IMPLICATIONS OF SIPP RECORD CHECK RESULTS FOR MEASUREMENT PRINCIPLES AND PRACTICE

Kent H. Marquis, Jeffrey C. Moore and Vicki J. Huggins [1]
Census Bureau, Washington DC 20233

KEYWORDS: Response Error, Forgetting, Memory Decay, Telescoping, Errors-in-Variables, Cognitive Strategies

1. INTRODUCTION

Measurement errors in household surveys are inevitable. Yet they are not always well understood and, if one wanted to do something about them, it is not always clear what to do. This is especially true for the Survey of Income and Program Participation (SIPP), a new government survey to provide policy planners with detailed longitudinal information about the economic circumstances of families and people in the United States. Our goals in this paper are to contribute to understanding SIPP response errors and to begin considering what to do about them.

We report results of a record check study covering 8 months of reported participation in 8 programs in 4 states. While response errors are rare, they have important distorting effects on estimates of means and correlations, especially when the estimates involve a measure of change in participation status.

We discuss reasons for why the errors occur, statistical strategies for minimizing their effects, and survey design strategies for averting their occurrence. We conclude by recommending an expanded use of administrative records by SIPP and research on procedural changes to control the measurement errors.

2. METHODS

Using a full design (see Marquis, 1978), the record check compares information from the administrative records of 8 programs with participation reported by SIPP households in 4 states: Florida, New York, Pennsylvania and Wisconsin. The programs are:

AFDC,	Aid to Families with Dependent Children
CSRET,	Federal Civil Service Retirement
FOOD,	Food Stamps
OASDI,	Social Security
SSI,	Supplemental Security Income
UNEM,	Unemployment Insurance
VETS,	Veterans Pensions and Compensation
WORK,	Workers' Compensation

We used the Census Bureau's computerized record linkage software (Jaro, 1989, LaPlant, 1989) to match SIPP interviews to administrative records, using variables such as social security number, name, address, and date of birth. We include responses from 2 consecutive interviews, each covering 4 months of program participation. [2]

We estimate errors in reports of program participation (receiving benefits from a program). The response error scores are derived by comparing responses from SIPP to the true values from administrative records. We discuss several kinds of response error, all defined from the 2 x 2 table in Figure 2.1. The letters in the table represent frequencies of reported and true characteristics. N is the sample size.

PARTICIPATION

REPORT-ED	TRUE YES	NO	
YES	a	b	
NO	c	d	
	a + c	b + d	N

Figure 2.1: Notation for cross-classified reported and true values.

The total number of WRONG ANSWERS (or misclassification errors) for a program is b + c. The rate of misclassification is (b + c) / N and the misclassification percent (or percent wrong) is [(b + c) / N] x 100.

The frequency of UNDERREPORT errors is c. The underreporting error rate, which is conditional on a true positive, is c / (a + c), and the percent of underreporting errors is 100 times the rate.

Similarly, the frequency of OVERREPORT errors is b, the rate is b / (b + d), and the percent is 100 times the rate.

For each program, we usually calculate descriptive statistics (e.g., percent wrong) for each month and report an average over the entire eight months (or other groups of time periods such as wave 1 and wave 2). Unless we say otherwise, the inferential statistics refer to these averages. [3]

We call the effect of response errors on a parameter estimate a bias. The bias is the difference between the parameter estimated with data containing response errors and the true parameter value. We will examine two kinds of parameter estimates, a mean and a correlation. The bias in the estimated mean is [(a + b) / N] - [(a + c) / N] or (b - c) / N. Dividing by (a + c) / N yields the percent bias. Our correlation bias estimate assumes (1) a particular measurement model for the participation variable and (2) that the other variable is measured without error (see Marquis and Moore, 1990, appendix for details). This is so we can show the pure biasing effect of the measurement error in the participation variable.

3. DESCRIPTIVE RESULTS

The average misclassification error percentages for monthly

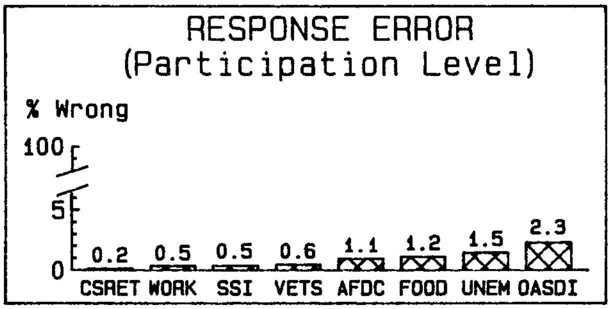

Figure 3.1: Average response error percentages for program participation level are very low.

