AN ADDITIVE MODEL OF RECALL ERROR: ANALYSIS OF SIPP DATA

Daniel H. Hill, Survey Research Center[1]

Interviewing in the Survey of Income and Program Participation (SIPP) is conducted on a four month rotating schedule and respondents are asked about their experiences over the last four months. One consequence of this design is that, for any particular calendar month, reports involve anywhere from one to four months of recall, depending upon which 'rotation group' the respondent has been assigned. Furthermore, since SIPP is a panel study, the extent of respondent 'conditioning' also varies from one rotation group to the next after the first four interviewing months. It is quite possible that the quality of the data is affected by both the length of recall and extent of conditioning. If so, efficient estimation of monthly population parameters would require that this heterogeneity of the data quality be taken into account.

The purpose of this paper is to capitalize on the SIPP design to test for the *existence of* (rather than the precise patterns of) differences in data quality which are systematically related to length of recall or extent of respondent conditioning. The paper is organized into three sections. In Section I we briefly describe the SIPP design and incorporate it into an additive model of recall error. In the next section we describe our sample and estimating procedures, while in the third section we present the empirical results.

Before going on to our description of the SIPP design, a couple of words on why length of recall and extent of respondent conditioning should affect data quality are in order. Length of recall affects data quality because as the recall period increases so does the probability that the respondent will fail to recall the particular events which are used to construct the full response to the survey question. Furthermore, respondent errors in the placing of events in time, also increase with length of recall, but at a decreasing rate. Respondent conditioning may also affect data quality, although it is difficult to know in what direction. On the one hand, as the respondent becomes experienced with the survey, he learns what is expected of him and can better prepare himself to provide accurate answers. On the other hand, as the novelty of the survey experience wears off, the respondent may be more willing to simplify reality in order to take short-cuts in the interviewing process.

Section 1
SIPP Design and a Model of Recall Error

The SIPP questionnaire is administered every four months to the same representative sample of adults in the U.S. Each respondent is asked about his earnings from each of his jobs in each of the four months of the reference period. These reports are taken as our dependent variables. To save costs in training interviewers, the sample is split into four random sub-samples. or 'rotation groups'. which are interviewed sequentially in a monthly rotating fashion. The first rotation group was interviewed in October 1983 and was asked about monthly earnings for the June through September period. The second rotation group was first interviewed in November and asked about the July through October period. Etc., etc. The result of this design is that, for any given calendar month, reports involve anywhere from one to four months of recall depending upon rotation group membership. Table 1 summarizes the relationship between rotation group and the amount of recall, as well as the number of times-in-sample, associated with each monthly earnings report for the September 1983 through May 1984 period.

TABLE 1
LENGTH OF RECALL

Reference Month	Length of Recall by Rotation Group			
	1	2	3	4
Sep ('83)	1	2	3	4
Oct	4	1	2	3
Nov	3	4	1	2
Dec	2	3	4	1
Jan	1	2	3	4
Feb	4	1	2	3
Mar	3	4	1	2
Apr	2	3	4	1
May	1	2	3	4

In order to see how this design can be used to address the question of whether differential recall errors represent a significant problem, it is necessary first to develop a model of the reporting process. Following, to the extent possible, O'Muircheartaigh's (1986) notation person j's earnings as reported during trial t for month m can be expressed as:

$$y_{mjt} = \bar{y}_m + \Delta_{mj} + \beta_j + \epsilon_{jt} \qquad 1)$$

where: $\bar{y}(m)$ is the true average earnings in month m of the population of inference; $\Delta(mj)$ is individual j's true deviation from this average; $\beta(mj)$ is the 'fixed response error' or bias; and $\epsilon(mjt)$ is the variable response error associated with trail t. Both the β and the variance of the ϵ can be expected to be affected by the length of recall and number of times in sample, while neither $\bar{y}$ nor Δ—the structural portion of the response—will be affected by these conditions of the interview.

The expectation of 1) across both trials and individuals is:

$$E_{\cdot tj}(y_{mjt}) = \bar{y}_m + \bar{\beta} \qquad 2)$$

where $\bar{\beta}$ is the average of the fixed response errors or the bias. For the reasons noted in the introduction, we would expect $\bar{\beta}$ to be a function of length of recall and number of times in sample. While we could attempt parameterizations of this function, a more flexible alternative is to treat the various combinations of length of recall and number of times in sample as distinct discrete shift parameters where $\bar{\beta}(cr)$ is the mean fixed response error associated with reports involving r months recall and c times in sample.

