holds for total income as well as for specific income types such as AFDC, wages, and short-term transfers; and, 2) Income levels were not significantly different between the recall periods.

III. THE AFFECTS ON LONGITUDINAL ESTIMATES

In the consideration of length of recall period it was determined that one important criterion in our decision should be the affect on longitudinal estimates from the SIPP. Longitudinal estimates from the survey are important because they represent the unique contribution that SIPP can make. Therefore, longitudinal estimates, such as transitions from one state to another or durations of time spent in a given state, became our focus, where a state of interest might be receiving poverty-level-income or participating in a government program. Conclusions from the ISDP Site Research Test suggest that longer recall periods reduce the proportion of respondents reporting some positive amount of income (wages, Social Security, AFDC, and residual short term transfers). Exits from and entrances into poverty would be affected by the non-reporting of income in a longer recall period because fewer income types, particularly of a temporary nature, would be reported. Thus we would expect that more persons would appear to be poor at any time and appear to remain poor, while in truth, they may have received some income that brought them temporarily above the poverty line. Variability of income, and the number of exits and entrances from/into poverty would be reduced by greater non-reporting of income receipts. However, the ISDP Site Research Test left the affects of recall length on other longitudinal estimates virtually unanswered. In this section, we simulate the affects of recall length on spell duration.

A. METHODOLOGY OF SPELL DURATION SIMULATION

To study the affect of recall length on spell duration estimates, we looked first at the median spell duration for program participation. We defined a program spell as a period of participation preceded by one or more months of non-participation. A spell is observed until it ends or until it is right-censored. A right-censored spell is a spell that is still in progress when the survey ends.

To predict the affect on median spell estimates if we change from 4-month to 6-month recall we needed the spell length distribution for SIPP data when using 6-month recall. There is no existing data, so our goal was to obtain useful information without conducting a field experiment. In order to simulate the 6-month recall spell length distribution, we needed:

- the underlying or true distribution of spell lengths.
- the respondent reporting pattern--how the underlying distribution is altered when using interviews with 4-month recall.

We first determine the affects that 4-month recall has on the underlying spell length distribution, and then from these 4-month recall affects, to predict what would happen if the SIPP used 6-month recall.

Current SIPP reported data is biased by the 4-month recall reporting-error pattern. In other words the underlying spell distribution has been altered because SIPP respondents do not have perfect memory or they are careless in their reporting. Since our reported data are already biased, our attempt to simulate the underlying distribution of spells and the respondent reporting-error pattern simultaneously was an iterative process. We worked backwards trying to reproduce the reported data. Our steps were:

- Assume a reporting-error pattern.
- Assume an underlying spell duration distribution.
- Apply the reporting-error pattern to the distribution.
- Compare the transformed distribution to the reported distribution.

1. Assume a reporting-error pattern

During an interview, SIPP respondents are required to recall 4 months of program participation status. To simplify our model we assumed that if a respondent misreported his participation status he did so in one of two ways: (1) he reported "not receiving" for the entire recall period although he received for at least one month in the period or (2) he reported "receiving" for the entire recall period when he did not receive for the entire period. (Such as a tendency, to over-report transitions between the last month of one interview and the first month of the following interview, and to under-report transitions between adjacent months covered by the reference period for a single interview, is also called seam phenomenon.) Because of these assumptions, spells that are misreported are assumed to be reported on one of the seam months (for 4-month recall a seam-month spell is a spell with a duration which is divisible by four).

The 1987 SIPP AFDC data were collected using 4-month recall. The peaks that occur every four months seem to imply some sort of misreporting.

Allocation to surrounding seams:

We made an assumption to determine how a misreported spell would be reported as a seam-month spell. We assumed that only a portion of the spells are misreported. For misreported spells: we assumed that a respondent would report "receiving" for the entire recall period if he was receiving in the month prior to the interview (the last month of a recall period), otherwise he would report "not receiving" for the entire recall period. In other words, the respondent would report for the entire recall period whatever his status was in the last month of the period.