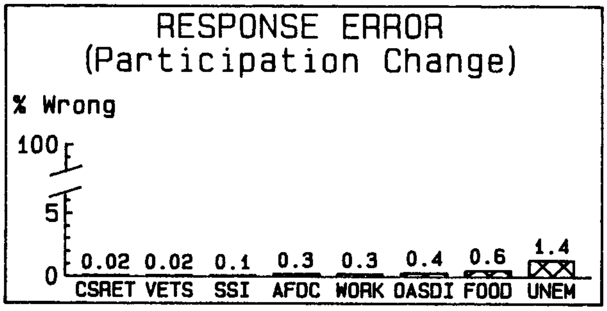

Figure 3.2: Average response error percentages for participation change are also very low.

reports of participation level and change in each of the 8 programs (% wrong) are very low. For participation level, in Figure 3.1, the averages range from 0.2 percent wrong for the CSRET program to 2.3 percent for OASDI. For participation change, in Figure 3.2, the range is even lower: from 0.02 for CSRET to 1.4 for UNEM. Thus, almost all respondents report participation status in each of the tested programs accurately almost all of the time.

3.2 Effects of Response Errors on Estimates

To see how these low response error percents impact the uses of the data, we look at biases in two kinds of estimates: the mean and the correlation. The mean estimate could be something like the proportion of the sample enrolled in the Food Stamps program in the month of June. The correlation estimate could be between employment status and participation in the Food Stamp program last month.

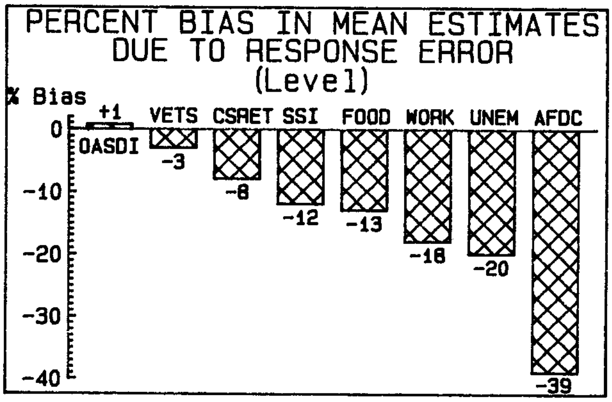

Figure 3.3: Response errors usually bias estimates of program participation levels in a negative direction.

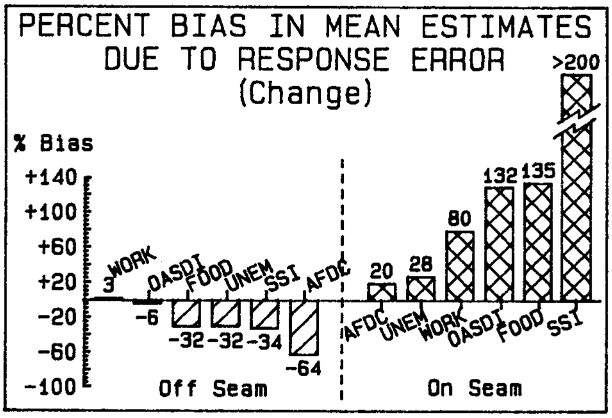

Figure 3.4: The sign of the change bias depends on whether change is measured on or off the seam.

For estimates of the mean participation level, the percent bias is usually negative, indicating that the estimated mean is usually lower than the true mean (Figure 3.3). Biases for some programs are substantial, such as the 18 percent underestimate of the WORK participation mean and the 39 percent underestimate of the AFDC mean. Biases for other programs are low, such as the 3 percent underestimate for the VETS program or the 1 percent overestimate for the OASDI program.

One problem that has been haunting SIPP for several years is that higher rates of change are measured between interviews compared to within an interview. A change between the two interviews is called an on-seam change. Change in any other pair of adjacent months is an off-seam change. Record check results in Figure 3.4 show that almost all of the off-seam biases are negative and all of the on-seam biases are positive. Too few program changes are measured off-seam and too many are estimated for the on-seam months.