With this notation in mind we are now ready to see how the SIPP rotating design allows investigation of our hypotheses. Table 2 illustrates some of what can and cannot be estimated using the SIPP design. The top panel includes the expected value of reports for September and October 1983 from each of the four rotation groups. In all there are eight linear equations in seven unknowns ($\bar{Y}(s)$, $\bar{\beta}(11)$, $\bar{\beta}(12)$, $\bar{\beta}(13)$, $\bar{\beta}(14)$, $\bar{Y}(o)$, and $\bar{\beta}(24)$). Unfortunately, however, these eight equations are not independent. Just as age, period and cohort effects are inseparably linked, so are month, recall and time-in-sample effects. We can not identify the individual parameters.

TABLE 2
LENGTH OF RECALL

Reference Month	Expected Value of Monthly Reports by Rotation Group			
	1	2	3	4
Sep ('83)	$\bar{y}_s+\beta_{11}$	$\bar{y}_s+\beta_{12}$	$\bar{y}_s+\beta_{13}$	$\bar{y}_s+\beta_{14}$
Oct	$\bar{y}_o+\beta_{24}$	$\bar{y}_o+\beta_{11}$	$\bar{y}_o+\beta_{12}$	$\bar{y}_o+\beta_{13}$

Sep ('83);$\bar{y}^*_s$;$\bar{y}^*_s+(\beta_{12}-\beta_{11})$;$\bar{y}^*_s+(\beta_{13}-\beta_{11})$;$\bar{y}^*_s+(\beta_{14}-\beta_{11})$
Oct;$\bar{y}^*_o+(\beta_{24}-\beta_{11})$;$\bar{y}^*_o$;$\bar{y}^*_o+(\beta_{12}-\beta_{11})$;$\bar{y}^*_o+(\beta_{13}-\beta_{11})$

Where $\bar{y}^* \equiv \bar{y} + \beta_1$

What we can identify is the fixed recall error of all but one of the recall groups relative to that of the one. This is illustrated in the bottom panel of Table 2. In essence if we take one month recall as our 'norm' (i.e. $y^* \equiv y + \beta(11)$) then we can see how reports involving more than one month's recall differ from this norm.[2] While this is less than we would like, it is enough for our immediate purpose of testing for the existence of recall error.

While we could investigate the significance of recall group bias simply by performing ANOVAs of mean reports, a more efficient and flexible alternative procedure is available which allows some investigation of recall error variances. Since we do not want to attribute differences in levels or variances to recall groups which are due to differences in the systematic portion (Δ) we first formulate a traditional human capital model of labor earnings which is appropriate for the situation in which levels in some months are zero According to this model earnings are determined by the level of individual investments in human capital. The two principal forms of these investments are formal education and on-the-job learning which is generally measure by years of experience. Thus:

$$\Delta_j = F(Ed_j, Exp_j) \simeq \sum_{i=0}^{k} (\delta_{Ei}\, Ed_j^i + \delta_{xi}\, Exp_j^i) \tag{3}$$

where Ed(j) and Exp(j) individual j's years of formal education and experience on the job, measured as deviations from average values, and the δ's are structural parameters relating earnings to human capital investments.

Of course, there is a very large number of other factors which will affect any given person's earnings, but, for the most part, the importance of each of these other factors, taken individually, is small. One important exception to this is race, the effect of which on earnings is far from negligible. All other factors can be collapsed into a single stochastic error term ψ_i. It is often argued that, as a result of the large number of excluded factors and the central limit theorem, this error term can be assumed to be normally distributed. It is also generally assumed that ψ is uncorrelated with earnings, education, experience and race and that its variance is constant.

With these assumptions the behavioral model becomes:

$$\Delta_j \simeq \sum_{i=0}^{k} (\delta_{Ei}\, Ed_j^i + \delta_{xi}\, Exp_j^i) + \delta_R\, Race_j + \psi_j \tag{4}$$

$$= \Delta x_j + \psi_j$$

where Δ is the vector of structural parameters, and x(j) is the vector of powers of individual j's education and experience, and his race.

