STATISTICAL POLICY
WORKING PAPER 33

Seminar on the

Funding Opportunity in Survey Research

Federal Committee on Statistical Methodology

Statistical Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget
October 2001

The Federal Committee on Statistical Methodology
(September 2001)

Members

Brian A. Harris-Kojetin, Chair, Office of Management and Budget

Susan W. Ahmed, Consumer Product Safety Commission

Wendy L. Alvey, Secretary, U.S. Census Bureau

Lynda Carlson, National Science Foundation

Cynthia Z.F. Clark, U.S. Census Bureau

Steven Cohen, Agency for Healthcare Research and Quality

Lawrence H. Cox, National Center for Health Statistics

Cathryn Dippo, Bureau of Labor Statistics

Zahava D. Doering, Smithsonian Institution

Robert E. Fay, U.S. Census Bureau

Ronald Fecso, National Science Foundation

Gerald Gates, U.S. Census Bureau

Barry Graubard, National Cancer Institute

William Iwig, National Agricultural Statistics Service

Daniel Kasprzyk, National Center for Education Statistics

Nancy J. Kirkendall, Energy Information Administration

Charles P. Pautler, Jr., Internal Revenue Service

Susan Schechter, Office of Management and Budget

Rolf R. Schmitt, Federal Highway Administration

Monroe G. Sirken, National Center for Health Statistics

Nancy L. Spruill, Department of Defense

Clyde Tucker, Bureau of Labor Statistics

Alan R. Tupek, U.S. Census Bureau

G. David Williamson, Centers for Disease Control and Prevention

Alvan O. Zarate, National Center for Health Statistics

Expert Consultant

Robert Groves, Joint Program in Survey Methodology

PREFACE

The Federal Committee on Statistical Methodology was organized by the Office of Management and Budget (OMB) in 1975 to investigate issues of data quality affecting Federal statistics. Members of the committee, selected by OMB on the basis of their individual expertise and interest in statistical methods, serve in a personal capacity rather than as agency representatives. The committee conducts its work through subcommittees that are organized to study particular issues. The subcommittees are open by invitation to Federal employees who wish to participate. Since 1978, 33 Statistical Policy Working Papers have been published under the auspices of the Committee.

Statistical Policy Working Paper 33 presents the proceedings of the "Seminar on the Funding Opportunity in Survey Research." We are indebted to all of our colleagues who assisted in organizing the seminar, and to the many individuals who not only presented papers but who also prepared these materials for publication.

Program on the Funding Opportunity in Survey Research Seminar

Bureau of Labor Statistics
2 Massachusetts Avenue, NE Washington, DC 20212

June 11, 2001

Welcoming Remarks

Welcoming Remarks to the FCSM Seminar On The Funding Opportunity In Survey Research June 11, 2001

Monroe G. Sirken
National Center for Health Statistics

Introduction

Good morning. As chair of the Research Subcommittee of the Federal Committee on Statistical Methodology (FCSM) that organized this meeting, I welcome all of you to this FCSM Seminar on the Funding Opportunity in Survey Research. Other members of the FCSM Subcommittee are Bob Fay, Alan Tupek; Nancy Kirkendall, and David Williamson.

If the venue of this Seminar could have been changed to Geneva, Switzerland, where Kathy Wallman is today attending a meeting of the Conference of European Statisticiains, she would be making these welcoming remarks. I am very sorry that she is not here or we are not there. Kathy is a prime architect of today's program, and I'm sure she regrets not being with us today.
.

This Seminar is being sponsored by the Survey Methodology and Data Collection Sections of the Washington Statistical Society (WSS), and the D.C. Chapter of the American Association for Public Opinion Research. Cynthia Clark, WSS president, is not here this morning, but will extend WSS greetings this afternoon after we return from lunch.

In 1998, the Funding Opportunity in Survey Research was established in the Methodology, Measurement, and Statistics Program, National Science Foundation (NSF). The Funding Opportunity is jointly funded by the NSF and a consortium of 13 Federal statistical agencies, with the FCSM Research Subcommittee serving as liaison between the two parties. The program awards grants to meritorious projects in statistical and survey research oriented to the needs of Federal agencies. It is vital to the program's success that findings of the research projects it supports are widely disseminated and discussed with the staffs of Federal agencies. Indeed, this

Seminar was organized to foster dialogues between the staffs of Federal agencies and the principal investigators of research projects being supported by the Funding Opportunity..

Agenda

This Seminar features the reports of four research projects that are currently being supported by the Funding Opportunity. Each report will be presented by the project's principal investigator, and discussed by staff of two Federal statistical agencies. These reports and discussants' remarks will be presented in sessions 2, 3, 4, and 5.. Sessions 2, 3 and 4 involve topics on the cognitive aspects of survey methods, and session 5 deals with a statistical topic:

Session 2. The Cognitive Basis For Seam Effects In Panel Surveys,
Session 3. A Computer Tool To Improve Questionnaire Design,
Session 4. Social Presence In Web Surveys, and
Session 5. A Unified Jackknife Theory In Small Area Estimation.

Also, this Seminar features two sessions about the Funding Opportunity in Survey Research whose name, I should note, was recently changed to the Funding Opportunity in Survey and Statistical Research to more appropriately reflect the broad scope of the program.. The present status and future prospects of the Funding Opportunity are discussed in sessions 1 and 6 respectively.

Session 1. The Funding Opportunity In Survey and Statistical Research
Session 6. The Future of the Funding Opportunity In Survey and Statistical Research

To further the Seminar's objective of fostering dialogues between principal investigators and staffs of Federal agencies, the final 15 minutes of each session is reserved for floor discussion.

Acknowledgements

I would be remiss if I failed to acknowledge the very generous support of many parties and people in organizing this Seminar.

The National Center for Health Statistics (NCHS) provided critical funding and logistical support for traveling out-of-town speakers and preparing the Seminar Proceedings. It is no exaggeration to say that this Seminar would not have been possible without the support of Edward Sondik, the NCHS Director.

The Washington Statistical Society was enormously helpful in disseminating information about the Seminar to the Washington community.

Royalties from publication of the Wiley book "Cognition and Survey Research" contributed the funding for refreshments at this Seminar. These royalties were originally assigned to a survey methods research fund at the National Foundation for the Centers for Disease Control and Prevention by the book's editors: Monroe Sirken, Douglas Herrmann, Susan Schechter, Norbert Schwarz, Judith Tanur and Roger Tourangeau.

Under a small contract from the NCHS, the Council of Professional Associations On Federal Statistics will collect the presentations made at this Seminar, and compile them into proceedings that will be published as a FCSM Statistical Policy Working Paper. Ed Spar is, with his usual thoroughness, overseeing this activity.

The Federal Committee on Statistical Methodology thanks all of these contributors. Also, the Committee thanks the chairpersons, presenters and discussants at this Seminar who willingly and graciously agreed to participate in this Seminar, and thanks Norman Bradburn for his help in the initial planning of the Seminar program.

Finally, personal thanks to Barbara Hetzler, NCHS, an invaluable administrative officer, for making the Seminar arrangements.

Session 1

The Funding Opportunity in Survey Research

The Funding Opportunity in Survey Research

Cheryl Eavey, National Science Foundation

In FY1999, NSF's Methodology, Measurement, and Statistics (MMS) Program, in collaboration with a consortium of federal statistical agencies represented by the Interagency Council on Statistical Policy (ICSP), conducted the first of three planned competitions for research on the development of new and innovative approaches to surveys. A total of $600,000 was set aside for successful proposals, $300,000 from the MMS Program and $300,000 from participating federal statistical agencies.[1] Priority was given to research proposals that were interdisciplinary in nature, had broad implications for the field in general, and had the greatest potential for creating fundamental knowledge of value to the Federal Statistical System.

NSF received twenty-eight proposals in response to its announcement. Proposals underwent a multi-tier review process that included both an outside experts panel and a panel of representatives from the participating federal statistical agencies. Of the twenty-eight proposals, six proposals (four projects) were selected for funding. The portfolio included work in small-area estimation, seam effects in panel surveys, the design of web surveys, and a computer tool that critiques survey questions. A more detailed list of the projects supported is available on the NSF home page at http://www.nsf.gov/sbe/ses/mms/survey99.htm.

Representatives of the participating agencies generally were pleased with the results of the first year's competition. The funded projects nicely blend fundamental research with some of the practical needs and concerns of the federal statistical agencies. One measure of the success of the FY1999 competition was the total amount contributed by the agencies. Collectively, the agencies exceeded their expected contribution by providing a total of $674,036 for the support of the four projects.

Another measure of success was the decision to continue the competition into its planned second and third years. NSF issued the announcement for Research on Survey and Statistical Methodology in FY2001. The announcement was similar to the FY1999 solicitation, except that the focus was expanded to include the development of methods for the analysis of survey data. As of late August 2001, the second year of the competition is near completion. NSF received sixteen proposals in response to the announcement and again anticipates funding four projects. All awards will be made by 1 October 2001. Total contributions from the agencies are expected to exceed the $600,000 originally pledged by NSF and the participating statistical agencies.

[1] In FY1999, participating federal statistical agencies included the Bureau of Economic Analysis, the Bureau of Justice Statistics, the Bureau of Labor Statistics, the Department of Transportation, the Division of Science Resource Statistics, the Economic Research Service, the Energy Information Administration, the National Agriculture Statistics Service, the National Center for Education Statistics, the National Center for Health Statistics, the Social Security Administration, and the U.S. Census Bureau.

The planned FY2002 competition will continue the second year's focus on survey and statistical methodology. The announcement, which is available on the NSF home page at http://www.nsf.gov/pubs/2000/nsf00147/nsf00147.htm, has a deadline for submission of proposals of 30 November 2001.

Session 2

Cognitive Basis for Seam Effect in Panel Surveys

SEAM EFFECTS FOR QUANTITATIVE AND QUALITATIVE FACTS

Lance J. Rips
Northwestern University

Frederick G. Conrad
Bureau of Labor Statistics

Scott S. Fricker
Bureau of Labor Statistics

Jennifer Behr
Yale University

SEAM EFFECTS FOR QUANTITATIVE AND QUALITATIVE FACTS[1]

The accuracy of answers to factual questions degrades over time. People's memory for an event becomes less accurate with the time since the event took place, so it's natural to expect the accuracy of survey responses that depend on such memories to decrease in the same way. If you ask about our income sources, health histories, or other biographical facts, you can probably expect better answers for last month's information than for that of the month before. Many studies of autobiographical memory document this decrease, though the rates of forgetting vary widely from one type of material to another (see Shum & Rips, 1999, and Tourangeau, Rips, & Rasinski, 2000, chaps. 3 and 4, for reviews of the memory literature as it bears on surveys). Of course, people aren't entirely at the mercy of memory, since they usually have ways of estimating or inferring information when memory fails. Response accuracy over time will then depend on the interplay of forgetting and its compensating strategies. Survey researchers face the job of understanding this interplay in order to estimate the true values of the information they seek.

The studies we report here explore a well-documented type of response error called the seam effect that occurs in panel surveys. The seam effect is time-dependent, since it exhibits a clear temporal profile, but its form is more complicated than a simple increase in errors over time. Our goal is to try to understand this effect by examining some of its components. The panel surveys in which the seam effect appears typically take place over a period of several years, which makes it difficult to study efficiently. We've therefore made use of a laboratory analog that produces seam effects and allows us to vary factors that might contribute to them. In this paper, we first describe the nature of the seam effect in actual surveys and the analog of the effect with which we've been working. Next, we briefly review earlier results using this method and then report two new experiments that extend these findings. Finally, we summarize our conclusions about the seam effect and possible ways to eliminate it.

Seam Effects in Panel Surveys

Seam effects occur in panel surveys that ask respondents about events from each of a series of subintervals within the survey's larger response periods. For example, the Survey of Income and Program Participation (SIPP) interviews respondents three times a year, but during an interview the respondents must report about income and employment for each of the past four months. We show this type of schedule in the upper panel of Figure 1. A respondent might be interviewed in May, for example, and provide answers during that interview about whether he or she received social security benefits during each of the months of January, February, March, and April. The same respondent would be re-interviewed in September and would then provide information about receiving social security benefits in each of May, June, July, and August; and so on.

The seam effect appears in plots of changes in the individual respondent's answers from one month to the next in this series. For example, we can count the number of times respondents changed their answer from "yes, I received social security benefits in January" to "no, I did not receive social security benefits in February" or the reverse change from "no" to "yes." If we then graph the total number of such month-to-month changes across the entire period of the survey, we get the type of scalloped profile that appears in the lower panel of Figure 1, which is taken from SIPP (Jabine, King, & Petroni, 1990). These data are month-to-month changes in reports of receiving social security and food stamps. For the respondent who is interviewed in May, September, and January, the first point on the x-axis corresponds to the change in response between January and February, the second point to the change between February and March, and so on for all pairs of adjacent months.

It's crucial that the change between months 1 and 2 depends on answers that come from the same interview (the May interview in our example), as does the change between months 2 and 3, and months 3 and 4. The change between months 4 and 5, however, is based on data from two separate interviews: Month 4 answers come from the first interview in this series (e.g., the May interview), whereas the month 5 answers come from the second interview (e.g., the September interview). Months 4 and 5 are on the "seam" between the response periods for the first two interviews, and these and other seam transitions appear in the figure at the positions of the dashed vertical lines. The seam effect is the finding that the number of changes at these seams is much greater than the number of changes between other pairs of adjacent months.[2]

A Laboratory Model of the Seam Effect

There is little doubt that the seam effect is due to the fact that data from the seam months come from two separate interviews while data from the nonseam months come from the same interview. We might say that the response period for the interviews (e.g., four months in SIPP) differs from the response period for the questions (one month), and this difference is responsible for the pattern in Figure 1. Similar differences in response periods appear in other surveys, such as the Consumer Expenditure Survey (CE). The issue for survey methodologists and cognitive psychologists is which factors associated with the change in interviews increase or decrease the size of the effect.