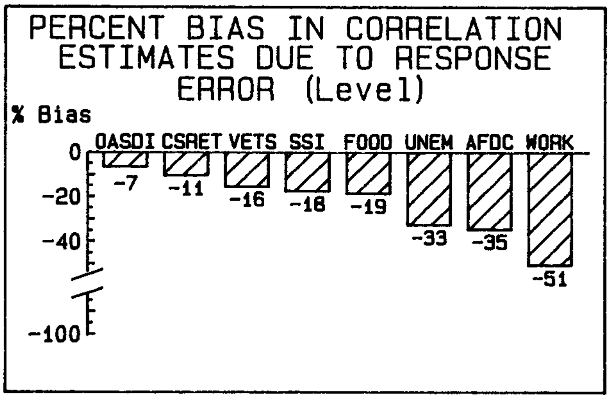

Figure 3.5: For measures of level, biases in estimated correlations due to response errors are small for some programs and quite serious for others.

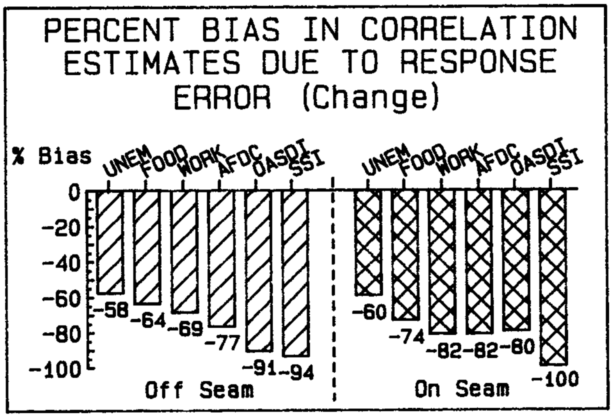

Figure 3.6: For measures of change, correlation biases are consistently large, regardless of whether they are measured on or off the seam.

Estimates of association may be a more typical use of SIPP data than estimates of means. In Figure 3.5, for participation level, we observe small to moderate percentages of bias in the correlation estimates for five programs and a major attenuation in the correlation estimates for 3 of the programs. In Figure 3.6, for participation change, we see that estimates are all substantially biased and that the bias does not depend on whether the change is on or off seam [4].

We have shown that while response errors occur at very low rates, they can have large effects on the kinds of estimates that analysts want to make from SIPP data. We will look at traditional models and assumptions about why people make response errors in order to understand the sources of response errors and to devise strategies to counteract or remove them.

4. CAUSES OF RESPONSE ERRORS

We would like to figure out the sources of the participation response errors so we can devise strategies to remove or counteract the causes. Here we examine characteristics of the error distributions, focusing on traditional models and assumptions about why people make response errors. We will examine both underreports and overreports for participation level.

A widely used approach to survey design uses assumptions from forgetting theory: that errors are mostly omissions or underreports, that underreporting gets worse as elapsed time increases, and that recall of recent events is accurate. The record check results, however, do not support forgetting theory. Although underreports dominate, all programs contain a relatively large number of overreports as well (Figure 4.1). If there were memory decay, we would see more underreporting of participation 4 months ago compared to last month. And the theory would also predict that the level of underreporting last month would be close to zero. Examining the results in Figure 4.2, we see that neither prediction is very accurate. For most programs underreporting rates are the same for recent and past events [5] and the underreporting rates for last month are often much larger than zero. Overreporting rates (not shown) generally do not differ between last month and 4 months ago either.

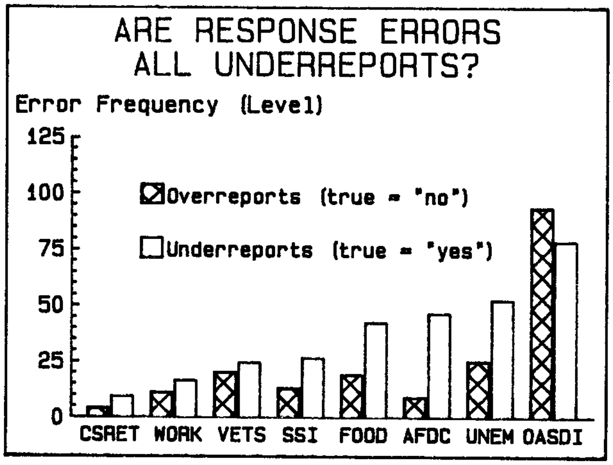

Figure 4.1: Although underreports usually predominate, all programs contain overreports as well.