This assumption was also used by Kalton, et al (1992). This type of response is sometimes referred to as a constant wave response. Under the constant wave response process, a respondent who starts a spell on a program in the middle of a wave will report being on the program for the whole of the wave, and the program start would appear to have occurred at 4 months of recall.

For 4-month recall, 25% of spells of one-month duration occur in the month prior to the interview. Thus, according to our assumptions, 25% of misreported one-month spells are reported as four-month spells and 75% are not reported. Fifty percent of two-month spells are active in the month prior to the interview month. So, 50% of misreported two-month spells are reported as four-month spells and 50% are not reported. And 75% of misreported three-month spells are reported as four-month spells and 25% are not reported.

A reasonable but arbitrary reporting-error pattern for four-month recall is 40% of spells duration one month, 35% of spells duration two months, 30% of spells duration three months and 25% of spells duration four months or longer are misreported.

The 4-month reporting pattern was extended to predict the 6-month reporting pattern. Therefore, for the misreported spells: percent not reported and percent reported as 6-month recall are 83 and 17 for spells of 1-month duration, 67 and 33 for spells of 2-month duration, 50 and 50 for spells for 3-month duration, 33 and 67 for spells of 4-month duration and 17 and 83 for spells of 5-month duration. An assumed reporting-error pattern for six-month recall is 50% of spells duration one month, 45% of spells duration two months, 40% of spells duration three months, 35% of spells duration four months, 30% of spells duration 5 months and 25% of spells duration six months or longer are misreported.

2. Assume an underlying spell duration distribution

Some of our early attempts to simulate the underlying spell distribution included cubic and exponential regressions fit to 1987 reported SIPP data, functions of the form 1000 / (X+c) for 'c' a constant, and 'X' the spell duration in months, and various uniform distributions. The resulting 4-month recall spell distributions were not as close to the reported spell distribution as was the technique that we are describing.

Step 1: Imputation of lost spells

From Vaughan and other research we know that longer recall lengths imply more lost short spells. Since reported data has been biased by the 4-month recall reporting pattern, some short spells are not reported. We decided to impute spells of duration less than four months. The reporting pattern was used to calculate the number of imputed lost spells.

Step 2: Data smoothing

To simulate the underlying spell distribution we used a modified data smoothing technique. SIPP 1987 reported AFDC data (after adjusting for lost spells) was smoothed using a three month moving average technique. (See figure 1)

We fitted two lines to this smoothed data. One line was fitted to the smoothed data for one month spells to seven month spells. The other line was fitted to the eight month spells to 27 month spells. This "segmented linear regression" was the assumed underlying "true" spell distribution.

3. Apply reporting patterns to the underlying distribution

We applied the 4-month and 6-month reporting patterns to our underlying distribution. Figure 2 shows the reported data and our simulated 4-month recall data for AFDC spell. These two graphs are quite similar and demonstrate that our simulations have reproduced the reported data very well.

We also applied the above reporting patterns to other models of the "true" distributions: Linear combinations of the "segmented" regression fit to AFDC data; 1000/(X + c), where c is a constant and X is the spell duration (in months); cubic regression fit to reported AFDC data; exponential fit to reported AFDC data. Then the same comparisons were made (bias in median spell, square difference, seam ratio). In all cases the original 4- and 6-month reporting patterns were applied to the distribution. These comparisons were made since the underlying distribution for other programs (such as Medicaid recipiency or months in poverty) might resemble these curves or some member of these families of curves.

We carried out the same data smoothing techniques for SIPP 1987 reported Food Stamps data and applied our reporting-error patterns. The procedure used for the food stamp simulation is identical to the one used for the AFDC simulation. Again, our simulations reproduced the reported data quite well.

B. Statistics Used for Comparison of the Spell Duration Distributions

We used the following measurements to compare the simulated spell distributions under 4-month recall versus 6-month recall:

1. Bias in the median spell duration: We calculated the median spell duration for each distribution. These medians were compared to the median from the underlying distribution to determine the bias.