Two general explanations of the seam effect are possible, and the literature on the seam effect implicates both. On one hand, the effect might be due to factors that enhance the difference at the seam months. Suppose, for example, that respondents gradually forget information during the period between interviews, as seems likely. Then answers to questions about month 4, the most recent month queried during the first interview, will draw on respondents' relatively rich memory for the events; but answers to questions about month 5, the earliest month queried during the second interview, will draw on relatively impoverished memory. Forgetting for the incidents in question could therefore contribute to the size of the seam difference. On the other hand, the effect could also be due to factors that minimize differences across nonseam months. Respondents might be biased, for example, to give the same answer about each month during an interview. When asked whether they received social security benefits during each of January, February, March, and April, they answer "no" to all four questions (or "yes" to all four) as a way of simplifying their task. These constant-wave responses reduce the changes for nonseam months, making the changes at the seam stand out (Kalton & Miller, 1991; Young, 1989).

To analyze the seam effect, we need to know the facts that respondents are trying to report. Unless we have access to the correct answers, it's difficult to know whether the seam effect is due to exaggerated changes at the seam months, to suppressed changes at the nonseam months, or both. For this reason, we've designed a new procedure that is in some ways a cross between a field study and a laboratory task. Figure 2 illustrates the basic schedule for the experiment. In this procedure, we mail information to respondents each week for eight consecutive weeks. This information is embedded in a questionnaire that they fill out and mail back to us within 24 hours. Respondents come into the lab at the end of the fourth week and again at the end of the eighth week, and during these test sessions, we ask them to report on the information they had seen in the questionnaires during each of the preceding four weeks. These two test sessions are our analogs to the survey interviews in SIPP, CE, and other panel surveys, dividing the interval into two response periods. The questions that we ask during these sessions provide the week-to-week data that we need in order to study the seam differences. Changes in respondents' answers between weeks 4 and 5 are the seam changes, coming from two different test sessions. Changes in answers between the other pairs of neighboring weeks (1-2, 2-3, 3-4, 5-6, 6-7, and 7-8) are nonseam transitions, coming from the same test session. The time scale of the design is in weeks rather than months to allow us to study seam effects more efficiently.

Previous Results

Our earlier experiments demonstrated that we could use this design to produce seam effects and to alter their size. In one such experiment (Rips, Conrad, & Fricker, 2000, Experiment 1), the questionnaires we sent to respondents during weeks 1-8 asked them yes-or-no questions about common activities they might have participated in that week. The questionnaire for week 1, for example, asked, During the last week, did you have the oil changed in your car?, During the last week, did you order a pizza for home delivery?, along with similar items. Each of the eight questionnaires contained a total of 50 questions about the occurrence of everyday events, which we had selected on the basis of norms of rated frequency of occurrence, duration, importance, and affective impact from an earlier study. We composed the questionnaires so that one group of items appeared during weeks 1, 2, 7, and 8, and a separate group of items appeared during weeks 3, 4, 5, and 6. Thus, the questionnaires in seam weeks 4 and 5 were identical, apart from the random order of the items in the questionnaires.

During the two test sessions, we asked respondents to think back to the questionnaires they had filled out in the past four weeks and to decide whether certain items had appeared on those questionnaires. In the first test session, for example, we gave respondents a list of questions and asked them whether each question had appeared in the questionnaire for week 4 (e.g., On the questionnaire for week 4, did you see: [the item about having] oil changed in your car?, ...did you see: [the item about ordering] a pizza for home delivery?, etc.). We next re-presented the same questions in a new random order and asked the respondents whether each item had appeared in the questionnaire for week 3. We then repeated this procedure for week 2 and week 1. The procedure for the second test session was identical, except that we asked respondents about the content of the questionnaires for weeks 8, 7, 6, and 5 (in that order).

The main data from this experiment come from the test sessions: respondents' answers about whether they remembered seeing individual items on the weekly questionnaires. We can look directly at the changes in their answers to these items to see whether they exhibit a seam effect. But because we also know which items actually appeared on the questionnaires, we can also evaluate their responses for accuracy. In fact, the data from this study produced a seam effect: The largest percentage of changes occurred between seam weeks 4 and 5. (These are changes from "yes, I saw that item on the week 4 questionnaire" to "no, I didn't see that item on the week 5 questionnaire" or the reverse change.) The actual items that respondents had seen during these two weeks were exactly the same; so the increase in changes at this point is entirely due to response error. By contrast, true changes in the questionnaire items had occurred between weeks 2 and 3 and again between weeks 6 and 7; however, the data showed no increase in the percentage of changed responses at these two points. Thus, the results showed an increase in changed responses where there were no objective changes (between weeks 4 and 5), but no change in responses where there were objective changes (between weeks 2 and 3 and between weeks 6 and 7).

What is responsible for the form of these data? It seems likely that forgetting contributes to the effect. Respondents were reliably above chance in their ability to recognize items from the questionnaires they had seen just before the test sessions in weeks 4 and 8 (63.5% correct), but fell to near chance accuracy for earlier weeks (e.g., 52.6% correct for weeks 3 and 7, where 50% is chance recognition). There is some evidence, however, that constant-wave responding also contributed to the effect. In 19.8% of cases during the first test session, respondents made positive constant-wave responses, saying that they had seen an item in all four preceding weeks. Similarly, 10.5% made negative constant-wave responses, saying that they had not seen an item in any of the four preceding weeks. During the second test session, the comparable statistics are 27.4% and 9.9%. Thus, on about 30% of occasions in each test session, respondents were making constant-wave responses. Because each test item appeared in exactly two of the four weeks during a response period, these constant-wave responses were incorrect for two of these weeks. Forgetting could be responsible for the negative constant-wave cases: Respondents may simply have been unable to remember an item on any of the last four questionnaires. It's more difficult, however, to account for the more numerous positive constant-wave responses. Some additional bias in favor of "yes" responses must be at work here.[3] We show in an earlier paper (Rips, et al., 2000) that a theory based on a combination of forgetting and positive constant-wave responding can account for the detailed results from this experiment.

Study 1: Seam Effects for Biographical Material

The results from our preliminary study depend on a rather unusual type of question. We asked respondents whether they recalled items from questionnaires – for example, whether they remembered seeing an item on the questionnaire for week 3 about having the oil changed in their car. We asked questions like these because they gave us experimental control over the to-be-recalled information and allowed us to determine the accuracy of the respondents' answers. But, of course, it is also important to know whether the results generalize to items closer to those of actual surveys. We want our conclusions to apply to survey questions about personal information, not just to questions about questions about such information.

To extend our study in a more naturalistic direction, we've modified the basic procedure in Figure 2 to ask respondents during the test sessions about personal incidents. In this new experiment, we again sent respondents weekly questionnaires about ordinary activities, such as whether they had the oil changed in their car or whether they had checked out books from the library last week. As in the study just described, respondents received eight of these questionnaires in consecutive weeks. They filled out the questionnaires, checking off "yes" or "no" to each item, and mailed the questionnaires back to us within 24 hours. The questionnaires themselves were similar to those of the preliminary experiment, except that there was no change of items from one questionnaire to the next: Respondents saw the same set of items (e.g., During the last week, did you take a day off work due to illness?) on each of the questionnaires they filled out. The respondents were 56 adults (average age 50.6) whom we had recruited through advertisements in local newspapers.

The respondents also took part in two test sessions, again following the pattern of Figure 2. This time, however, we asked respondents about the actual incidents they had described earlier. In the first test session, we asked respondents, for example, whether they had taken a day off work due to illness during week 4, whether they had checked out a book from the library during week 4, and so on. There were 60 questions in all. Half these questions the respondents had answered in their earlier questionnaires, and half were new. We then asked the same set of questions about week 3, week 2, and finally week 1. We conducted the second test session in the same way, asking about weeks 8, 7, 6, and 5, in that order. The test sessions in this experiment, then, asked respondents directly about their own individual activities rather than about whether they had seen a questionnaire item about the activity. In this procedure, of course, we have no absolute knowledge of whether their answers to these questions were correct or incorrect, but data from the questionnaires can provide a partial check on accuracy. Since respondents filled out the questionnaires near the time the target events took place, answers on the questionnaires should be more accurate than answers to the same questions during the test sessions.

The design of this experiment gives us two ways to look at the week-to-week changes in respondents' answers. First, we can examine the transitions as they appear in the test sessions: The percentage of times that respondents said, during these sessions, that they had participated in an activity during week k but not during week $k + 1$ (or the opposite change). These data are analogous to those of the preliminary study and to those of the panel surveys, in that there are separate seam transitions (weeks 4-5, where the data come from different test sessions) and nonseam transitions (weeks 1-2, 2-3, 3-4, 5-6, 6-7, and 7-8, where the data come from the same test session). We plot these data as filled circles in Figure 3, and they show a modest seam effect, with a reliable difference in the number of changed responses from week to week. The second perspective on the changes comes from responses to exactly the same questions on the weekly questionnaires. These data appear in Figure 3 as open circles, and as we might expect, they show no increase in the percentage of changed responses for seam weeks. The procedure here is similar to a hypothetical panel survey that interviewed respondents monthly and asked during each interview about the preceding month alone. In the terms we introduced earlier, the results from the test sessions (filled circles) have different response periods for the interview (four weeks) and for the questions (one week), whereas the results from the questionnaires (open circles) have the same response period for both (one week in each case). The difference between the two curves in Figure 3 illustrates the effect of separating these response periods.

Once again, part of the seam effect is probably due to memory. If we gauge accuracy by the difference between a respondent's answer in the test session and the answer to the same question on the relevant weekly questionnaire, then we find that accuracy is generally quite high – overall, 86.2% of answers matched. Accuracy in the test sessions, however, decreased with time in a regular way. Accuracy was 91.2% for the events of weeks 4 and 8 (the most recent weeks in the two response periods) and declined to 82.4% for events from weeks 1 and 5 (the earliest weeks in the two periods). There is also evidence for constant-wave responding, but the number of incorrect constant-wave responses appears to be smaller than in our preliminary study. In the first test session, respondents made positive constant-wave responses for 16.2% of the items and made negative constant-wave responses for 54.8%. These figures are high, but they may simply reflect the true proportion of times the respondents had taken part in each of the activities during all four of the preceding weeks or had taken part during none of the preceding weeks. To check for bias, we can again compare these percentages to those for the weekly questionnaires. These data show that respondents had answered "yes" in all four questionnaires for 14.0% of the items and had answered "no" in all four questionnaires for 47.2% of items. Thus, constant-wave responding was only slightly (though reliably) more common in the test session than in the original questionnaires. Results were similar for the second test session: Respondents made positive constant-wave responses for 15.2% of items and negative constant-wave responses for 58.8%, only somewhat higher than the 13.3% positive and 53.8% negative constant-wave responses in the questionnaires.

This experiment suggests that we can detect seam effects for personal events using our procedure. The effect is smaller, though, than those of the preliminary experiment in which we used more artificial items. Respondents' memory for the personal events is much better than for the artificial ones, at least if memory is evaluated relative to answers on the earlier questionnaires. Even after four weeks, accuracy is quite good for the everyday events we used, and this may have decreased respondents' tendency to rely on constant-wave responding and other strategies that could increase the size of the seam effect.[4] This, of course, does not imply that seam effects will also be small in surveys that ask about personal events. The longer response periods of actual panel surveys may decrease memory, and the structure of the survey interview may increase constant-wave responding, as we are about to see. Moreover, some of the questions in panel surveys seek quantitative information rather than the sort of qualitative (yes/no) answers that we have looked at so far. Seam effects may be different when respondents have to come up with a number (e.g, the amount of a purchase or the amount received from a source of income) rather than simply deciding whether or not an event happened. We report one further study of quantitative responses before returning to implications for survey methods.

Study 2: Seam Effects for Quantitative Information

To study seam effects for quantitative information, we used a second variation on our standard procedure. In the new experiment, we again sent respondents weekly questionnaires, but for a period of six rather than eight weeks. Test sessions occurred at the end of weeks 3 and 6. The schedule was similar to that of Figure 2, then, but condensed by two weeks. The more important difference between the studies concerns the questions we asked in the questionnaires and test sessions. The questionnaire items were all of the type: During the last week..., did you (or someone in your household) spend more or less than $X on Y? Please circle either "More" or "Less" or "Did not purchase," (e.g., During the last week..., did you ...spend more or less than $2 on milk and cream from the grocery or convenience store? During the last week..., did you... spend more or less than $17 on electricity for your home?) We based these questionnaire items on ones that appear in CE. We asked an individual respondent about the same items (e.g., milk and cream, electricity, etc.) on each questionnaire. The specific amounts, however, varied for some items. For half the questionnaire items, we asked about the same dollar amount each week (e.g., respondents might be asked on each questionnaire whether they spent more or less than $2 for milk and cream that week); for the remaining questionnaire items, the amount changed from week to week.[0]

During the test sessions, we asked respondents to recall the dollar amounts they had seen on the questionnaires. For example, one item in the first test session was: On the third week's questionnaire, which you filled out on ..., when you were asked about milk and cream, what was the dollar amount you were asked about? Respondents' answers to these questions provided that data that we used to analyze the seam effects. In addition, we varied the way in which we grouped the questions during the tests. In earlier research (Rips et al., 2000), we had found larger seam effects when respondents had to answer all the question about a given topic one after another, and we were interested in determining whether the same would be true for the quantitative questions in this study. For this reason, half the respondents answered the test questions in an order blocked by item: In the first test session, these respondents answered the question about milk-and-cream for week 3, week 2, and week 1; then they were asked the question about electricity for weeks 3, 2, and 1, and so on. In the second test session, they answered the question about milk-and-cream for weeks 6, 5, and 4; then the question about electricity for weeks 6, 5, and 4; and so on through the full set of items. The remaining respondents answered the test questions in an order blocked by week: During the first test session, these respondents answered all the questions about week 3, then all the questions about week 2, then all the questions about week 1; in the second test session, they answered all the questions about week 6, then week 5, then week 4. Fifty-four adults participated in this study. We recruited them in the same way as before, but none had been in the earlier experiment.