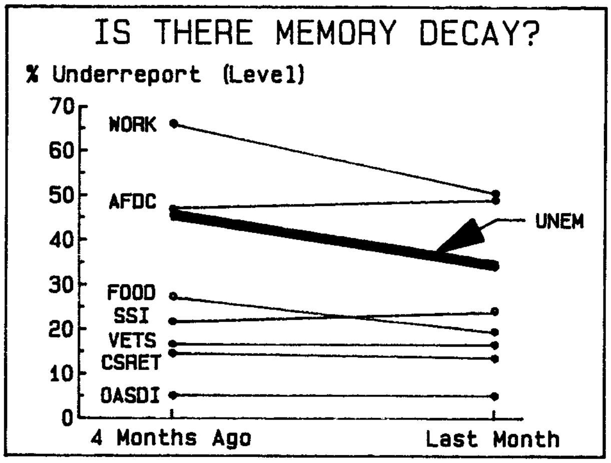

Figure 4.2: Participation underreports for "4 months ago" versus "last month" show little evidence of memory decay.

External telescoping errors are a concern to many panel survey designers. In the case of SIPP, respondents may telescope instances of past program participation into the first interview (wave 1) reference period because the interview is "unbounded." Since the interviewer reminds the wave 2 respondent what was reported in wave 1, it is unlikely that the respondent will report participation that truly happened in wave 1 incorrectly as occurring in wave 2. If external telescoping is responsible for SIPP response errors, we should see much more overreporting in wave 1 (unbounded) than in wave 2 (bounded). The results, in Figure 4.3, indicate that the wave-specific overreport rates do not differ significantly for any of the 8 programs. The trend for more overreporting in wave 2 is contrary to the hypothesis.

4.4 Additional Results

Other results mentioned in a recent research report (Marquis and Moore, 1990) include:
- In Pennsylvania, many respondents report AFDC benefits as General Assistance benefits.
- A small number of households confuse Social Security and Supplemental Security benefits.
- Some apparent errors in reporting Food Stamp recipiency are merely mistakes in reporting the correct official recipient in the household.

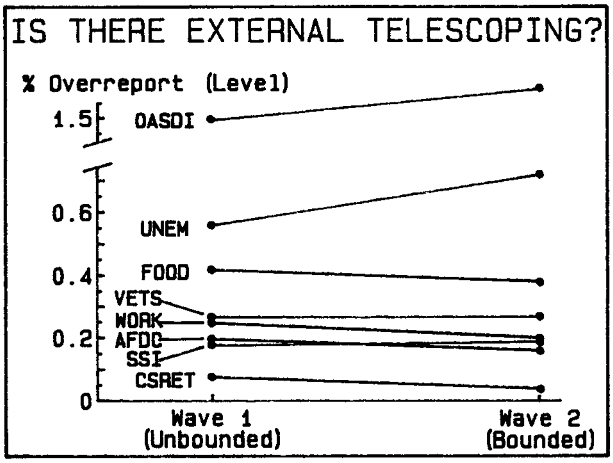

Figure 4.3: External telescoping into an unbounded (Wave 1) reference period does not explain observed overreporting error.

- Aside from the above three instances, a comprehensive search failed to reveal other instances of respondent confusion about program names or official recipient designations.
- Since average overreport rates are the same 4 months ago as last month, internal telescoping is not a major determinant of the observed overreporting errors.
- Respondents may learn to deliberately underreport Workers' Compensation and Unemployment Compensation participation because Wave 2 underreporting rates are higher than Wave 1 underreporting rates for those two programs.
- People did not increase their participation in the tested transfer programs in wave 2 compared to wave 1, so measurement did not result in a detectable behavior change.
- Interviewer effects were at the same low levels as found in most major surveys (one or two percent of total variance).
- In general, the directional error levels do not differ by self and proxy status although trends indicate more underreporting by proxy and perhaps more overreporting by self respondents.

To summarize, traditional hypotheses about the sources of survey response errors really do not account for the patterns of errors observed in the SIPP record check study.

5. STATISTICAL CORRECTION [6]

We mention strategies that might be used to correct for response errors providing we are comfortable making the required assumptions, can obtain the extra measurements required, and can demonstrate satisfactory performance in evaluation studies. We divide the discussion into procedures that correct individual responses directly and procedures that operate at the macro level.

5.1 Micro-Level Corrections

Edits: Respondents sometimes misreport the name of the program. Confusion has been observed between AFDC and general assistance (e.g., Klein and Vaughan, 1980) between means-tested and service-connected veterans programs (Vaughan, Lininger and Klein, 1983), and between Social Security and Supplemental Security (Vaughan, 1978). The edit approach to correcting program name confusion errors entails getting additional data via questionnaire (and/or from past and future waves concerning, for example, personal characteristics that determine program eligibility) and using logical "edit" rules to verify or reassign participation to the correct program.