The combined measurement and behavioral model is obtained by substituting Δ from equation 4) into equation 1) to yield the following additive model of recall error:

$$y_{mj} = \beta_{cr} + \Delta x_i + \psi_i + \tilde{\epsilon}_{mj} \tag{5}$$

In addition to the assumptions identified above, identification requires that the measurement errors ϵ be uncorrelated with the behavioral errors (ψ), the determinants of earnings (x), and actual earnings.[3]

Section II
Data and Estimation

The model was estimated using data for prime-age (25–55 years old) males who had at least two months with some employment in the first three waves of the 1983 SIPP panel. In all there were approximately 6300 such individuals in rotation groups one through three.[4] In order to eliminate confounding effects of proxy respondents, however, the sample was limited to those men who provided their own reports in each of the three interviews. This apparently innocuous restriction resulted in approximately seventy percent of the cases being eliminated from our sample. Finally, roughly ten percent of the cases with imputations on wage and salary items, as well as cases with self-employed income, in any one month were filtered from the sample. The result of these eliminations is a rather special subsample of the population which is of a quite manageable size (1378 cases). Since it is, to a certain extent a 'self-selected' subsample,[5] inferences to the overall prime-age male workforce should be quite guarded. Nevertheless, unless there are different mechanisms operating in the various rotation groups which determine self- versus proxy-reporting behaviors, the behavioral model should still be common to each rotation group,[6] and tests of the effects of recall and conditioning should remain valid. Indeed, if there are systematic difference in the selection mechanism then they should show up as rotation group effects—the significance of which we will test in the following section.

Estimation was performed by comparing the product moment matrix implied by the model presented in equation 5) with the actual product moment matrix calculated from the sample. The product moment matrix implied by the model is:

$$\Sigma \;=\; n \left[\begin{array}{c:c} [\Gamma X]'X\Gamma + \Psi + \sigma_\epsilon & X'X\Gamma \\ \hdashline \Gamma X'X & X'X \end{array}\right] \tag{6}$$

where $\Gamma' = [\Delta|\beta]$ and $X' = [x|R]$. The submatrix R is a (9x3) matrix composed of dummy variables for rotation group membership in each of the nine months.

The concentrated log-likelihood of the model given the sample is:

$$L = \mathrm{Log}|\Sigma| + \mathrm{tr}(S\Sigma^{-1}) - \log|\Sigma| - C - 1 \tag{7}$$

where C is the rank of S and Σ. We should note that with R containing dummy variables for all three rotation groups in our sample, X matrix is singular. We *must*, therefore, impose constraints on the β's to obtain unique solutions. The constraints we choose are those cross-month constraints which allow us to examine the hypotheses of recall bias, conditioning, and, alternatively, rotation group bias.

Once the constraints on the β are imposed, the above function can be minimized with respect to the β, δ, ψ and $\sigma(\epsilon)$ to yield full information maximum likelihood estimates. The various hypotheses regarding recall and conditioning biases can be tested by performing likelihood ratio tests using the minimum value of equation (7) under the alternative sets of constraints placed on the β's.

In actual practice we use the LISREL algorithm of Jöreskog and Sörbom (1976) to perform the estimation. Sufficient statistics for this consist of the (weighted) means and covariances of the sample. This formulation of the model is especially convenient for testing the various hypotheses regarding the form of the fixed response error. For testing hypotheses regarding error-variances, an alternative structure of the LISREL model is more convenient. This involves treating each recall group as a separate sample and estimating grouped systems of monthly data. The structural parameters (except the constant) are constrained to be equal across groups, while the relative measurement error variances are allowed to vary across groups.

Section III
Empirical Results

Table 3 presents the estimates of the mean difference in reporting bias in each recall/TIS group from that of the first recall and TIS group (i.e. $\beta(jk) - \beta(11)$) under various hypotheses. In addition to these estimated relative bias estimates (and their standard errors computed under the assumption of simple random sampling), the table provides the value of the likelihood function (and relevant degrees of freedom) from which likelihood ratio tests of the hypotheses can be performed. The first row of results in Table 4 refer to the hypothesis that length of recall is the only factor affecting reporting bias. Implementing this hypothesis involves relaxing three of the original 27 over-identifying restrictions incorporated in the model. When this is done twice the value of the likelihood function declines from its fully restricted value of 28.9 (not shown) by eight and one half units. Since twice the value of the likelihood function is distributed χ-square with degrees of freedom equal to the number of over-identifying restrictions, it is apparent that significant improvements in the goodness of fit are accomplished by allowing for recall-bias effects. Further significant improvements in goodness-of-fit are obtained when both time-in-sample and length of recall are allowed (see column 2). We accomplish this by permitting the effects of the recall group membership to vary from one TIS to the next. In all, this involves relaxing ten of the original 27 over-identifying restrictions, and results in a 19.4 unit decrease in the value of the fitting function.

While the recall and TIS effects are significant as a group, the estimated individual coefficients are sufficiently imprecise as to make it difficult to interpret their pattern. Only the coefficient on four-months recall when wave effects are allow is sufficiently large in relation to its estimated standard error to attain statistical significance. That it is significantly negative, is consistent with the type of memory model suggested by Sudman and Bradburn (1964) in which recall errors are assumed to stem from, in our application, a tendency for respondents to fail to recall more distant paychecks or to mis-place the occurrence of these payments in time. The Sudman and Bradburn model, however, would suggest a pattern of reported levels which would decline monotonically with length of recall. The point estimates in Table 3, if plotted against length of recall, would provide a very distinct impression of a 'saw-tooth' pattern. This pattern may reflect a tendency for respondents in the various recall groups to systematically misplace weekly paychecks from one month to the next.