2. Square difference: Which recall length distribution is closer to the underlying distribution? The measure that we used was the square difference. The square difference was defined as the sum of the squared differences (of spell duration frequencies) between the underlying distribution and the recall distribution.

3. Seam ratio: A major concern was that the longer recall length might exacerbate the seam problem. This was addressed with a statistic which compared the relative size of the seam months to the non-seam months. This ratio was the average number of spells per seam month divided by the average number of spells per non-seam month. It was computed for both recall length distributions.

C. Results of Spell Duration Simulation

Table 1 shows the results of our simulation on Food Stamps spell duration. (Results on other characteristics are not shown due to space limitations.) The bias in the median spell was calculated by subtracting the median spell duration of the underlying distribution from the median spell duration of the recall-length distribution. In all cases the bias for 4-month recall was smaller. The bias introduced by using 4-month recall ranged from 7 to 34 percent of the underlying or "true" median. For 6-month recall the bias ranged from 16 - 37 percent of the underlying median. The ratio of 6-month / 4-month bias ranged from 1.04 to 2.80. Since the median is fairly insensitive to changes in the distribution, other estimates would probably be more biased by the recall lengths.

The ratio (6-month/4-month) of the square difference, ranges from 1.55 to 1.90. This implies that the 4-month recall distribution is substantially closer to the underlying distribution than is the 6-month recall distribution. Hence, our simulations point to better data quality when 4-month recall is used.

The seam to non-seam ratio was smaller for 4-month recall, regardless of the underlying distribution. The 6-month ratio was between 26 and 34 percent larger than the 4-month ratio. This implies that the seam problem will be exacerbated by changing to 6-month recall.

From Vaughan (1989) it was anticipated that 6-month recall would report approximately 95 percent of the spells that 4-month recall reported. Our simulations showed 6-month recall reporting 93 percent of the total spells that 4-month recall reported.

IV. AFFECTS ON MEAN SQUARE ERROR

Currently, the SIPP interviews respondents three times a year. If the SIPP used a 6-month recall, then only two interviews are needed per year. This savings would allow us to increase the sample size by as much as fifty percent and, of course, decrease the variances of our estimates. However, SIPP estimates are already biased by the 4-month recall, and from the last section and previous research we believe that changing to 6-month recall will increase these biases. An important question then becomes: What affect would the larger sample size have on the mean square error (MSE = variance + bias^2) of SIPP estimates?

Under the same cost limitations, by changing from 4-month recall to 6-month recall, we could actually increase our sample by at most fifty-percent. Therefore, the following calculations are based on this best-case scenario for the 6-month recall.

We estimated SIPP variances for sample sizes of n=20000, 30000, 50000, and 75000 households. For example, if SIPP had a sample size of 20,000 households using 4-month recall then we were interested in the variance if the sample size is increased to 30,000 households under 6-month recall.

Table 1 shows the results from our MSE simulations for percent of persons ever participating in Medicaid. We made an attempt to include estimates with different characteristics in these simulations (some key statistics for each subgroup: total or whites, blacks and all others). We also included cross-sectional and longitudinal estimates in our simulations. (Due to space limitations, other tables are not shown.)

The top row of the table is the percent bias in p, when using a 4-month recall. We believe that biases as high as 15% and 20% are reasonable for some SIPP estimates. Note that the first column gives the MSE for unbiased estimates, therefore these entries are the variances for the different sample sizes. The row headings are for the additional bias that would be introduced by using 6-month recall. A ratio of 1.10 means the switch to 6-month recall increased the 4-month recall bias by 10 percent.

The first row of MSEs in the table uses n=20,000 and the various bias assumptions. This row is intended to be the 4-month recall MSEs. Below each 4-month recall MSE are the MSEs with n=30,000 for ratios of bias ranging from 1.00 to 1.30. These entries correspond to the 6-month recall MSEs under various additional bias assumptions. The idea is to compare the first entry in the column to the other entries in that column to determine the effect on the MSE when changing to 6-month recall under the different additional bias assumptions.