Coder and Ruggles (1988), for example, have developed and evaluated a procedure to distinguish participation in AFDC from participation in the local general welfare program, and to remove cases that do not belong in either program. Macro-level evaluation for AFDC was favorable but it is not clear whether false positive errors were reduced and it is not clear whether there was an increase, decrease, or no change in the false negative errors for either AFDC or the other programs.

Raking Ratio Estimation: Huggins and Fay (1988) describe the use of Internal Revenue Service (IRS) data in connection with raking estimation procedures (e.g., Brackstone and Rao, 1976) to improve the quality of SIPP estimates subject to the effects of measurement and other errors. The technique works by adjusting the sample weights assigned to individual people to force consistency between sample estimates of marginals and corresponding population totals for cross classified variables. The procedure is analogous to the iterative proportional fitting algorithm for contingency tables, which yields maximum-likelihood estimates for hierarchical factorial log-linear models. For a sample of SIPP cases matched to IRS records, Huggins and Fay prepared "population" controls from the IRS data, implemented the estimation for selected SIPP characteristics, and analyzed the effects of the reweighting (which were quite favorable for person-level income). Their paper makes suggestions for further research using the procedure.

Administrative Records: It is possible to match records from SIPP questionnaires to appropriate administrative records and to substitute data of higher measurement quality into the individual questionnaire records. Nevertheless, administrative records are not routinely available from other agencies. If available, complete data are not ready immediately, and accurate matching delays availability further. Administrative records of high quality do not exist for all characteristics of interest to SIPP, and we do not yet have experience obtaining and using records from most states and many other federal programs. If it were possible to implement a comprehensive record check for a sample of the survey cases in a timely fashion, Marquis et al. (1981) discuss several ways of using such data to correct statistical estimates. We turn to the general question of adjusting statistical estimates next.

5.2 Macro-Level Corrections

Earlier we showed that measurement error will produce biased estimates of association such as a correlation. We mention two general ways to introduce corrections into estimates of relationships based on additional information about the measurement errors: variance-covariance matrix correction and instrumental variables.

Variance-Covariance Matrix Correction: Perhaps the most widely known procedure is to use a reinterview to learn about the variance of the measurement error distribution and introduce this information into a variance-covariance matrix before making a relationship estimate. Fuller and Hidiroglou (1978) present the general theory that has been implemented in SUPER CARP (Hidiroglou, Fuller and Hickman, 1980) and PC CARP (Fuller, 1986), computer software for estimation using survey data that contain measurement errors. Fuller (1987) further discusses the theory and an application for the case of labor force status classification (also generally applicable to SIPP program participation). [7] The procedure assumes a measurement error model and makes implicit multivariate assumptions. Although most applications of the variance-covariance adjustment approach use a reinterview, one might estimate the measurement error variances using other approaches, such as redundant questioning within a single interview (internal consistency), overlapping the reference periods covered by adjacent panel interviews or record checks on a sample of the survey observations.

Instrumental Variables: The instrumental variable correction strategy (e.g., Johnson, 1963; Fuller, 1987, Marquis et al. 1981) devotes additional measurement resources to measuring, constructing and using another variable which is assumed to be correlated with the variable of interest but uncorrelated with the measurement error in the variable of interest. The instrumental variable is used in a system of regression equations to produce asymptotically unbiased estimates of the regression parameters of interest, subject, of course, to the validity of the assumptions.

In practice, one cannot use the instrumental variables strategy with dichotomous variables since it is not possible to meet both critical assumptions simultaneously [8]. Nevertheless, the strategy might be useful in SIPP for analyses involving continuous variables, such as total income or dollar amounts of monthly program benefits.

To recap, methods exist to correct for measurement error in survey data. If SIPP deems a method's assumptions to be reasonable, then the method becomes a candidate for empirical evaluation, using research that obtains both the kinds of data needed to make the correction (e.g., reinterview data) and criterion data needed to evaluate the efficacy of the entire procedure (e.g., administrative record data).

6. SURVEY DESIGN CHANGES TO MINIMIZE ERRORS

Measurement errors are the result of human behavior. If we change the behavior we may change, or even prevent, the measurement errors. The survey design problem is to learn which survey conditions are producing the erroneous behavior so that the conditions can be changed. In parallel with the previous section, our goals here are to mention some of the important design remedies for response errors and to urge further consideration of them--in the form of research and implementation of suitable strategies.