An alternative hypothesis which is closely related to the recall error hypothesis, however, is that rotation group membership, *itself*, is associated with the response errors. A comparison of the values in the third row of Table 3 with those of the first indicates that we can reject this alternative interpretation of data. The χ-square statistic associated with the rotation group hypothesis is nearly four units larger than that associated with the recall error hypothesis yet involves the same number of over identifying restrictions.[7] Furthermore, relaxing the restrictions necessary to implement the former hypothesis does not significantly improve the overall goodness of fit, whereas relaxing those associated with the recall error hypothesis does.

Table 3
Estimated Biases Various Reporting Bias Hypotheses

	Hypothesis	
	Recall Bias Only	Recall and TIS Bias
$\beta_{12} - \beta_{11}$	-7.61 (10.42)	20.21 (30.28)
$\beta_{13} - \beta_{11}$	9.01 (10.05)	37.32 (55.15)
$\beta_{14} - \beta_{11}$	-19.17 (10.07)	- (-)
$\beta_{21} - \beta_{11}$	0 (-)	-32.83 (100.09)
$\beta_{22} - \beta_{11}$	-7.61 (10.42)	-52.67 (75.98)
$\beta_{23} - \beta_{11}$	9.01 (10.05)	-39.17 (52.47)
$\beta_{24} - \beta_{11}$	-19.17 (10.07)	-61.40* (30.23)
$\beta_{31} - \beta_{11}$	0 (-)	66.53 (51.72)
$\beta_{32} - \beta_{11}$	-7.61 (10.42)	26.01 (58.99)
$\beta_{33} - \beta_{11}$	9.01 (10.05)	39.87 (68.38)
$\beta_{34} - \beta_{11}$	-19.17 (10.07)	-3.36 (80.61)
χ-Square	20.4	9.4
d.f.	24	17

SRS Standard Errors in Parentheses.
Fully restricted χ-square: 28.9 (d.f.=27).
χ-square for Rotation bias only: 24.3 (d.f.=24). χ-square for Rotation and Wave bias: 11.1 (d.f.=19).

The preceding analysis suggests that there is significant differential bias resulting from length of recall and extent of conditioning in the SIPP reports of monthly earnings Furthermore, this differential bias is above and beyond that which can be accounted for by differences in subsample education, age and race, which are 'controlled' in our analysis. This analysis does not, however, shed light on the relative importance of differential bias and differential error-variance. In order to address this question it is necessary to analyze each recall group as a separate system of equations and impose the common structure of the behavioral model across these systems. Various hypotheses can then be incorporated in this group of systems by altering the cross-group constraints.

Table 4 presents the results of such analyses performed on the September, January and May earnings data for the same subsample employed in our preceding analysis.[8] Looking only at the first three rotation groups we are able to test hypotheses regarding the differential bias and error-variance for two and three months recall relative to one month recall.[9] The first row presents the value of the likelihood function obtained when neither differential bias nor error-variance are allowed. The second and third rows report the likelihood function values when bias only and bias and error-variance are allowed, respectively. As in the earlier analysis allowing for differential bias associated with length of recall and number of times in sample results in a relatively large and significant decrease in the value of the fitting function. This improvement in the goodness of fit, however, is not nearly as dramatic as that obtained when the restrictions that the error-variances are identical across recall and TIS groups are relaxed. Removing these over identifying restrictions results in a drop in the log-likelihood value of nearly fifty points (49.5). Since, under the null hypothesis that these improvements are due solely to chance, twice the decline in the log-likelihood is distributed χ-square with twelve degrees of freedom, we must conclude that differences in the reporting error variances across recall and TIS groups are quite important and extremely significant.

The pattern of relative recall-error variances for the various recall/TIS groups are quite interesting. For all three TIS groups, the rotation group associated with two months of recall has significantly lower estimated error variances than either the one or the three month recall groups. Although this initial decline in error-variance with recall seems quite peculiar, it is consistent with certain models of telescoping, such as Sudman and Bradburn's, based on "Weber's law of perceived time", according to which errors in the placement of events in time increase logarithmically with elapsed time. If the reporting process calls first for bounding the calendar month in perceived time and then summing the individual recalled paychecks received in that perceived period, then the first recall period will be longer in actual elapsed time than subsequent periods. If all people telescope at the same rate, are paid on the same schedules, do not forget entire paychecks, and are interviewed at the same time then this process would result in monotonically declining biases but relatively stable error-variances. To the extent that people do vary with respect to telescoping rates, pay schedules, and interviewing dates, however, then error variances can be expected to decline with recall length so long as the basic logarithmic pattern of telescoping errors holds.