Assume SIPP, under 4-month recall, has a sample size of 20,000 households and a bias of 5 percent, then under 6-month recall n=30,000 households and a larger bias is incurred. For most of the large subgroup estimates that we examined, the gain in precision from the larger sample is lost if the bias increases by 10 percent or more. In other words for most of the estimates that we looked at the MSE is dominated by the bias.

Assume SIPP, under 4-month recall, has a sample size of 20,000 households and a bias of 20 percent. Then in most cases examined, the change to 6-month recall cannot compensate for any non-trivial increase in bias. Even if there is no additional bias incurred under 6-month recall the absolute gain in precision is minimal.

Assume SIPP, under 4-month recall, has a sample size of 50,000 households and a bias of 5 percent. Then the gain in precision from a switch to 6-month recall is negligible for most of the large subgroup estimates that we looked at. For most estimates, any non-trivial (10 percent or more) bias incurred by switching to 6-month recall is enough to nullify the benefit from a larger sample.

V. CONCLUSIONS

The conventional Census Bureau response to a question such as whether to change from 4 month recall to 6 month recall is to conduct an experiment. In this case we have been able to make a decision by doing other things that were cheaper and quicker. The most important things we did were to conduct a simulation (discussed in section III) and to conduct a mean square error sensitivity analysis (discussed in section IV). These showed that 6 month recall resulted in generally much larger biases than 4 month recall, and that bias dominated the mean square error for the relevant characteristics and sample size. The literature review (discussed in section II) was also very useful and influential in this decision making.

The simulation and mean square error analysis were actually used in conjunction with another innovation. We applied several methods from Neustadt and May (1986). These authors suggest a variety of techniques for a sound decision-making process. Their emphasis is on correct use of historical information. Petroni (1992) used the methods of noting the likenesses and differences in apparent analogies, reviewing the issue history, and listing key elements for the decision as to whether information on the element is known, only presumed, or completely unclear. This latter method proved particularly valuable. In Petroni's original listing of key elements, there were only 5 known elements and 15 that were presumed or unclear. After our investigation was completed, many more elements are now known. Of particular importance in deciding whether to conduct an experiment is that an experiment would probably not help in providing more knowledge for any of the unclear or presumed elements.

In summary, on the basis of the simulation, the mean square analysis, overview that an experiment would not help with unclear and presumed elements, and the literature review, the Census Bureau decided to continue with 4-month recall without conducting a costly experiment.

REFERENCES

Burkhead, Daniel and Coder, John (1985), "Gross changes in Income Recipiency from the Survey of Income and Program Participation." Proceedings of the American Statistical Association, Social Statistics Section, 351-356.

Hill, Daniel H. (1987), "Response Errors Around the Seam: Analysis of Change in a Panel with Overlapping Reference Periods," Proceedings of the American Statistical Association, Survey Research Methods, 210-215.

Kalton, Graham, Miller, David P., and Lepkowski, James (1992), "Analyzing Spells of Program Participation in the SIPP," Final Report for Joint Statistical Agreement 90-36 between the Bureau of the Census and the Survey Research Center, University of Michigan.

Kobilarcik, Edward L., Alexander, Charles H., Singh, Rajendra P., and Shapiro, Gary M. (1983) "Alternative Reference Periods for the National Crime Survey," Proceedings of the American Statistical Association Survey Research Methods Section, 197-202.

Mathematica Policy Research (1979), "Survey of Income and Program Participation Site Test Analysis: The Evaluation of Experimental Effects on Data Quality," Task 2 Report Submitted to U.S. Department of Health and Human Services.

Neter, John and Waksberg, Joseph (1964), "Reporting Errors in Collection of Expenditures Data by Household Interviews: An Experimental Study," Journal of the American Statistical Association March 1964, Vol. 59, pp. 18-55.

Neustadt, Richard and May, Ernest (1986), "Thinking in Time," The Free Press.

Petroni, Rita (1992), "SIPP: Recall Length Study," Internal Memorandum from Rajendra Singh to Senior

Management Redesign Team, United States Bureau of the Census.