6.1 Reducing Recall Difficulties

Shorter Reference Periods: Normally, surveys seek to use the shortest reference (or recall) periods possible to minimize the effects of memory decay. The memory assumptions are that recall is very good for recent time periods, that the errors are mostly false negatives, and that the false negatives increase as the length of the recall interval increases. As we showed above, however, the record check results suggest that these assumptions do not apply to SIPP participation reporting. Thus, we might not reduce errors importantly by shortening the reference period.

More Memory Cues: Memory retrieval can often be improved by furnishing additional cues to help the search process. For example, it is well known that a recognition task ("Have you seen this before?") is much more likely to get successful retrieval than unaided recall ("What did you see?"). But reducing false negatives may increase false positives (Marquis, Marshall, and Oskamp, 1972). Later we suggest that a better way to improve "recall" is to get respondents to use their records.

Dependent Interviewing: If the interviewer reminds the respondent of what was reported in the last interview, this may help anchor events in time. But positive benefits depend on the validity of at least 2 assumptions: That the information from the last interview is correct and that the procedures do not encourage remaking the same errors in the current interview. The lack of memory decay effects causes us to question the first assumption. Marquis and Moore (1989) show that the covariance of response errors declines slightly between interviews and suggest that dependent interviewing might have the undesirable effect of increasing the error covariance. Dependent interviewing, then, is unlikely to cause major reductions in SIPP response errors.

Respondent Rules: It is possible that some measurement errors arise because the interviewer does not interview the most knowledgeable person(s) in the household. It is assumed that the best information about a person comes from that person directly. What would happen if SIPP adopted a more stringent rule such as all self-responses? Marquis and Moore (1990) present a detailed analysis of errors made by self and proxy respondents for each of the 8 record checked programs. Although the data are not from an experimental design, the trends suggest that underreport error levels would be reduced only minimally--nowhere near zero--by an all-self-response rule, while misclassification and overreport error levels might actually increase. Below we imply that none of the household members possess the necessary understanding and skill to report correctly; to do so they need both training and restructured tasks. If true, a simple change in the respondent rule would not greatly affect measurement errors.

In sum, we have discussed a set of procedural changes that one might consider implementing in the presence of known response error. However, for the SIPP application, the assumptions underlying these procedures may not hold. We need to identify other procedural changes that are more likely to succeed.

6.2 Preliminary Cognitive Research Results

During the spring of 1989 professional staff members accompanied SIPP interviewers to nonsample households. They interrupted the interviews at appropriate places and used cognitive techniques to learn whatever the respondent could reveal about the answering processes. They wrote summaries of the important verbal interactions which they observed.

One of the main conclusions from the summaries is that many respondents adopt a simple heuristic or rule of thumb to quickly answer questions about recurring events in the four-month reference period (such as monthly income sources and amounts). Respondents use the simple rule as a substitute for detailed, direct recall and as a substitute for checking their personal records. A second general conclusion is that, while individual respondents sometimes had difficulties with particular questions, comprehension was not a pervasive, general problem. Instead of being caused by memory decay, forgetting, telescoping, deliberate lying, and the like, the measurement errors may be caused by trying to reconstruct a complex past using too simple a rule. If the hypothesis is correct, then a different set of remedies may be appropriate. We suggest 3:

Household Planning: Inform respondents of our detailed data requirements and teach them how to meet them. At the start of a panel, interviewers should make sure that respondents understand what we want them to accomplish and how we want them to do it. Interviewers and respondents should work out who will do what, how to keep and interpret financial records and how to deal with topics for which no records exist.

Quality-Focused Interviewing: We need to reorient interviewers to the importance of response quality. They should be taught how to recognize the inappropriate response strategies and how to steer respondents toward better tactics. Supervisors should put much less emphasis on avoiding refusals, less emphasis on interviewing efficiency and more emphasis on communicating the data goals and methods of achieving them.

Questionnaire: Interviewers need to explain the goals of each section in the questionnaire. Questions need to be reorganized and reworded to be consistent with the section's goals. To improve recall of when things happen, make more use of landmark events in the respondent's life. Both interviewers and respondents should be allowed flexibility in how they approach each section.

7. CONCLUSIONS AND IMPLICATIONS

We have described the results of record check research for SIPP which has yielded information about measurement errors in reports of program participation level and change. When we looked at how these measurement errors might affect statistical estimates, we learned that the effects could be considerable, both for estimates of means and for estimates of correlations. This conclusion prompted us to go on to review strategies for mitigating the effects of the measurement errors on estimation. Our considerations included both statistical correction strategies, that attempt to correct existing errors, and design alteration strategies, that attempt to prevent errors from occurring initially.