The questions about quantitative information produced clear seam effects. Figure 4 plots these new data. The y-axis of this figure shows the mean absolute value of the change in the dollar amounts from week to week. For example, if a respondent said that the questionnaire for week 1 had asked whether s/he had spent $1 for milk and cream and that the questionnaire for week 2 had asked whether s/he had spent $5 for milk and cream, then the change for this item would be |1-5| = 4. Figure 4 shows these absolute changes, both for respondents whose questions were blocked by item (filled circles) and for those whose questions were blocked by week (open circles). It's easy to see that while both ways of grouping the questions produced seam effects, the effect was larger for blocking by item. This agrees with our earlier results for qualitative responses (Rips et al., 2000).

Respondents' accuracy for the amounts was quite low overall: They recalled the correct dollar value for only 12.1% of items. Accuracy also decreased significantly over the response periods, although this decrease was relatively small, probably because of a floor effect. On average, respondents were correct for 14.4% of items during the most recent weeks of the response periods (weeks 3 and 6) and for 9.2% for the earliest ones (weeks 1 and 4). This decrease was about the same whether the questions were blocked by item or blocked by week during the test sessions.[6]

The more interesting findings, however, concerned constant-wave responding. We counted a respondent as making a constant-wave response during a test session if he or she gave the same answer (dollar amount) each time we asked about an item. For example, if a respondent said during the first test session that he or she was asked about spending more or less than $2 for milk and cream on the week 1 questionnaire, $2 on the week 2 questionnaire, and $2 on the week 3 questionnaire, this was scored as a constant-wave response. (A nonresponse on all three questionnaires was not scored as constant wave.) In these terms, respondents made constant-wave responses to 36.0% of the items during the test sessions. As we noted earlier, the correct amount was actually constant from week to week for half the items and varied for half; so a constant response was appropriate for the former items and incorrect for the latter. Respondents' answers, however, were not greatly different for these two item classes. When a constant response was the correct answer, respondents gave constant answers for 44.2% of items; when a constant response was incorrect, they gave constant answers for 33.8% of items. The number of constant-wave responses did, however, depend on whether the questions were blocked by week or blocked by item. When blocked by item (e.g., all questions about milk-and-cream appeared together in the test), 51.4% of responses were constant wave. But constant-wave responses decreased to only 18.5% when the items were blocked by week (e.g., all questions about week 3 appeared together).

These results suggest that grouping questions about the same topic encourages respondents to give the same answer to each item. If consecutive questions ask respondents about the amount for milk-and-cream in week 3, milk-and-cream in week 2, and milk-and-cream in week 1, it's tempting for these respondents to give exactly the same answer each time. Placing these questions in different parts of the test, as we did when questions were grouped by week, greatly decreases this tendency. This difference clearly contributes to the larger seam effect when questions were grouped by item rather than by week, as seen in Figure 4. In earlier research (Rips et al., 2000), we had also obtained greater accuracy when questions were grouped by week than when the same questions were grouped by item. This was not true in the present study: Respondents were correct on 10.9% of questions when these questions were blocked by week and on 13.1% when blocked by item (a small but marginally significant reversal). This difference between experiments may be due to the fact that in the earlier study none of the items had correct answers that were constant across weeks, while in this study half had correct constant answers.

Simulations of the Seam Effect

The exact form of the seam effect differs in these studies, probably as the result of the relative contributions of memory, constant-wave responding, and other factors. To see why, consider respondents in a SIPP-like survey who are faced with yes/no questions about whether they received some benefit. Respondents' memory for the event will be most accurate for the periods just preceding the interview, declining through the response period, probably at a decreasing rate. If the tendency toward constant-wave responding is moderate, this will create sizeable changes in responses for the months just preceding the interview (when memory is fading most rapidly). It will also produce a sizeable seam effect, since the seam data come from the most recent month of the first interview (when memory is strongest) and the earliest month of the second interview (when memory is weakest). The result will be asymmetric curves, such as the ones in Figure 5a, which come from simulations based on the assumptions just outlined.[0] The figure shows that the degree of asymmetry in the curves – the amount by which the right-most point on the curve is higher than the left-most point – depends on the strength of memory for the target events. If memory for the event is initially weak (front part of the figure), then the asymmetry is relatively mild, whereas if memory for the event is initially strong (rear part of the figure), the curve is much more clearly asymmetrical.

Respondents' tendencies to make constant-wave responses can also affect the global shape of the function. As an extreme case, if memory is negligible for the events in question and if respondents always make a constant wave response, then the function will be perfectly symmetrical, as shown at the front of Figure 5b. The only opportunity for changing a response in this situation occurs for cases in which a respondent makes one constant-wave response (e.g., "yes") during the first interview, and a different constant wave response ("no") for the second. If respondents are less willing to make constant-wave responses, the asymmetry will increase accordingly, as Figure 5b also shows.

These simulations suggest that the asymmetrical curve from our first study (filled circles in Figure 3) may have been the result of the respondents' fairly accurate memory for the everyday, personal events we asked about. The curve shows the rise in the middle and end that we see in Figure 5a. Asymmetries also appeared in our second study when we grouped questions by week (open circles in Figure 4). These asymmetries largely disappeared, however, when we grouped questions by item (filled circles in Figure 4). Grouping by item probably encourages constant-wave responding, increasing the symmetry of the curve. The differences between the two curves in Figure 4 are similar to the differences in the curves of Figure 5b, where we have deliberately varied the underlying rate of constant-wave responding. SIPP also groups questions by item – for example, asking all the questions about receiving one type of benefit before asking about other types – and this may help account for the symmetric curves in the SIPP data that we glimpsed in Figure 1.

Summary and Implications for Surveys

All the studies we have conducted to date have obtained seam effects – larger changes in responses when the data come from two different interviews than from the same interview. In most of these studies, the key questions that produced the seam effect concerned information that we had supplied. However, the first experiment we reported here extends our finding to naturalistic events, similar to those in actual surveys. This study also compares a situation in which the response period of the interviews is the same as that of the questions to one in which the interviews' response period is longer than that of the questions. As Figure 3 illustrates, only the second type of schedule produced a seam effect. The seam effect is the result of economizing on the number of interviews: By interviewing every four months and asking questions about each month in the preceding interval, these surveys produce response errors that would probably not have occurred if the interviews had been conducted every month.

The studies we described here also show that seam effects appear both for questions about quantities (amounts paid for goods in this case), as well as questions about the occurrence or nonoccurrence of events. In addition, the size of the seam effect for both quantitative and qualitative information depends on question order. When respondents receive questions about the same content for one temporal interval after another, it's easy for them to give the same answer to each item in the series. These constant-wave responses, in turn, increase the seam effect.

Our data show that separating questions about the same topic can reduce the size of the seam effect. The results are not so clear about the effect of this manipulation on accuracy. As we mentioned earlier, the outcome on accuracy may depend on the pattern of events that the survey questions target: Separating questions about the same topic may be helpful when the true answers vary from one response interval to another. It may be of less help when true answers are in fact constant across intervals. We believe similar caution is probably warranted for other methods for reducing seam effects. We can probably reduce seam effects by counteracting biases in respondents' answers, such as the constant-wave tendency, but we need to be careful that in doing so we don't also introduce other sources of error.[8]

For example, SIPP has begun dependent interviewing to help reduce the size of the seam effect. In the second and later interviews, respondents are told about the information they provided in the previous interview before they answer related questions about the current response period (e.g., Last time you said you had job X. Do you still hold that job?). It seems likely that dependent interviewing can smooth seam transitions by reminding respondents of their previous answers. In some cases, this could also provide a memory prompt for information that respondents might not otherwise remember. In other cases, though, giving respondents their own earlier answers may simply bias them toward giving the same answer in the current round of questions, providing an anchor for their judgments (Wilson, Houston, Etling, & Brekke, 1996). Although this would minimize the seam effect, it could lead to equally inaccurate responses. We need to check empirically in each case to see whether reducing the seam effect does more harm than good.

References

Jabine, T. B., King, K. E., & Petroni, R. J. (1990). Quality Profile, Survey of Income and Program Participation. Washington, D. C.: Bureau of the Census.

Kalton, G., and Miller, M. E. (1991). The seam effect with social security income in the Survey of Income and Program Participation. Journal of Official Statistics, 7, 235-245.

Marquis, K. H., & Moore, J. C. (1989). Response errors in SIPP: Preliminary results. Proceedings, 1989 Annual Research Conference, U. S. Bureau of the Census. (Arlington, VA, March 19-22, 1989). Washington D.C.: U. S. Bureau of the Census.

Rips, L. J., Conrad, F. G., & Fricker, S. S. (2000). Unraveling the seam effect. Proceedings of the Section on Survey Research Methods, American Statistical Association. Alexandria, VA: American Statistical Association.

Ruben, D. C., & Wetzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. Psychological Review, 103, 734-760.

Shum, M. S., & Rips, L. J. (1999). The respondent's confession: Autobiographical memory in the context of surveys. In M. Sirken, D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, & R. Tourangeau (Eds.), Cognition and survey research (pp. 95-109). New York: Wiley.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. Cambridge, England: Cambridge University Press.

Wilson, T. D., Houston, D. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. Journal of Experimental Psychology: General, 125, 387-402.

Young, N. (1989). Wave-seam effects in the SIPP. Proceedings of the Section on Survey Research Methods, American Statistical Association. Alexandria, VA: American Statistical Association.

Footnotes

1. NSF Grant SES-9907414 supported the research reported here. We are grateful to Jami Barnett for her help with these studies. We also thank audiences at the American Statistical Association and the Washington Statistical Society for comments on previous reports about this project. Gordon Willis and Elizabeth Martin commented on this paper, and their remarks will appear along with it in the published version. In fairness to these discussants, we have left the text of the paper in the form that they read it; however, we have added several footnotes addressing a few of the problems the discussants raised. The new footnotes appear in angle brackets. Correspondence about this paper should be sent to Lance Rips, Psychology Department, Northwestern University, 2029 Sheridan Road, Evanston, IL 60208.

2. <We note that seam effects can be quite large. For example, Kalton and Miller (1991) present SIPP data showing that 98.3% of respondents report no change in social security benefits from one month to the next when the months are off-seam; however, only 34.4% report no change across seam months. Seam effects also appear in a wide range of variables, including such important characteristics as employment status and total family income (Young, 1989). Of course, whether survey researchers need to worry about these differences depends on their purposes. But those who use panel surveys to make inferences about changes (e.g., changes in social security or food stamp benefits) need to be cautious, unless overestimates of changes between seam months exactly balance underestimates between nonseam months.>

3. Memory may still have a role to play, however, in explaining the positive bias. Respondents had seen all items prior to the test sessions – although, of course, not on each week – so the items may have seemed familiar to them. "Yes" responses based on familiarity could explain positive constant-wave answers.

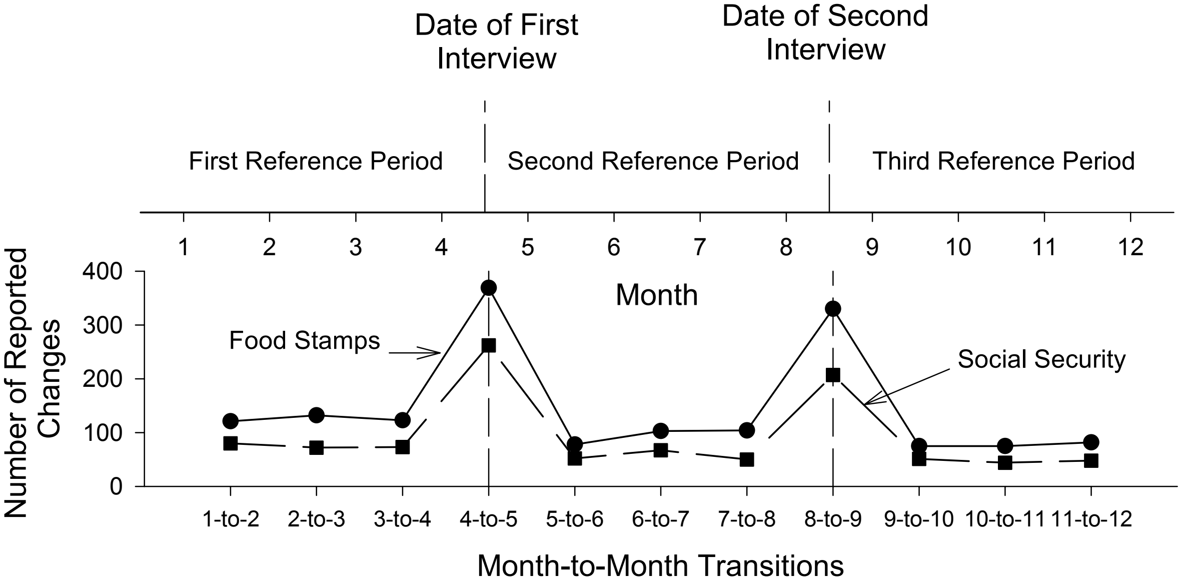
4. < Figure 3 shows that the test-session data underestimate the number of changes both across seam and nonseam months. This finding contrasts with the results of our preliminary study (Rips et al., 2000) and with results from SIPP (Marquis & Moore, 1989). In the earlier work, seam months produced overestimates of the number of changes, whereas nonseam months produced underestimates. The difference between the present study and the earlier ones is probably due to the particular distribution of the personal events we tested here. For example, there were 49 cases in which an event occurred to a respondent in both weeks 4 and 5, but the respondent failed to report it for week 5. This type of error might be due to forgetting and would serve to inflate the seam change. However, there were also 61 cases in which an event did not occur in week 4, did occur in week 5, but was not reported for either week. This could again be due to forgetting, but it would serve to deflate the seam change. Because of the larger number of events of the second type, incorrect reporting tended to produce too few changes at the seam. Underestimates were less severe, however, for seam months than for nonseam months.>