Vaughan, Denton R., "Reflections on the Income Estimates From the Initial Panel of the Survey of Income and Program Participation (SIPP)", Survey of Income and Program Participation Working Paper Series No. 8906, United States Bureau of the Census.

Table 1: The 4-Month vs. 6-Month Recall Length Comparisons
Simulated Food Stamps Spell Duration
Median = 4.95

	Median	Percent Bias	Square Difference	Average Seam	Average Non-Seam	Seam/ Non-Seam
4-month recall	5.91	19	9023829	1241	584	2.13
6-month recall	5.99	21	14664914	1469	538	2.73
ratios (6mo/4mo)		1.08	1.63	1.18	0.92	1.29

Table 2 Comparison of Mean Square Errors
Percent (p) of persons ever participating in Medicaid (p=7.3)

Total bias by recall length and sample size (n)		4-Month recall bias (percent of p)				
		0	5	10	15	20
4-month recall with n=20000		0.06	0.19	0.59	1.36	2.19
6-month recall with n=30000						
Ratio of 6-month to 4-month bias	1.00	0.06	0.17	0.57	1.34	2.17
	1.05	0.06	0.19	0.63	1.36	2.39
	1.10	0.06	0.20	0.69	1.49	2.62
	1.15	0.06	0.22	0.75	1.63	2.86
	1.20	0.06	0.25	0.81	1.77	3.11
	1.30	0.06	0.27	0.94	2.07	3.66
4-month recall with n=50000		0.03	0.16	0.54	1.23	2.16
6-month recall with n=75000						
Ratio of 6 month to 4-month bias	1.00	0.02	0.16	0.56	1.22	2.16
	1.05	0.02	0.17	0.61	1.35	2.37
	1.10	0.02	0.19	0.67	1.47	2.60
	1.15	0.02	0.20	0.73	1.61	2.84
	1.20	0.02	0.22	0.79	1.75	3.09
	1.30	0.02	0.25	0.92	2.05	3.43

Figure 1: AFDC Spell Duration
Reported vs. "Smoothed"

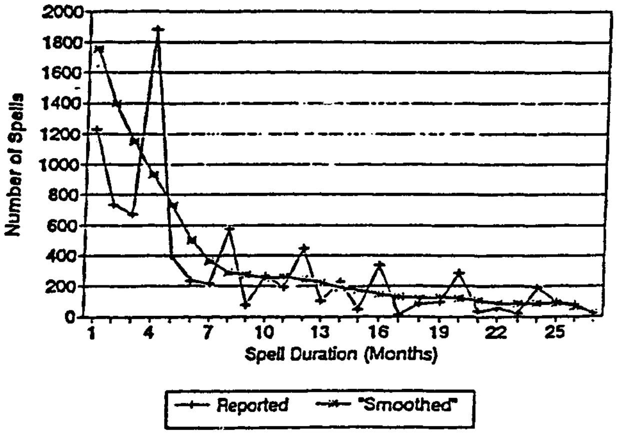
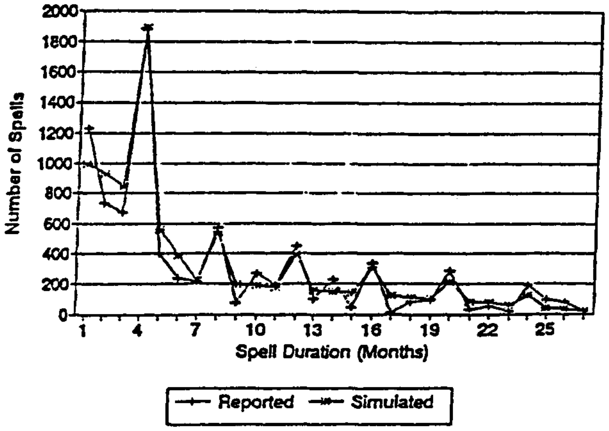


Figure 2: AFDC Spell Duration
4-Month Recall


Number of Spells
Spell Duration (Months)
Reported
Simulated