Contemporary quality management theory (e.g. Deming, 1982, Crosby, 1984, Juran, 1988), recommends constantly monitoring the quality of the data product, making after-the-fact corrections as necessary, and constantly improving the process design to eliminate measurement errors before they occur. Monitoring quality involves learning as much as possible about the errors, both in a descriptive sense for statistical correction strategies, and in a causal sense for improving the process.

SIPP is in a unique and advantageous position to adopt modern quality assurance procedures because it potentially can use administrative record data to regularly monitor the quality of its priority measurements--program participation and income. For other surveys, high quality administrative record data are not always available.

Administrative record data can also serve an important function in research to develop and evaluate statistical procedures to correct for errors. There are a number of possible correction procedures to be adapted and evaluated. And the evaluation concerns not only the quality of the corrections each strategy produces, but also the quality of its assumptions and the ability of the error measurement procedure that it relies on (e.g., reinterview) to yield correct descriptions of the error characteristics of interest. Beyond this, administrative record data can be very helpful in evaluating selected design features, such as the length of the recall period and the respondent rule.

So our view of the implications of the record check results for measurement principles and practice should now be clear:

1. Measurement errors can cause quality problems for survey data products.
2. Survey designs need to be expanded to include measures of the measurement errors.
3. Survey procedures need to include techniques to correct for measurement errors and to alter the processes that produce the errors.
4. The monitoring, product correcting, and process changing need to be a continuous, high priority part of the entire survey operation.

8. REFERENCES

ABOWD, J. and ZELLNER, A. (1985), "Application of Adjustment Techniques to U.S. Gross Flow Data," Proceedings of the Conference on Gross Flows in Labor Force Statistics, U.S. Department of Commerce and U.S. Department of Labor, Washington, pp. 45-61.

BRACKSTONE, G. and RAO, J. (1976), "Raking Ratio Estimators," Survey Methodology, Vol. 2, pp. 63-69.

BURKHEAD, D. and CODER, J. (1985), "Gross Changes in Income Recipiency from the Survey of Income and Program Participation," Proceedings of the Social Statistics Section, American Statistical Association, pp. 351-356.

CODER, J. and RUGGLES, P. (1988), "Welfare Recipiency as Observed in the SIPP," Survey of Income and Program Participation Working Paper No. 8818, U.S. Census Bureau, Washington, DC.

CROSBY, P. (1984), Quality Without Tears, McGraw-Hill, NY.

DEMING, W. E. (1982), Quality, Productivity and Competitive Position, M.I.T. Center for Advanced Engineering Study, Cambridge.

FELLEGI, I. and SUNTER, A. (1969), "A Theory for Record Linkage," Journal of the American Statistical Association, Vol. 64, pp. 1183-1210.

FULLER, W. (1986), PC CARP, Statistical Laboratory, Iowa State University, Ames.

FULLER, W. (1987), Measurement Error Models, Wiley, New York.

FULLER, W. and CHUA, T. (1985), "Gross Change Estimation in the Presence of Response Error," Proceedings of the Conference on Gross Flows in Labor Force Statistics, U.S. Department of Commerce and U.S. Department of Labor, Washington, pp. 65-77.

FULLER, W. and HIDIROGLOU, M. (1978), "Regression Estimates After Correcting for Attenuation," Journal of the American Statistical Association, Vol. 73, pp. 99-104.

HIDIROGLOU, M., FULLER, W. and HICKMAN, R. (1980), SUPER CARP, Survey Section, Statistical Laboratory, Iowa State University, Ames.

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

HUGGINS, V. and FAY, R. (1988), "Use of Administrative Data in SIPP Longitudinal Estimation," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 354-359.

JARO, M. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," Journal of the American Statistical Association, Vol. 84, pp. 414-420.

JOHNSON, J. (1963), Econometric Methods, McGraw-Hill, New York.

JURAN, J. (1988), Juran on Planning for Quality, The Free Press, New York.

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

LaPLANT, W. (1989), "Users' Manual for the Generalized Record Linkage Program Generator," Statistical Research Division, U.S. Census Bureau, Washington, DC.

LEMAITRE, G. (1988), "The Measurement and Analysis of Gross Flows," Labour and Household Surveys Analysis Division Staff Report, Statistics Canada, Ottawa.