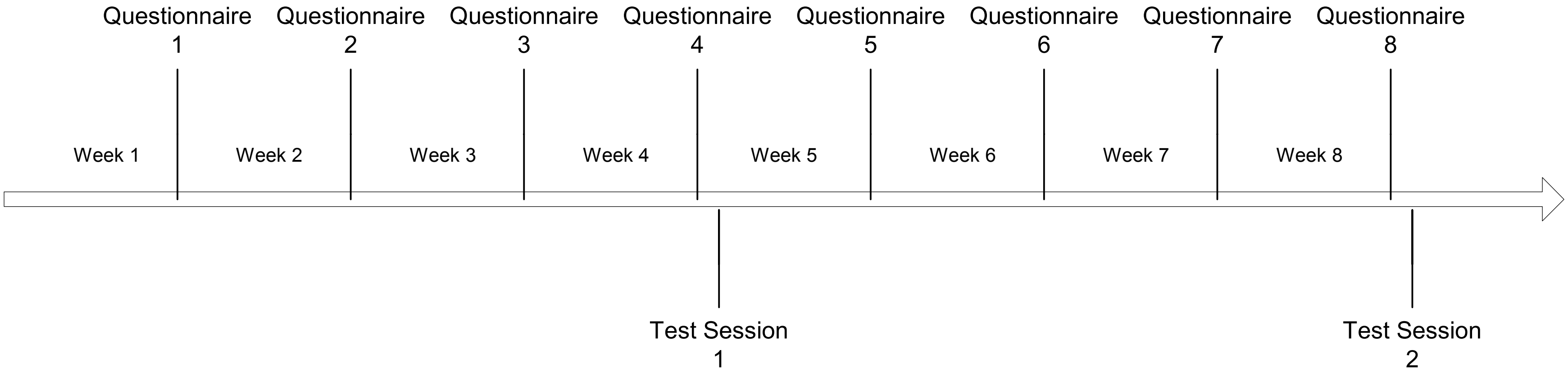
. 5. The amounts for these variable items changed according to four patterns. One group of items increased in amount for weeks 1-3 and increased again for weeks 4-6; a second group increased for weeks 1-3 and then decreased for week 4-6; a third group decreased, then increased; and a fourth group decreased then decreased. For example, if the item about milk and cream was in the increase-increase group for a specific respondent, that respondent was asked on the first questionnaire, Last week, did you spend more or less than $1 for milk or cream...?, on the second questionnaire, Last week, did you spend more or less than $2 for milk or cream...?, and on the third questionnaire, Last week, did you spend more or less than $4 for milk or cream...?. This same sequence was then repeated for weeks 4-6. A respondent saw an equal number of items in each of the four groups. Individual items were rotated through the groups across respondents.

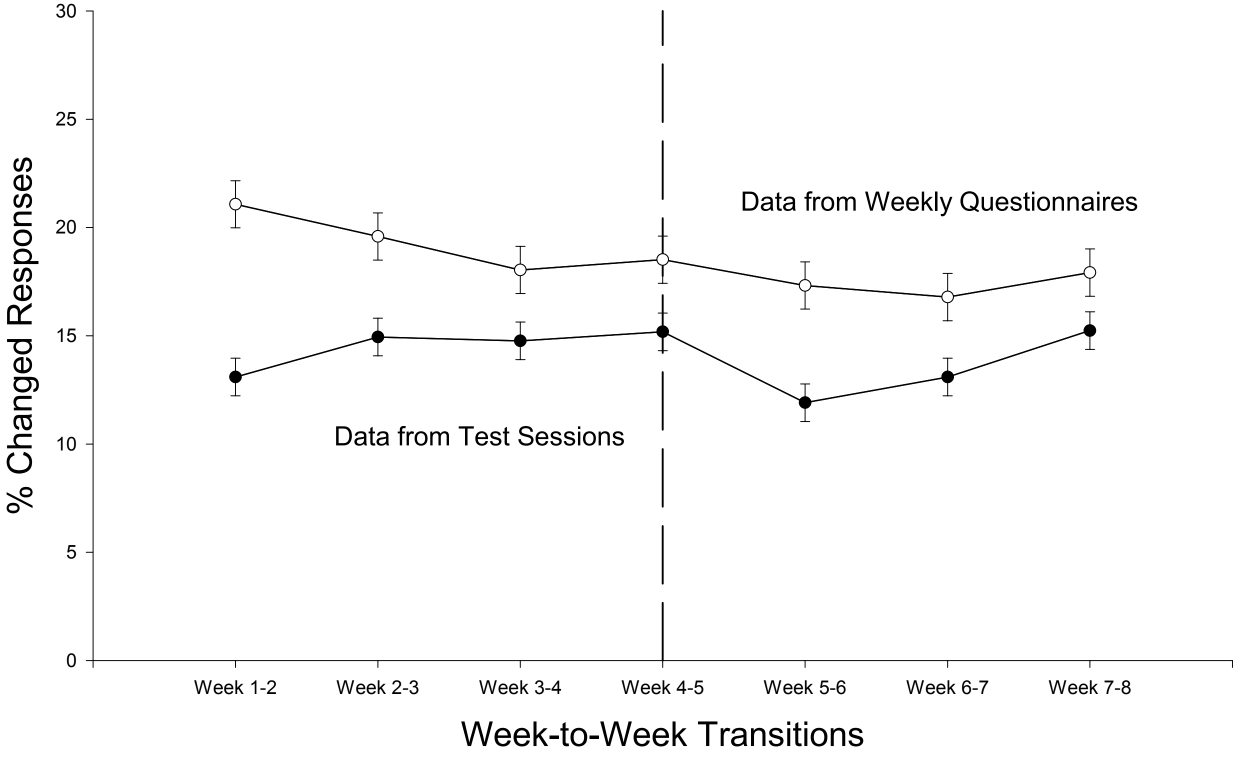
6. <These accuracy figures indicate whether the reported amount exactly matched the amount on the questionnaire. We can construct a more sensitive measure of accuracy by calculating the average absolute deviation (in dollars) between the recalled amount and the true amount. For weeks 3 and 6, the weeks just before the test sessions, the average deviation was $5.88. For weeks 1 and 4, the earliest weeks, the average deviation increased to $6.53. This difference again supports the hypothesis that respondents were forgetting the correct amounts over the reference period. Of course, when respondents no longer remember the amount, they can use a variety of strategies to produce an answer, including constant-wave responding, estimating a usual value for the product or service, or even sheer guessing. The point of the present paper is that the seam effect depends both on forgetting and on the nature of these alternative strategies, as we attempt to show in the following section of this paper.>

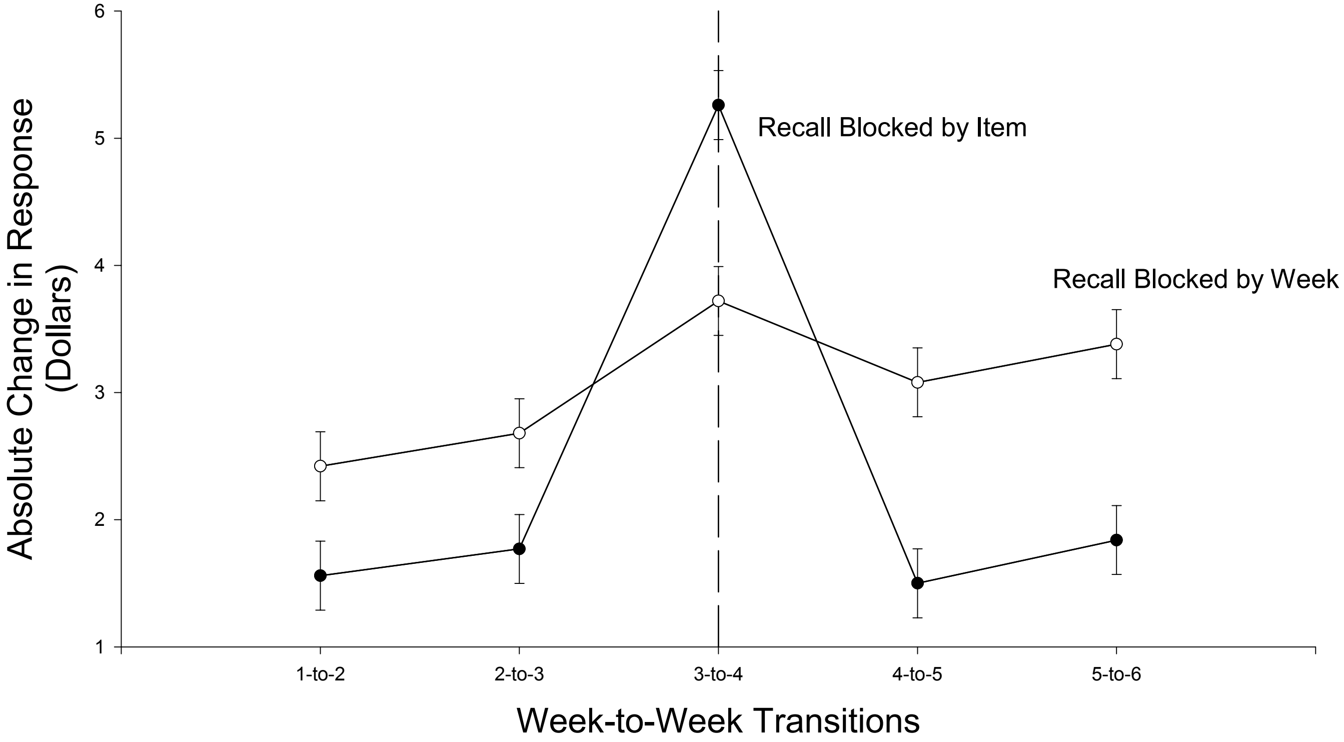
0. 7. For purposes of these simulations, we assumed that forgetting followed a negative logarithmic function of time ($y = a - b*\ln(t)$), in line with earlier work on long-term forgetting (Rubin & Wetzel, 1996). The exact form of the function is probably not crucial, however, as long as memory decreases at a steep rate at first and more gradually thereafter. We also assumed that when a particular piece of information is forgotten, a respondent can rely on one of two strategies. First, the respondent can interpret forgetting as negative evidence (i.e., failure to remember indicates that the event in question didn't occur), and answer "no." Second, the respondent can make a constant-wave response, answering "yes" for all earlier intervals or "no" for all earlier intervals in the response period. The functions in Figure 5 then depend on three parameters: the forgetting parameters (a and b in the equation above), and the probability of a constant wave response. Figure 5a varies the first of these parameters, and Figure 5b varies the third.

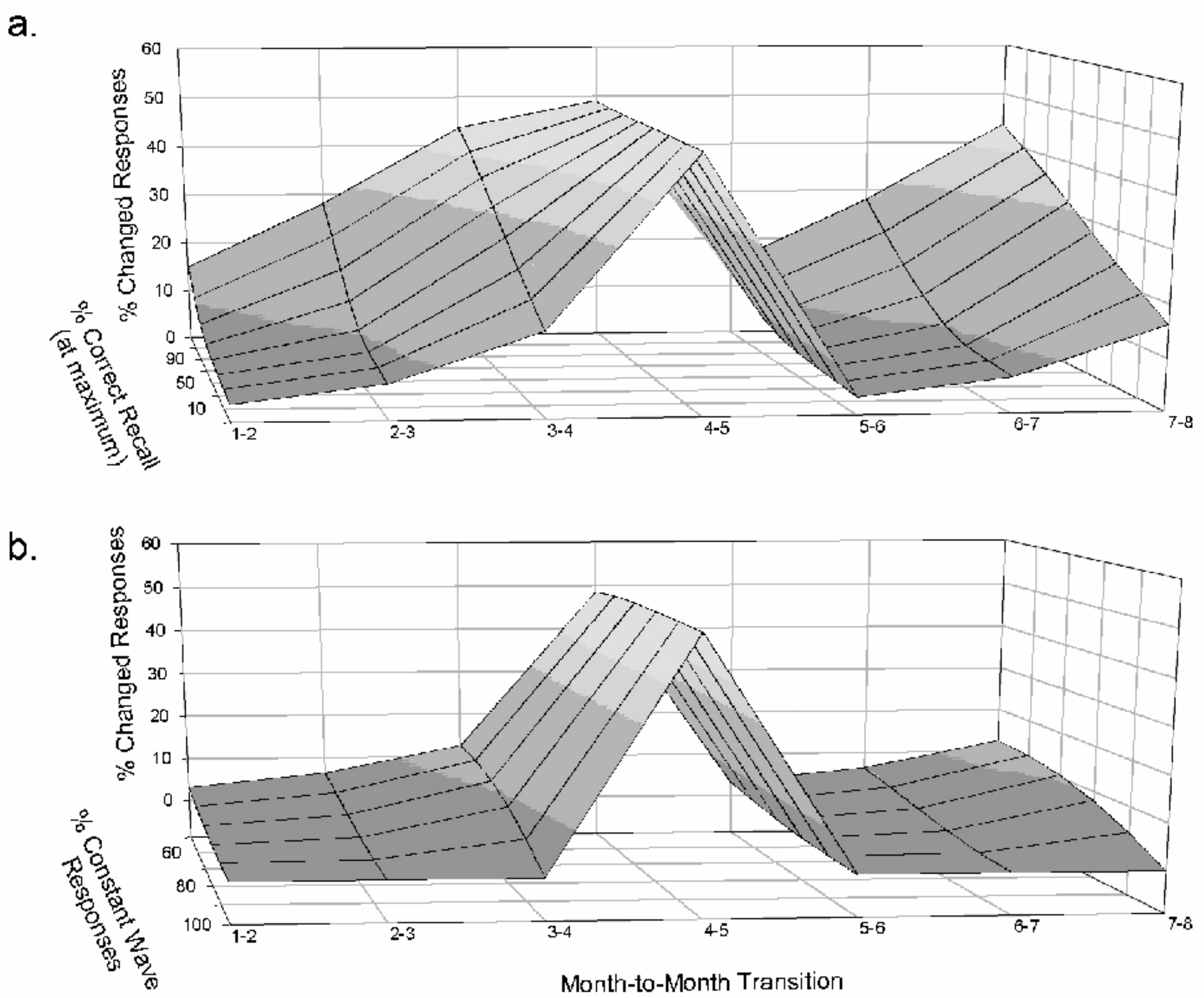
8. <It may seem disappointing that the size of the seam effect isn't a clear indicator of overall data accuracy. However, the seam effect depends on the variability of responses from month to month, and we shouldn't expect variability (or stability) to correlate perfectly with the responses' correctness. We hope that the technique we are developing here can serve as a useful way of finding out how proposed methods for reducing the seam effect will impact other aspects of data quality. As such, methodologists could use the technique alongside other procedures, such as cognitive interviewing, for anticipating problems in actual surveys.>