Eventually, according to these models, telescoping errors will be overwhelmed by errors of omissions which increase monotonically with length of recall, and we should see a reversal in the direction of the pattern of biases and variances over time—something which seems to hold in the bottom panel of Table 4.[10]

Table 4
Estimated Relative Error Variances

	Goodness of Fit		
	χ-Square	d.f.	
Fully Restricted	130.5	48	
Bias Only	118.4	42	
Bias and Variance	19.4	30	
	Sep	Jan	May
$\sigma^2_{\epsilon(12)} - \sigma^2_{\epsilon(11)}$	-2798** (978)	-7775** (1222)	-4779** (1041)
$\sigma^2_{\epsilon(12)} - \sigma^2_{\epsilon(11)}$	-1758 (1016)	-4277** (1347)	-564 (1217)

Conclusions

In this initial analysis of SIPP earnings data we have found evidence of significant differential reporting bias and error-variance associated with length of recall and extent of respondent conditioning. The alternative hypothesis that data quality is a function of rotation group membership *itself* is not supported by the data.

For the particular model, subsample and measures we have examined, the statistical importance of the differential relative error-variances is much greater than that of the differential relative biases. Unlike our earlier work with the PSID Validity Study (see Duncan and Hill, 1985), however, we can not say reporting-error variance is in an absolute (nor even a mean-squared-error) sense more important than reporting bias.

The implication of these findings is that efficient estimation of monthly population parameters from the SIPP will require some corrections for the differential quality of the data from the various rotation/recall groups. Because there is evidence of significant differential bias as well as differential error variance, a proper treatment of the problem may require obtaining validating data for the SIPP instrument.

References

Duncan, Greg J. and Daniel H. Hill, (1985) "An Investigation of the Extent and Consequences of Measurement Error in Labor Economic Survey Data", *Journal of Labor Economics*.

Jöreskog, Karl, and Dag Sörbom (1976), *LISREL III: Estimation of Structural Equation Systems by Maximum Likelihood Methods*, (Chicago: National Educational Resources).

O'Muircheartaigh, Colm, (1986) "A General Model of Response Errors", *Response Variance in Surveys*, (Chicago: American Statistical Association Tutorial, August 1986).

Sudman, Seymour and Bradburn, Norman (1964), *Response Effects in Surveys*, (Chicago: Aldine Publishing Co.)

Footnotes

[1]The author would like to thank Bob Groves, Graham Kalton and Dan Kasperzyck for their helpful comments on earlier drafts of this paper.

[2]One month recall is equivalent to the CPS methodology.

[3]While these assumptions are frequently made in measurement error models, the latter two are particularly unfortunate in light of our previous findings the reporting error of annual earnings is negatively correlated both with actual earnings and experience (See Duncan and Hill, 1985).

[4]Rotation group four is omitted in the present analysis because the Wave 3 questionnaire was never administered.

[5]By this we mean that the individual's own behavior determines, in part, whether he is available at the time of the interview or whether someone else must report as a proxy for him.

[6]Initial selectivity bias analyses show only negligible differences between rotation groups in the self/proxy probit estimates. The dominant determinants in these models are relation to reference person, type of family, marital status, and employment status.

[7]A slightly different answer suggests itself when the combined hypotheses of recall and conditioning are contrasted with rotation group and wave hypotheses. In this case we can not reject one of the joint hypotheses in favor of the other. The fact that the wave and TIS hypotheses are operationally equivalent means that these joint hypotheses are, perhaps, too closely related to be differentiable.

[8]We limit our attention to this relatively small subset of months because the computational costs of examining groups of large systems is beyond our current resources.

[9]With these three rotation groups we can also estimate the differential bias and error variance of three and four months recall relative to two month recall using August, December and April.

[10]An alternative hypothesis is that the observed pattern of variances is not really related to length of recall so much as it is to rotation group membership. We might be able to test this alternative by repeating the analysis for October of 1983 and February 1984 when reports from rotation group 2 involve only one month's recall and those of rotation group 1 involve four months of recall. If the Sudman-Bradburn recall model is generating the data then we would expect to observe the same 'v' shaped pattern of estimated response variances. Unfortunately, we have not yet been able to develop a formal comparative test of these alternative models.