LOFTUS, E. and MARBURGER, W. (1983), "Since the Eruption of Mt. St. Helens, Has Anyone Beaten You Up? Improving the Accuracy of Retrospective Reports with Landmark Events," Memory and Cognition, Vol. 11, pp. 114-120.

MARQUIS, K. (1978), "Inferring Health Interview Response Bias from Imperfect Record Checks," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 265-270.

MARQUIS, K., DUAN, N., MARQUIS, S. and POLICH, M. (1981), Response Errors in Sensitive Topic Surveys: Estimates, Effects, and Correction Options, R-2710/2-HHS, the RAND Corp., Santa Monica, CA.

MARQUIS, K., MARSHALL J. and OSKAMP, S. (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 MOORE, J. (1989), "Response Errors in SIPP: Preliminary Results," Proceedings of the Fifth Annual Research Conference, Census Bureau, Washington DC, pp. 515-536.

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

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

MOORE, J. and MARQUIS, K. (1989), "Using Administrative Record Data to Evaluate the Quality of Survey Estimates." Survey Methodology, Vol. 15, 129-143.

NELSON, D., McMILLEN, D., and KASPRZYK, D. (1985), "An Overview of the Survey of Income and Program Participation, Update 1." SIPP Working Paper Series, No. 8401, Washington, DC: U.S. Bureau of the Census.

POTERBA, J. and SUMMERS, L. (1985), "Adjusting the Gross Changes Data: Implications for Labor Market Dynamics," Proceedings of the Conference on Gross Flows in Labor Force Statistics, U.S. Department of Commerce and U.S. Department of Labor, Washington, pp. 81-95.

VAUGHAN, D. (1978), "Errors in Reporting Supplemental Security Income Recipiency in A Pilot Household Survey," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 288-293.

VAUGHAN, D., LININGER, C. and KLEIN, R. (1983), "Differentiating Veterans' Pension and Compensation Income in the 1979 ISDP Panel," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 191-196.

YOUNG, N. (1989), "Wave Seam Effects in SIPP," Proceedings of the Section on Survey Research Methods, American Statistical Association, (Forthcoming).

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 Census Bureau.

[2] Our analyses exclude some sample persons as follows: a. About 2700 children under age 15 because this age group is not interviewed, b. about 350 adults who refused to report their Social Security number, c. about 500 adults for whom all 8 months of data are unavailable (due to deaths, moving, refusing, etc.) and d., from analyses involving the AFDC, FOOD, UNEM, and WORK programs, about 2700 New Yorkers because of unresolved issues concerning the quality or availability of administrative data from that state. For the Federally administered programs (CSRET, OASDI, SSI, VETS) the total number of sample persons in the analyses is about 7550, for the other programs, which are administered by the states, there are about 5200 persons in the analyses.

[3] For the hypothesis tests and other "within person" comparisons, most inferences are based on paired-comparison t-tests that take into account the correlation of the observations for each person over time. We reject the null hypothesis for p .05. We discuss other inferential procedures as they are used. For all of our inferential statistics we assume simple random sampling although the SIPP sample design is more complex than this. As a result, our population variance estimates and corresponding p-values are likely to be slightly underestimated for the individual monthly or program-specific analyses. However, we feel that our stated conclusions, based on consistent patterns across programs and time periods, would not change if we were to take the complex sample design into account in our variance estimates.

[4] We have omitted two programs, CSRET and VETS that had no true change in at least one pair of months.

[5] For each program, the analysis is based on all people who could have underreported (true participation = "yes") either "4 months ago" or "last month" in a wave. Significance testing is for each wave separately, taking account of the within-person correlation of observations over time where appropriate. We report the average underreport percent over waves in Figure 4.2. The t-value for the wave 2 UNEM difference is the only one exceeding 2.00. That difference is not significant if we take account of the design effect. Numbers of people included in these analyses, by program and wave are: AFDC=111,108 CSRET=69,69 FOOD=215,205 OASDI=1467,1499 SSI=118,121 UNEM=193,203 VETS=149,150 and WORK=42,34.

[6] This section is based, in part, on Marquis, Duan, Marquis, and Polich (1981, Part IV).

[7] For additional ideas about applications to the labor force status classification issue, see Abowd and Zellner (1985), Fuller and Chua (1985), Porterba and Summers (1985) and Lemaitre (1988).

[8] Assuming a "true score plus error" measurement model, the correlation between the true value and the response error is negative for dichotomous variables. Thus, a variable cannot be correlated with the true value and uncorrelated with the measurement error as required by the instrumental variables approach.