Date of First Interview
Date of Second Interview
First Reference Period
Second Reference Period
Third Reference Period
1
2
3
4
5
6
7
8
9
10
11
12
Month
Number of Reported Changes
400
300
200
100
0
Food Stamps
Social Security
1-to-2
2-to-3
3-to-4
4-to-5
5-to-6
6-to-7
7-to-8
8-to-9
9-to-10
10-to-11
11-to-12
Month-to-Month Transitions


Questionnaire 1
Questionnaire 2
Questionnaire 3
Questionnaire 4
Questionnaire 5
Questionnaire 6
Questionnaire 7
Questionnaire 8
Week 1
Week 2
Week 3
Week 4
Week 5
Week 6
Week 7
Week 8
Test Session 1
Test Session 2


30
25
20
15
10
5
0
% Changed Responses
Data from Weekly Questionnaires
Data from Test Sessions
Week 1-2
Week 2-3
Week 3-4
Week 4-5
Week 5-6
Week 6-7
Week 7-8
Week-to-Week Transitions


Absolute Change in Response
(Dollars)
6
5
4
3
2
1
Recall Blocked by Item
Recall Blocked by Week
1-to-2
2-to-3
3-to-4
4-to-5
5-to-6
Week-to-Week Transitions


a.
60
50
40
30
20
10
0
% Changed Responses
% Correct Recall
(at maximum)
90
50
10
1-2
2-3
3-4
4-5
5-6
6-7
7-8
b.
60
50
40
30
20
10
0
% Changed Responses
% Constant Wave
Responses
60
80
100
1-2
2-3
3-4
4-5
5-6
6-7
7-8
Month-to-Month Transition

Rips, Conrad, Fricker, and Behr
Seam Effects for Quantitative and Qualitative Facts

Discussion

Gordon Willis, National Cancer Institute

In general terms, Rips, Conrad, Fricker, and Behr are to be commended for their careful description and analysis of the Seam Effect in panel surveys, and for the way in which they tie this conceptually to "Constant Wave Effects" operating at individual interview points. At a more specific level, my comments begin with a basic question: Is the Seam Effect a worthy protagonist in a research study that endeavors to point the way toward reduction in response error through investigation of the cognitive aspects of survey responding? Potentially, such a study represents an attempt to examine a problem observed in a particular survey environment (here, the Survey of Income and Program Participation, or SIPP), to determine whether there are consistent and modifiable sources of measurement error that reflect the operation of comprehension, recall, and decision processes. Further, to the extent that the Seam phenomenon is pervasive, such research could possibly elucidate cognitive processes that extend to a variety of surveys. Therefore, a vital consideration I pose in evaluating the Rips, et al. investigation is whether the study investigates a phenomenon that a) is non-trivial, b) extends to a range of surveys, and c) is similar in nature to related problems of a cognitive nature that afflict surveys.

Is the Seam Effect an "interesting" phenomenon?

As to whether the Seam Effect is typically of sufficient magnitude to be considered a problem, as opposed to simply a minor source of noise in an otherwise meaningful data distribution, the answer is provided in a separate review by Tourangeau, Rips, and Rasinski (2000), who convincingly portray the sometimes large magnitude of the Seam phenomenon – in particular, the number of changes in status reported between seam months may be as much as 12%, as opposed to a base level of 1-2% between non-seam months.

The Seam Phenomenon also appears to extend beyond the SIPP, as Seam Effects have emerged in the Panel Survey of Income Dynamics, and the Income Development Survey Program. The general paradigm studied conceivably has application to surveys other than those involving income and program participation. For example, health surveys involving cancer risk factors sometimes ask, at one interview point, for reports of status for a series of previous sub-intervals, such as usual weight during a ten-year period, or self-report of male sexual function following treatment for a number of months following prostate cancer. Such procedures do satisfy the procedural requirements necessary, in theory, to produce Constant Wave Effects within interviews, or Seam Effects over repeated interviews. I would therefore propose that a solution to the Seam Effect puzzle is potentially important enough to warrant a systematic examination, especially to the extent that the results are generalizable.

Attempts to generalize the Seam Effect

However, the Rips et al. attempt to directly generalize the Seam Effect to domains beyond that of income and program participation is risky, if this extension is carried out without sufficient consideration of how the cognitive aspects of different behaviors may vary. In particular, the attempt to cover everyday activities such as ordering pizza and changing the oil in one's car, under the Seam Effect umbrella, is problematic. The Seam Effect is hypothesized by the authors to be mediated by Constant Wave Effects, in which the report for an initial sub-interval serves as an anchor that in turn diminishes respondents' tendencies to vary their reports concerning other time intervals. However, Constant Wave effects seem much more likely to occur for some behaviors than others, in fairly predictable ways. For example, if someone changed the oil in a vehicle in one month, it is very unlikely he or she did so the next month as well, so that one might expect, in anything, a "negative wave effect" for that item (so that a positive report in one month would *decrease* reporting of the same behavior in surrounding months). I suggest that the authors need to consider the nature of the behavior studied, and how this may impact the potential for Seam Effects, as they attempt to extend its reach beyond reporting on items such as the receipt of social security benefits and employment.

Does reducing the Seam Effect translate into lower response error?

A key rationale for studying the Seam Effect seems to be that, if this effect is a symptom of cognitive problems, then factors that reduce the magnitude of the Effect will also be likely to improve data quality. For example, in a previous paper reporting on an initial experiment done as part of this research study, Rips, Conrad, and Fricker (2000) found that the use of backward-to-forward temporal ordering of recalled events reduced the magnitude of the Seam Effect, relative to forward ordering, and this finding suggested that the backward order alleviated error. Such a result is important from a methodological perspective, because it potentially provides a proxy measure of survey data quality (the size of the Seam Effect), obviating the need for direct measurement through more expensive procedures such as individual level response validation.

However, the status of Seam Effect-as-proxy-for-quality has become very murky, given the further research that Rips et al. have carried out as part of the current investigation. In particular, the authors find that in some cases manipulations that reduce the size of the Seam Effect may also reduce rather than improve response accuracy. For example, asking a series of survey questions in which related behaviors were organized by time, rather than by topic, did "break up" the tendency for Seam Effects to occur for each topic, but also adversely affected respondents' reporting. As an extreme example, it may be possible to eliminate a Constant-Wave-based Seam phenomenon through the use of random ordering of questions, preventing one response from becoming an anchor for others on similar topics. However, it is doubtful that such a practice would actually result in improved reporting, as this would serve only to treat a possible symptom (the Seam) but not the underlying disorder (response error).

So, it seems that the Seam Effect is not in itself necessarily a direct measure of data quality, and the authors conclude that: "We need to check empirically in each case whether reducing the Seam Effect does more harm than good." However, the need to check in each case (that is, for each new survey or new question environment) effectively nullifies the primary advantage we have for assessing the presence of a Seam Effect in the first place; if its magnitude cannot be mapped directly to data quality, then the fact that it may be modifiable is perhaps interesting, but of very limited practical utility to survey researchers. In order to move this research forward, one would perhaps need to determine the types of situations in which Seam Effects indicate good versus poor quality data. However, pursuing this road further risks a type of reductionism in which the phenomenon under study is increasingly narrowed in scope so that it retains only academic interest, and is found to be too complex and multivariate to have the type of generality that made it appear attractive initially.

Implications of the Rips et al. study

Despite the limitations of the research, I believe that there are a number of interesting implications that are not directly stated, but that stem from the overall findings:

1. The study of general cognitive processes. The fact that the Seam Effect at the outset appeared to be somewhat straightforward in basis and amenable to study appeared to render it a good candidate as an application of CASM (Cognitive Aspects of Survey Methodology) in which a general law of cognitive functioning as related to questionnaire design could be explicated as the result of the research effort. That is, the intent is to demonstrate that Seam Effects are produced by cognitive mechanisms that can be modified in predictable ways by particular design modifications, that employing these design rules will ameliorate the Seam Effect, and that as a byproduct, survey data quality improvements will be realized. The fact that studying this effect did not lead to such a generalizeable rule, such as "reverse temporal ordering of recall is superior to forward ordering" is consistent with other (somewhat frustrating) failures to produce generally applicable rules of questionnaire design through experimental cognitive study. Thus, questionnaire design and evaluation practice continues to be largely an empirical issue, where the factors that impact on design decisions related to a particular survey instrument are complex, and represent the mutual effects of a number of opposing considerations. This is perhaps why practitioners continue to evaluate questionnaires through empirical techniques such as cognitive interviewing, as opposed to simply relying on a bible of design rules. I do not argue that design rules are useless – simply that they must be regarded as a general starting point, and are insufficient in themselves when we partake of questionnaire design "down in the trenches."

2. Survey responding is problem solving. The point has been made many times that survey respondents do not simply directly output information from memory, as these memories are queried by our survey questions. Rather, in the face of partial and difficult-to-retrieve information, respondents make use of processes such as complex estimation, background knowledge of probability, and other heuristics, in order to produce responses that they feel are reasonable. The detailed study of the Seam Effect by Rips and colleagues further demonstrates this effect. When deciding on how to answer a string of questions over sub-intervals related to the same behavior, respondents consider issues such as the likelihood of an event or behavior occurring in month N, given that it had (or had not) occurred in month N-1. In addition to direct retrieval of specific memories, processes such as knowledge of regularity, frequency, and patterning of particular behaviors (whether ordering pizza, changing the oil, or receiving social security benefits) drives the process by which the respondent attempts to reach a suitable level of accuracy under conditions in which memory itself is insufficient. We must continue to be reminded that answers to survey questions are often not so much reported from storage as they are synthesized on the spot from a variety of information sources.

3. Respondent consistency. The finding of a Constant Wave Effect within interview, but sometimes virtually random perturbation in response between interviews, also has two important implications:

a) Consistency of responses within interview is no assurance of response accuracy. Survey designers sometimes are led to believe that their survey questions "work" because no obvious problems arise, and the responses to related items do not illustrate gross inconsistency. However, a generalization of Constant Wave Effects may be "Consistent Answer Effects" in which respondents strive to maintain a coherent picture ("Because I said X to the previous question, now I better answer Y"). So, especially during pretesting and evaluation, it makes sense to delve into the basis for each answer, rather than simply accepting a seemingly solid and consistent facade.

b) On the other hand, we should not necessarily expect great consistency *between* interviews, even for information which should not have changed between interview. This particular issue arises perennially when we conduct reinterview studies to assess question reliability, or where a longitudinal study requires an answer to the same question at multiple time points. A consistent concern is that at time T2, the respondent is simply recalling the answer he/she gave at time T1, rather than recalling the answer anew. However, if respondents' behavioral tendencies with respect to the Seam Effect can be used as a guide, then perhaps those worries are unfounded; if respondents are not even consistent when we *do* want them to be, then perhaps they also are not attempting (or able) to be consistent when we *don't* want them to be.

To return to the initial question posed – Why study the Seam Effect under the rubric of CASM research? – Perhaps the answer is not that this will provide a means for reducing error by finding ways in which to ameliorate this effect, but rather, that it provides a rich source of data indicating how respondents make decisions as they answer survey questions, specifically about a series of past sub-intervals. The fact that they may be inclined to engage in Constant Wave behavior, when the behavior is viewed as likely to be constant in nature, sensitizes us to the need to emphasize the veracity of the initial response reported, and leads us to consider whether it is even advisable to request information from the respondent that may be severely tainted by other reports they have just given. As a means for investigating this phenomenon further, I advocate additional research which attempts to directly determine the effects of economizing survey reporting by obtaining monthly (or other periodic) information on a less-than-monthly basis. In particular, by comparing the results of a procedure in which some respondents actually are asked the repeated questions (on program participation, ordering pizza, etc.) monthly, and others are asked for the same information, but periodically (e.g., quarterly), we can determine the direct effects of the use of the latter procedure.

References

Tourangeau, R., Rips, L.J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge Cambridge University Press.

Rips, L.J., Conrad, F.G., & Fricker, S.S. (2000). Unraveling the Seam Effect. *American Statistical Association, Proceedings of the Section on Survey Research Methods*

Remarks on "Seam Effects for Quantitative and Qualitative Facts"

Elizabeth Martin
U. S. Census Bureau

It may be useful to begin with a brief recapitulation of the research you've just heard reported. Rips and his colleagues are attempting to reproduce or simulate the seam effect in the lab. To do this, they in effect miniaturize everything--months are transformed into weeks, and the four month wave is transformed into a four (or three) week wave. They ask respondents about two different kinds of events. First, every week they send questionnaires to be filled out, then at the end of four weeks they interview respondents about what questions appeared in the questionnaires. Let's call this "questionnaire recall." Second, they also ask respondents--both in the questionnaires, and in the end-of-wave interview sessions--about ordinary events that may have happened during each week of the "wave." Let's call this "ordinary event recall." "Transitions" are measured as week-to-week changes in respondents' reports. So, if a respondent said a particular question appeared in one of the weekly questionnaires but not the next one, that counts as a transition. For ordinary events, if a respondent said a particular event happened one week and not the next, that counts as a transition. For both types of events, they have measures of truth. For the first, they know which questionnaires they sent out, so they know which questions were in fact asked each week. For the second type of event, they have the responses to the weekly questionnaires as a check on the accuracy of respondents' reports about the same events at the end of the four week "wave."

I think there are three questions we need to ask about the research.

- First, have they produced seam effects in the lab?
- Second, does their laboratory version of the seam effect reproduce or match what we know about essential features of the survey phenomenon?
- Third, if we are satisfied that their laboratory simulation reproduces the survey reporting phenomenon in critical ways, what light does their research shed on the cognitive underpinnings of seam bias?

I would answer the first question with a skeptical "maybe." Despite the authors' interpretations, I do not see evidence of a seam effect for the second type of "ordinary event recall." Compared to the weekly questionnaires which serve as the measure of truth, the test sessions produce consistently lower estimates of week-to-week changes in responses. But we know from record check studies of income reporting that, compared to truth, the number of transitions at the seam is too high, and the number of within-wave transitions is too low (Moore and Marquis, 1989). Hence we should observe the lower line spike up above the top "truth" line at the seam, but it doesn't.

For the other type of "questionnaire recall" the results do seem to show a seam effect which is very much affected by the structure of questioning during the test session interview. For example, respondents were asked in the test session to recall the dollar amounts asked about in questionnaire items for each of the weeks of the "wave." When the recall questions were organized by week, the seam effect is slight; when organized by topic (i.e., respondents were asked to recall the dollar amount referred to in a particular question in each of the weekly questionnaires) it is very large.

The authors theorize that the seam effects in their lab studies, and in surveys such as SIPP, are produced by the combined effect of recall for events in relatively recent time periods and estimation of events in distant time periods, which respondents do not remember well.

But are Rips and his colleagues measuring recall in their laboratory simulation? The events for which they find seam effects--recall of questions asked in questionnaires--are very ephemeral and inconsequential, in contrast to receipt of income, say. The authors do not provide evidence to support the premise that any recall at all is involved in this reporting task. It seems plausible that the task of "recalling" which of eight or four weekly questionnaires contained a particular item is pure guesswork, and that the seam effects they produce are a consequence of artificial constraints upon the consistency of guesses across the weekly time periods. Answers to the following questions would shed light on whether respondents are engaged in recall or guessing:

1.) What fraction of correct responses should be expected by chance? The test sessions only asked about events (questionnaire items) that really had appeared in the weekly questionnaires, and the results suggest respondents were biased toward positive answers to the questionnaire recall questions. It might be useful to include in the test sessions questions that ask respondents to "recall" items that never appeared in any questionnaire, to learn whether respondents are as likely to say they saw a questionnaire item that never appeared in any questionnaire, as one that did. If so, then it's difficult to interpret the results as being about something other than guessing.

2.) Were respondents given the option of responding "don't know" to the questionnaire recall questions, and what fraction did so (or did not respond)?

3.) Did the researchers conduct any debriefings or think-alouds with respondents to learn how they attempted to solve the questionnaire recall task?

4.) What is the correlation between respondents' reports and truth? (I suspect it is close to zero.)

If the questionnaire recall task does not involve recall, then the answer to the second bulleted question above is "no," because the laboratory version of the income reporting task does not match what is known about the survey phenomenon. Income reports may be characterized by a good deal of error, but no one doubts that income receipt is memorable and that reporting income involves recall.

In seeking to reproduce the seam bias phenomenon, it would be useful to review what is known or suspected about the seam bias as it affects income reporting. I was surprised the authors had not done this. We have a good deal of evidence about the seam bias from the record check studies of the Survey of Income and Program Participation (SIPP), conducted by Kent Marquis and Jeff Moore. That research suggests that:

1.) The seam bias appears to involve both underreporting of true changes within a wave, and overreporting of changes between waves (Moore and Marquis, 1989).

2.) Reporting accuracy (i.e. low underreporting error) varies a good deal by program, but the seam effect turns up even for very accurately reported events, such as Social Security income (see Marquis and Moore, 1990).

3.) It is not clear that more recent events (e.g., income receipt one month ago) are more accurately reported than more distant events (e.g., income receipt four months ago) (see Marquis and Moore, 1990; table 4.2). This finding is counterintuitive, and other research suggests that recall for income receipt does deteriorate over time. For example, Kalton and Miller (1991) find that a one-time increase in Social Security payments was less likely to be recalled and reported the longer the time interval between the occurrence of the increase and the interview date. The possible role of memory decay in producing the seam bias is an important question for research. The authors beg the question by assuming that memory decay produces better recall for recent events than for more distant ones, taking the evidence of a seam effect as support for this explanation. However, any factor or process that increases the consistency of reporting across weeks within a wave, and/or that reduces the consistency of reporting between waves, could produce a seam effect. It would be useful to contrast their hypothesis with alternative hypotheses about the underlying cognitive processes that may account for the seam bias phenomenon. Existing research has implications for theorizing about the cognitive processes underlying seam effects, and should be taken into account in research on the topic. Results of methodological studies which have attempted to correct the seam bias (see, e.g. Moore, Marquis, and Bogen, 1996) are also pertinent and should be considered in developing cognitive theories of the seam effect and proposing solutions for it.

The authors need to reexamine the task they are using in their laboratory simulation, which should more closely mimic the survey task that gives rise to the seam bias phenomenon. The character of the events being reported about in surveys such as the SIPP is quite different from the questionnaire events for which the authors find seam effects in their laboratory simulation, and the differences almost certainly affect the recall strategies employed by respondents. Income receipt is, for most of us, pretty memorable and consequential. For most of us, it is temporally regular, and patterned in some way. It depends on external, continuing sets of conditions and life circumstances--having a job, being eligible and enrolled for food stamps or social security, and so on. When these conditions change, then income receipt changes--one loses a job, stops receiving wage income, becomes eligible for unemployment compensation, loses eligibility after a certain number of weeks, and so on. Behind the month-to-month changes in recipiency are real transitions in life circumstances which are meaningful to respondents. The fact that income receipt is associated with meaningful transitions influences the response strategies available to respondents as they report income. They can try to reconstruct the timing of a change in income source or amount using associated life events and changes as anchors and landmarks. (Of course, the fact that such recall strategies are available to respondents does not necessarily mean they employ them, or that they produce accurate reports if they do.) Such strategies are unavailable to respondents in this study, because the events (recall of questionnaire items) are meaningless and the "transition" from one item to another in different weekly questionnaires is completely artificial and arbitrary, from the respondent's point of view. (I find it difficult to imagine strategies for answering these questions other than constant wave responding or random guessing.) Thus, the cognitive processes respondents engage in during the questionnaire recall task seem considerably different from the cognitive processes involved in reporting income. For this reason, the answer to the third bulleted question in my opinion is "no," this research has not (yet) shed light on the cognitive underpinnings of the seam bias phenomenon in surveys. However, this is an interim assessment; by better integrating knowledge from the existing methodological literature, and by reexamining and corroborating their assumptions about the nature of the response task and the cognitive processes that respondents engage in, and by exploring respondents' response strategies more directly, the authors would make useful contributions to our understanding of this difficult survey problem.

References

Kalton, G. and Miller, M. E. (1991) “The seam effect with Social Security income in the Survey of Income and Program Participation.” Journal of Official Statistics 7(2):235-245.

Marquis, K. H., and Moore, J., (1990), "Measurement Errors in SIPP Program Reports." Proceedings of the 1990 Annual Research Conference, pp. 721-745.

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

Moore, J. C., Marquis, K. H., and Bogen, K. (1996), The SIPP Cognitive Research Evaluation Experiment: Basic Results and Documentation. Washington DC: Census Bureau.

Session 3

A Computer Tool to Improve Questionnaire Design

FCSM Seminar on the Funding Opportunity in Survey Research
Introduction to Session 3, "A Computer Tool To Improve Questionnaire Design"

Chair, Robert Parker, U.S. General Accounting Office

The subject of our 3rd session today is "A Computer Tool To Improve Questionnaire Design" and features a paper by a number of faculty members of the University of Memphis, headed by Professor Arthur Graesser, who will make this morning's presentation. Before introducing our speaker, I'd like to say that I am very pleased to be chairing this session, because I strongly support research designed to help reduce nonsampling errors and to increase response rates. The work described in this paper looks like a promising step in that direction, and I look forward to hearing comments from our discussants.

Now let me introduce our speaker. Professor Graesser is presently a full professor in the Department of Psychology and an adjunct professor in Mathematical Sciences at the University of Memphis. He is currently a co-director on the Institute for Intelligent Systems and director of the Center for Applied Psychological Research. Dr. Graesser received his Phd in psychology from the University of California at San Diego and has as his primary research interests cognitive science and discourse processing. He is currently editor of the journal Discourse Processing. In addition to publishing over 200 articles, he has written 2 books and edited several others.

Our first discussant will be Terry DeMaio, a principal researcher in the Census Bureau's Center for Survey Methods Research. She has been at the Census Bureau for 25 years, working on research issues related to nonresponse and questionnaire design. She currently heads a group that conducts research on the Bureau's demographic surveys. Terry received her graduate training in sociology at University of Indiana.

Our second discussant will be Fran Featherston. Fran is a senior survey researcher at the General Accounting Office and has extensive research in the design and analysis of a wide variety of surveys. Fran received a Phd in political science from the University of Michigan.

I want to thank Professor Graesser and our two discussants for their presentations. I also would like to add additional comments on the QUAID computer tool. First, it would seem to me that this tool would be useful not only to survey designers but also to managers in statistical agencies. Using QUAID, or some derivative program, for all surveys could provide managers with the knowledge that the questions in their surveys have been designed in a way to reduce comprehension problems by respondents. Second, it would seem that a next major step in the development of QUAID would be the ability to apply it simultaneously to groups of similar questions on a single survey.

A Computer Tool to Improve Questionnaire Design

Arthur C. Graesser, Ashish B. Karnavat, Frances K. Daniel, Elisa Cooper,
Shannon N. Whitten, and Max Louwerse
University of Memphis

Paper presented at the Funding Opportunity in Survey Research Seminar, June 11, 2001, Bureau of Labor Statistics, Washington, DC.

Send correspondence to: Arthur C. Graesser, Department of Psychology, 202 Psychology Building The University of Memphis, Memphis, TN 38152-3230, (901) 678-2742, 901-678-2579 (fax), a-graesser@memphis.edu.

This research was funded by grants awarded to the first author by the Statistical Research Division of the United States Census Bureau (1998-1999, 43-YA-BC-802930) and the National Science Foundation (SBR 9977969).

Abstract

We have developed a computer tool (called QUAID) that assists survey methodologists who want to improve the wording, syntax, and semantics of questions on surveys and questionnaires. QUAID stands for “Question Understanding Aid.” The input to QUAID consists of a question on a questionnaire, whereas the output is a list of potential problems with the question, including: (1) unfamiliar technical term, (2) vague or imprecise relative term, (3) vague or ambiguous noun-phrase, (4) complex syntax, and (5) working memory overload. QUAID is now available on the web (www.psyc.memphis.edu/quaid.html). This web facility encourages researchers to send us problematic questions so that we can iteratively assess and improve the performance of QUAID. We have performed analyses that assess how well QUAID diagnoses these five problems with questions, sampled from a corpus of 11 surveys provided by the US Census Bureau. We have also collected eye- tracking data while college students answer 69 questions.

Introduction

Questions on a survey should elicit valid and reliable answers from respondents in a short amount of time. The goals of validity, reliability, and efficiency cannot be met if respondents have trouble comprehending the questions. So how do survey methodologists identify questions that are difficult for respondents to comprehend? One method is to have experts identify particular problems with questions (Lessler & Forsyth, 1996). A second approach is to collect verbal protocols from respondents as they answer questions (Willis, DeMaio, & Harris-Kojetin, 1999); some of the problems with questions can be articulated by respondents. A third approach is to observe behaviors, such as pauses or requests for clarification, that suggest that the respondents are struggling with a particular question (Fowler & Cannell, 1996; Schober & Conrad, 1997).

A fourth approach is to build a computer model that identifies problems with questions in a theoretically principled or systematic fashion (Graesser, K. Wiemer-Hastings, Kreuz, P. Wiemer-Hastings, & Marquis, 2000). Building such a computer requires the coordination of several fields, including computer science, computational linguistics, discourse processing, cognitive science, and survey methodology. This fourth approach was pursued in the present project. We have developed a computer program (called QUAID) that critiques questions on different comprehension problems.

Researchers in CASM (Cognitive Aspects of Survey Methodology) have adopted models that dissect different stages question-answering (Jobe & Mingay, 1991; Lessler & Sirken, 1985; Sudman, Bradburn, & Schwarz, 1995; Schwartz & Sudman, 1996; Tourangeau, 1984; Sirken, Hermann, Schechter, Schwarz, Tanur, & Tourangeau, 1999). The stages included in most of these models are question interpretation, memory retrieval, judgment, and response selection. The inaccuracy and variability of question interpretation among respondents is known to be one of the serious sources of error that threaten the reliability and validity of answers to questions (Fowler & Cannell, 1996; Groves, 1989; Lessler & Kalsbeck, 1993; Schober & Conrad, 1997). Therefore, revising questions to minimize interpretation problems is one important strategy for reducing measurement error. QUAID was designed to diagnose interpretation problems, as opposed to other stages of questions answering (memory retrieval, judgment, and response selection).

QUAID stands for Question Understanding Aid. It has particular modules that critique each question on potential comprehension difficulties at various levels of language, discourse and world knowledge. The critique identifies words that are unfamiliar to most respondents, vague predicates (verbs, adjectives, adverbs), ambiguous noun-phrases, questions with complex syntax, and questions that overload working memory (Graesser, K. Wiemer-Hastings, Kreuz, P. Wiemer-Hastings, & Marquis, 2000). The identification of such problems should be useful to the survey methodologist if the computer tool can accurately identify the questions with potential problems and can point out what the problems are. Some of these problems might otherwise be missed because of fatigue or training deficits in the survey researcher who writes, revises, and pretests the questions. Most survey researchers do no have extensive training in linguistics, discourse processing or cognitive science, so QUAID should be a valuable augmentation to the standard tools of the survey methodologist.

This paper is a progress report on our development and evaluation of QUAID. Section 1 presents a succinct overview of QUAID. The is a web facility that survey methodologists can use to obtain critiques of questions with QUAID. Our hope is that survey methodologists use this web facility and send us problematic questions; these will be used in future tests and refinements of QUAID. The second section reports a recent evaluation of the performance of QUAID. That is, how well can it accurately discriminate questions with particular problems, when compared to expert evaluations as a gold standard. The third section describes an eye tracking study that recorded the eye fixations and eye movements while respondents answer survey questions. We are currently assessing the extent to which eye tracking patterns reveal problems with questions.

QUAID (Question Understanding Aid)

This section briefly describes the QUAID computer tool. QUAID can handle 5 problems with questions, as described shortly. The questionnaire designer first types a question into QUAID. Then QUAID critiques the question on the 5 different components. There are three levels of each critique that vary in specificity, from succinctly identifying a problem to a lengthy description of the nature of the particular problem.

Graesser's previous research has identified 12 problems with questions that periodically occur in surveys (Graesser, Bommareddy, Swamer, & Golding, 1996; Graesser, Kennedy, P. Wiemer-Hastings, & Ottati, 1999). Many of these problems have been incorporated in various analytical coding schemes of survey methodologists. The current version of QUAID reliably handles the five problems below.

(1) Unfamiliar technical term. There is a word or expression that very few respondents would know the meaning of.

(2) Vague or imprecise predicate or relative term. The values of a predicate (i.e., main verb, adjective, or adverb) are not specified on an underlying continuum (e.g., *try, large*, *frequently*).

(3) Vague or ambiguous noun-phrase. The referent of a noun-phrase, noun, or pronoun is unclear or ambiguous (e.g., *items, amount, it, there*).

(4) Complex syntax. The grammatical composition is embedded, dense, structurally ambiguous, or not syntactically well-formed.

(5) Working memory overload. Words, phrases, or clauses impose a high load on immediate memory.

When a question is submitted to QUAID, there are three slots of information that get entered: Focal Question, Context, and Answer Options. The Focal Question is the main question that is being asked. The Answer Options (if any) are the response options that the respondent selects. The Context slot includes sentences that clarify the meaning of the question and instructions on how the respondent is supposed to formulate an answer. The content of the 3 slots is illustrated in the following question.

> FOCAL QUESTION: From the date of the last interview to December 31, did you take one or more trips or outings in the United States, of at least one mile, for the primary purpose of observing, photographing, or feeding wildlife?
>
> CONTEXT: Do not include trips to zoos, circuses, aquariums, museums, or trips for scouting, hunting, or fishing.
>
> ANSWER OPTIONS: YES_____ NO_____

QUAID's critique of each question is a list of problems it identified. For example, if a question had a one problem with each of the 5 categories, QUAID would print out the following five summary messages:

> UNFAMILIAR TECHNICAL TERM: The following term may be unfamiliar to some respondents: <unfamiliar technical term>
>
> IMPRECISE RELATIVE TERM: The following term refers implicitly to an underlying continuum or scale, but the point or value on the scale is vague or imprecise: <problematic term>
>
> VAGUE OR AMBIGUOUS NOUN-PHRASE: The referent of the following noun may be vague or ambiguous to the respondent: <problematic term>
>
> COMPLEX SYNTAX: The question is either ungrammatical or difficult to parse syntactically.
>
> WORKING MEMORY OVERLOAD: The question imposes a heavy load on the working memory of the respondent.

In addition to this short feedback, there are two additional levels of extended help that define each problem more completely and that give examples of particular problems. This extended help allows the survey methodologist to dissect and repair the problem with a particular question.

It is beyond the scope of this paper to provide the technical details of how QUAID identifies problems (see Graesser, et al., 1996, 1999; 2000). QUAID adopts both theoretical and empirical criteria when deciding whether a question has a problem. Regarding theory, the process of developing QUAID involved exploring a large space of features, modules, and mechanisms in computational linguistics that are potentially diagnostic for identifying a particular class of problems with questions. For example, in the case of syntax, there were metrics that computed the number of constituents at the top level of a parse, the number of subordinate clauses, the number of relative clauses, and so forth (see Jurafsky & Martin, 2000 for recent developments in computational linguistics and natural language processing in artificial intelligence). We used correlation analyses to explore which of the alternative measures of syntactic complexity best predicted the ratings of syntactic complexity that were provided by language experts. As another example, unfamiliar technical terms were identified by accessing computer lexicons that specify the frequency of words in the English language.

QUAID currently runs on a Pentium computer with a Linux operating system. The software includes a number of processing modules written in different computer languages (Java, LISP, C). QUAID is currently available on the web (www.psyc.memphis.edu/quaid.html), available to the public for free. However, individuals will not be able to use QUAID unless they provide us their names, address, email, telephone number, and other pertinent information. QUAID users must also agree to our analyzing their questions for research purposes, in exchange for their free use of the facility. The originator of the questions will be kept anonymous, in compliance with the ethical use of human subjects in research. We will use these questions for the evaluation and refinement of QUAID. QUAID currently handles only one question at a time, whereas a future version of QUAID will accommodate a set of survey questions.

Performance of QUAID when Compared to Human Experts as a Gold Standard

This section discusses how well QUAID fares in detecting problems with questions when using human experts as the standard for a correct identification of a problem. So truth is defined as the judgment of human experts. It should be noted that a problem spotted by human experts may be a continuous variable, rather than a discrete variable (i.e., problem versus no problem). Thus, a question Q is said to have problem P on a continuum that varies from 0 to 1.0; this we define as problem score. Intermediate values of the problem score reflect differences among experts and different strengths of problemhood within the judges. We considered different thresholds of the problem score when declaring whether there is a problem with a question. That is, a question Q was said to have problem P if the problem score of experts met or exceeded some threshold T.

Graesser et al. (2000) conducted a study that assessed how well experts can identify the five problems with questions. Experts evaluated a corpus of 550 questions on the five problems (2750 judgments altogether). The three experts were extensively trained on the problems with questions and had a graduate degree in a field that investigated the mechanisms of language, discourse, and/or cognition. The experts judged whether or not each question had each of the 5 problems. The following rating scale was used in making these judgments: 1 = definitely not a problem, 2 = probably not a problem, 3 = probably a problem, and 4 = definitely a problem. The problem score was computed as: (sum of expert ratings – 3) / 9. A question was defined as having a problem P if the problem score $\geq$ threshold T.

Eleven surveys were selected for testing QUAID. These included: *Hunting and Fishing Questionnaire*, third detailed interview, 1991 (form FH-3C); *Nonconsumptive User's Questionnaire*, Third Detailed Interview, 1991 (form FH-4C); *1993 Survey of Working Experience of Young Women* (form LGT-4161); 1996 *American Community Survey* (form ACS-1); *United States Census 2000 Dress Rehearsal* (form DX-2); *Adolescent Self-Administered Questionnaire: Survey of Program Dynamics* (form SPD-18008); 1998 *National Health Interview Survey Basic Module: Adult Core* (version 98.1); 1998 *National Health Interview Survey Basic Module:Household Composition* (version 98.1); 1998 *National Health Interview Survey: Child Prevention Module* (version 98.1); *Crime Incident Report: National Crime Victimization Survey* (form NCVS-2); *Survey of Program Dynamics: Adult Questionnaire*. These surveys were furnished by the United States Census Bureau.

Signal detection analyses were performed on the data after we classified questions as being problematic versus non-problematic for any given criterion threshold T. Using the terminology of signal detection theory, a target item is a question that human experts regard as a problem (given threshold T) whereas a nontarget item is a question that human experts regard as nonproblematic. The following metrics can then be computed.

Hit rate = p(computer sees problem | human sees problem)
False alarm rate (FA) = p (computer sees problem | human sees no problem)
d' score = computer's discriminative ability to identify problem, in theoretical standard deviation units

A high *d'* score means that the QUAID tool does an excellent job discriminating between questions that are problematic versus non-problematic, at least according to the standard of the human experts. The d' score is a pure measure of the ability of QUAID to discriminate problems with questions, after controlling for guessing biases. Another useful measure is called a **problem likelihood**, which is the proportion of questions that are classified as problematic according to the experts (given some threshold T on the problem scores).

There have been previous evaluations of QUAID on the corpus of 550 questions provided by the US Census Bureau (Graesser et al., 2000; Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, & Kreuz, 2000). These previous evaluations support the claim that QUAID has discriminative validity in identifying all five problems with questions, as defined by the experts. Table 1 summarizes the results of the evaluation reported in these studies. Table 1 presents the different performance measures for the 5 categories of problems with questions. These include the hit rates, false alarm rates, *d'* scores, and problem likelihood scores. We selected suitable threshold values of problem scores that optimized hit rates, *d'* scores and problem likelihood scores.

Table 1: Comparison of QUAID and human experts in detecting problems with questions

	Hit Rate	False alarm Rate	d' score	Problem likelihood
(1) Unfamiliar technical term	.86	.41	1.31	.09
(2) Vague or imprecise relative term	.94	.53	1.48	.10
(3) Vague or ambiguous noun-phrase	.95	.61	1.37	.04
(4) Complex syntax	.29	.03	1.33	.07
(5) Working memory overload	.29	.04	1.20	.08

Several conclusions can be drawn from the data in Table 1. First, the QUAID tool was able to discriminate problematic questions because the *d'* scores were significantly above zero. Second, the hit rates and false alarm rates had remarkably different patterns among the five classes of questions. The hit rates were quite high for the first 3 problem categories (.86 to .95), but so were the false alarm rates (.41 to .61). QUAID does a good job in detecting these classes of problems but at the expense of generating false alarms that may not be problematic under more careful analysis. So the survey methodologist would have many questions flagged as problems, but would have to spend extra time rejecting many questions that are not problematic. An improved QUAID needs to have computational methods of not being fooled by false alarms. In contrast, problem 4 (complex syntax) and problem 5 (WM overload) had low hit rates and extremely low false alarm rates. In these cases, QUAID needs to have more sensitive algorithms and metrics for picking up problematic questions. For all 5 problems, the problematic likelihood scores were quite low (ranging from .04 to .10). Thus, only 1 out of 10 to 25 questions suffered from a particular problem.

During the course of our research project, we have been exploring improved computational procedures for identifying problems with questions. We recently have been particularly interested in improving the complex syntax evaluator because it had previously shown a poor ability to detect problematic questions. In order to provide a more sensitive assessment, we desired a sample of questions that were more evenly split between problematic and nonproblematic questions. Therefore, we selected a sample of 94 questions from the original 550 questions in the question corpus; this restricted corpus had a higher incidence of problematic questions. First, we selected the top 50 problematic questions, using problem score measures that integrated over the 5 problems. Second, we randomly selected 50 questions from the sample of 550; 6 of these were in the first set of problematic questions, so we ended up with 94 questions in total.

Table 2 presents the recent performance evaluation of QUAID. The old version of QUAID is compared with the revised version of QUAID. Table 2 also contrasts a lower threshold (T = .33) with a higher threshold (T = 44) of problem scores. As the threshold gets higher, the greater extent to which expert judges believe there is a problem with a question. As the threshold increases, there automatically is a lower problem likelihood score; when averaging over the 5 question problems, the problem likelihood scores were .38 and .18 for the low versus high thresholds, respectively. Similarly, the *d'* scores generally increase as a function of higher thresholds (as do hit rates and false alarm rates). So when the experts have a stronger belief there is a problem with a question, the accuracy of QUAID shows a similar improvement.

The most interesting data contrasts the performance of the old versus the revised version of QUAID. We spent considerable effort improving the syntax component and that clearly paid off. The hit rates and *d'* scores increased dramatically for syntactic complexity. In the future, we plan on giving greater attention to working memory overload module, now that that there has been reasonable progress on syntactic complexity. This is because one aspect of working memory load consists of syntactic complexity. In contrast to the dramatic increases in the performance of the syntax module, there were modest gains in unfamiliar technical terms and vague/ambiguous noun-phrases. The vague and imprecise relative term component is almost finished, so improvements are not anticipated on that module.

There is some question of what performance index to maximize in our QUAID tool. We plan on having two versions of QUAID, one that maximizes hit rates and one that maximizes *d'* discrimination. If we maximize on hit rate, then QUAID will identify most of the problems, but at the cost higher false alarms. So QUAID will alert the survey methodologist that there might be a problem, but the survey methodologist will have to make frequent decisions that these potential problems should be dismissed. If we maximize on *d'* scores, then QUAID will be identifying problems less often, but the decisions will be more accurate. The use of the different versions will depend on the goals of the survey methodologist (i.e., completeness versus timeliness).

There is one fundamental problem with using the expert ratings as the gold standard of spotting problems with question interpretation. The experts have only moderate agreement on the identification of these problems (see Graesser et al., 2000) and they miss many of the subtle analyses of language, discourse, and world knowledge. Therefore, we need a more objective measure of identifying questions with particular problems. Our hope is that eye tracking data will provide a more objective measure. Therefore, we conducted a study on eye tracking during question answering. This is reported in the next section.

Table 2: Recent comparison of QUAID and human experts in detecting problems with questions

	Hit rate	False alarm rate	d’ score	Problem likelihood
(1) Unfamiliar technical term				
Threshold = .33				
Old QUAID	.82	.44	1.06	.30
Revised QUAID	.96	.71	1.20	.30
Threshold = .44				
Old QUAID	.93	.49	1.50	.15
Revised QUAID	1.00	.75	1.64	.15
(2) Vague or imprecise relative term				
Threshold = .33				
Old QUAID	.77	.50	.74	.38
Revised QUAID	.77	.50	.74	.38
Threshold = .44				
Old QUAID	.90	.52	1.29	.22
Revised QUAID	.90	.52	1.29	.22
(3) Vague or ambiguous noun-phrase				
Threshold = .33				
Old QUAID	.88	.64	.82	.46
Revised QUAID	.90	.56	1.13	.46
Threshold = .44				
Old QUAID	1.00	.73	1.70	.08
Revised QUAID	1.00	.71	1.76	.08
(4) Complex syntax				
Threshold = .33				
Old QUAID	.28	.16	.42	.41
Revised QUAID	.62	.38	.62	.41
Threshold = .44				
Old QUAID	.39	.15	.76	.24
Revised QUAID	.91	.34	1.75	.24
(5) Working memory overload				
Threshold = .33				
Old QUAID	.40	.12	.90	.37
Revised QUAID	.40	.12	.90	.37
Threshold = .44				
Old QUAID	.63	.12	1.50	.20
Revised QUAID	.63	.12	1.50	.20

Eye Tracking While Answering Questions

The collection of eye tracking data provides a different method of diagnosing problematic questions with respect to question interpretation. Eye tracking patterns serve as a sensitive index of on-line comprehension processes. If a question is difficult to comprehend, then there should be a high density of multiple fixations on words and regressive eye movements. Words that are difficult to interpret should have long fixation times. We collected eye tracking data in order assess whether the problems identified by QUAID are manifested in eye movements and gaze durations.

We conducted a study on 9 college students who read and answered 69 questions selected from the corpus of 550 survey questions. The 69 questions included 45 problematic questions and 24 random questions. We had to exclude questions that were too long to fit on a computer screen. The eye tracking equipment was an Applied Science Laboratory Model 501 eye tracker with a head mounted device. Thus, the respondents could move their heads while reading and answering the questions.

During each trial, the participant advanced to the next question by hitting a bar in presence of a READY signal. Then the question appeared on the screen. The participant read the question and answered the question aloud. We recorded the eye tracking data while they read the question, audio recorded their answers, and videotaped the computer screen. The eye tracking portion of the study lasted 30 minutes, 10 for calibration of the eyes and 20 minutes for collecting data on the 69 questions. There were 6 different random orders of the questions. After collecting the eye tracking data, the participants completed a Wechsler Abbreviated Scale of Intelligence (Psychological Corporation, 1999) and an information sheet about demographic information and university training.

One index of comprehension difficulty is multiple fixations on a word. If comprehension runs smoothly, the reader would move ahead in a linear fashion, with only one eye fixation per word. However, there will be multiple fixations and regressive eye movements to the extent that there are problems interpreting words, noun-phrases, clauses, and sentences. The index of comprehension difficulty was therefore scored as number of eye fixations per word, given that there was at least one fixation on the word.

Table 3 shows this fixation frequency index for the content words of one of the questions. Content words include nouns, pronouns, adjectives, and main verbs. The function words and other minor words were not counted because they are known to have short fixation times. The fixation frequencies clearly increase as the readers progress further in the sentences, when the working memory load is higher and the syntactic complexity is more taxing. The mean fixation frequencies were 1.14, 1.44, 2.08, and 2.57 for the content words on lines 1, 2, 3, and 4, respectively. Table 4 shows that gaze durations on individual words show the same pattern. The mean daze durations (measured in milliseconds) are 225, 290, 397, and 633 milliseconds for lines 1, 2, 3, and 4, respectively.

We are currently analyzing fixation frequencies and gaze durations of the words in the 69 questions. The mean fixation frequency per content word (and mean gaze duration) should be significantly higher for the problematic questions than the nonproblematic questions. Moreover, gaze durations should be comparatively high for unfamiliar technical terms, unclear relative terms, and vague or ambiguous noun-phrases. Regressive eye movements should occur at points in the sentence when the syntactic complexity and/or working memory load are high. These predictions are currently being tested in our laboratory.

Table 3: Fixation frequencies for content words in an example question.

```
       1.27         1.00 1.27    1.00
Do the people who do not live and eat

   2.52 1.00  1.33 1.00   1.33
at your house have direct access from the

 1.70                2.00    2.55
outside or through a common hallway to a

 2.77    2.70    2.25
separate living quarter?

1.00  1.33 5.73      12.10 3.00
Yes;  No;  Refused;  Don’t know
```

Table 4: Gaze durations for content words in an example question.

```
       310          190  220     180
Do the people who do not live and eat

   500  240   290  200    220
at your house have direct access from the

 210                 400     580
outside or through a common hallway to a

 760     490     650
separate living quarter?

120   290  880       2600  530
Yes;  No;  Refused;  Don’t know
```

References

Fowler, F.J., & Cannell, C.F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In N. Schwarz & S. Sudman (Eds.), Answering questions: Methodology for determining cognitive and communicative processes in survey research (pp. 15-36). San Francisco, CA: Jossey-Bass.

Graesser, A.C., Bommareddy, S., Swamer, S., & Golding, J.M. (1996). In N. Schwartz and S. Sudman (Eds), Answering questions: Methodology for determining cognitive and communicative processes in survey research (pp. 143-174). San Francisco: Jossey-Bass.

Graesser, A.C., Kennedy, T., Wiemer-Hastings, P., & Ottati, V. (1999). The use of computational cognitive models to improve questions on surveys and questionnaires. In M.G. Sirken, D.J. Hermann, S. Schechter, N. Schwarz, J.M. Tanur, & R. Tourangeau (Eds.), Cognition and survey methods research (pp. 199-216). New York: Wiley.

Graesser, A.C., Wiemer-Hastings, K., Kreuz, R., Wiemer-Hastings, P., & Marquis, K. (2000). QUAID: A questionnaire evaluation aid for survey methodologists. Behavior Research Methods, Instruments, and Computers, 32, 254-262.

Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., & Kreuz, R. (2000). The gold standard of question quality on surveys: Experts, computer tools, versus statistical indices. Proceedings of the Section on Survey Research Methods of the American Statistical Association. (pp. 459-464).

Groves, R.M. (1989). Survey errors and survey costs. New York: Wiley.

Jobe, J.B., & Mingay, D.J. (1991). Cognition and survey measurement: History and overview. Applied Cognitive Psychology, 5, 175-192.

Jurafsky, D., & Martin, J.H. (2000). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, NJ: Prentice. .

Lessler, J.T., & Forsyth, B.H. (1996). A coding system for appraising questionnaires. In N. Schwartz and S. Sudman (Eds), Answering questions: Methodology for determining cognitive and communicative processes in survey research (pp. 259-292). San Francisco: Jossey-Bass.

Lessler, J.T., & Kalsbeek, W. (1993). Nonsampling error in surveys. New York: Wiley.

Lessler, J.T., & Sirken, M.G. (1985). Laboratory-based research on the cognitive aspects of survey methodology: The goals and methods of the National Center for Health Statistics study. Milbank Memorial Fund Quarterly/Health and Society, 63, 565-581.

Psychological Corporation (1999). Wechsler Abbreviated Scale of Intelligence. San Antonio, TX: Harcourt Brace.

Schober, M.F., & Conrad, F.G. (1997). Does conversational interviewing reduce survey measurement error? Public Opinion Quarterly, 60, 576-602.

Schwarz, N. & Sudman, S. (1996)(Eds.), Answering questions: Methodology for determining cognitive and communicative processes in survey research. San Francisco, CA: Jossey-Bass.

Sirken, M.G., Hermann, D.J., Schechter, S., Schwarz, N., Tanur, J.M., & Tourangeau, R. (1999)(Eds.), Cognition and survey methods research. New York: Wiley.

Sudman, S., Bradburn, N.M., & Schwarz, M. (1995). Thinking about answers: The application of cognitive processes to survey methodology. San Francisco: Jossey-Bass.

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T.J. Jabine, M.L. Straf, J,M. Tanur, and R. Tourangeau (Eds.), Cognitive aspects of survey methodology: Building a bridge between disciplines. Washington, DC: National Academy of Sciences.

Willis, G.B., DeMaio, T.J., & Harris-Kojetin, B. (1999). Is the bandwagon headed to the methodological promisef land? Evaluating the validity of cognitive interviewing techniques. In M.G. Sirken, D.J. Hermann, S. Schechter, N. Schwarz, J.M. Tanur, & R. Tourangeau (Eds.), Cognition and survey methods research (pp. 133-153). New York: Wiley.

Discussion: "A Computer Tool to Improve Questionnaire Design"

Theresa J. DeMaio
U. S. Census Bureau

The QUAID model is certainly a computational challenge, and interesting from the point of view of cognitive linguistics. And from the perspective of a survey methodologist, it has the potential to be a useful diagnostic tool. I'd like to focus my comments today on three aspects of the paper: the choice of human experts for comparison of problem detection with the QUAID computer model, the results of the comparison with human experts, and the search for a new gold standard.

Problem Detection by the Computer Model and Human Experts

In developing the computer model, Graesser and his colleagues have defined the truth as the correct identification of problems by human experts. In the results they report here, and in results they have reported in previous papers (Graesser, Kennedy, Wiemer-Hastings and Ottati, 1999; Graesser, Wiemer-Hastings, Wiemer-Hastings, and Kreuz, 2000), they compare the performance of the computer model in identifying questionnaire problems of specific types against the performance of human judgement. Their judges had graduate degrees in a field that investigated the mechanisms of language, discourse, and/or cognition. I think this would be a relevant criterion for the judges if the tool was for a purpose related to these disciplines. But since this is a tool to be used by survey methodologists, it would be much more appropriate if the results of the computer model were compared to evaluations of the same questions by questionnaire design experts in the field of survey methodology, who have familiarity with and expertise in identifying problems with survey questions experienced by respondents.

Their use of language experts makes an implicit assumption that the use of language in survey questions is the same as all other questions, and I think we know that this is not necessarily the case. "How many people live in your house or apartment?" may be interpreted differently when a survey interviewer talks to a recent illegal immigrant than when the immigrant is speaking to a friend or relative. Perhaps the results of the comparison of the computer with language experts would be the same as a comparison with questionnaire design experts – I wouldn't want to make predictions about the extent of any differences – but I would definitely feel more comfortable about the utility of QUAID as a diagnostic tool for surveys if I could see some data about how it compares to survey methodologists' evaluations of survey questions.

Results of the Comparison with Human Experts

I view QUAID as a preliminary questionnaire design tool, one that would be useful in identifying major problems in draft questionnaires during the initial questionnaire development process. As such, I don't see it as a competitor to either verbal protocols from respondents during think-aloud interviews or coding of the interaction between respondents and interviewers during field interviews. To my mind, its use would precede either of these two methods. It is more similar to an expert review and cognitive appraisal methods. So a questionnaire designer might want to make a choice between QUAID, expert reviews, or questionnaire appraisals (Lessler & Forsyth, 1996; Willis & Lessler, 1999) in the early stages of questionnaire development.

In this context, I was interested in the last column of Table 1, in which the authors note the problem likelihood, that is, the likelihood that each of the five problems of interest was identified in a question. These scores ranged from .04 (which means that a problem of this kind was detected in 1 out of 25 questions) to .10 (which means that a problem was detected in 1 out of 10 questions). Summed together, the problem likelihood that any of these five problems would be identified is .38, or 4 out of every ten questions. This is an upper bound, since more than one of these problems could apply to a single question. These scores seem very low to me. The questions came from 11 survey questionnaires conducted by the Census Bureau, and I'd like to think that Census Bureau surveys are this good, but I don't really believe it.

Research has been conducted on the expert review methodology and the questionnaire appraisal system, which I said I view as QUAID's main competitors, and these methods identify a much higher percentage of questionnaire problems. In 1991, Presser and Blair (1994) conducted experimental research in which expert reviews were conducted, along with other pretest methods. Two independent expert reviews were conducted on a 140-item questionnaire. One of the expert reviews identified 182 problems, and the other identified 140 problems.

More recently, Jennifer Rothgeb and her colleagues (Rothgeb, Willis, & Forsyth, 2001) presented a paper at AAPOR last month in which they compared expert reviews with questionnaire appraisals. For an 83-item questionnaire that was rated on a problem scale of 0 to 3, the expert review yielded a mean problem score of 1.55 (that is, items were found to be problematic half the time) and the questionnaire appraisal yielded a mean problem score of 2.93. In other words, almost all the time, items were found to be problematic.

None of these comparisons are exactly equivalent, but there is enough similarity in the objectives and methods that I would expect a higher problem yield from QUAID. The greatest portion of the problems identified in both these research efforts dealt with question meaning, and four out of the five problems included in QUAID deal with question meaning as well. One difference between the QUAID results and the other research is that survey experts conducted the expert reviews and the questionnaire appraisals, while this was not the case for the gold standard for the QUAID evaluation. Perhaps there is some unique expertise that questionnaire design experts bring to bear when evaluating survey questions that is different than the experience of linguists.

Since questionnaire appraisals and expert reviews identify more problems than cognitive interviews or behavior coding (Presser and Blair, 1994; Rothgeb et al, 2001), it seems that the knowledge of the survey experts leads them to identify potential problems that are not evidenced by respondents themselves. This is the equivalent of the False Alarm rate calculated by Graesser and his coauthors. I am not bothered by that as much as I am by the relatively low problem likelihood. I would urge them to focus their attempts to improve QUAID in that area. My perspective on this comes from my view that this is a tool for the initial stages of questionnaire development. Suspected problems that don't turn out to be serious can be addressed, but serious problems that never get detected can jeopardize a data collection effort.

What Should the Gold Standard Be?

Graesser and his colleagues have a question about what the gold standard should be for comparing the QUAID results against. They think that judgements of experts in language, cognition and world knowledge are problematic because they are not stable across multiple experts. They also consider getting feedback from respondents, but find this to be lacking because their judgements can be "insensitive to problems that allegedly exist." So they are moving on to eye tracking as an objective measure of on-line comprehension processes.

Eye tracking is an interesting notion. Cleo Redline, one of my colleagues at the Census Bureau, is investigating its use as a vehicle to evaluate visually administered instruments, and she also recently presented a paper at AAPOR (Redline and Lankford, 2001). Her research to date has focused on skip instructions on paper questionnaires, and she is planning to expand to studies of response to automated questionnaires and websites. Her concentration is on navigational issues, and keeping track of respondents' eye movements as they find their way through a questionnaire or a website makes intuitive sense.

I wonder, however, whether this technique can really be an objective measure of comprehension, as Graesser asserts. It seems to me that a big assumption must be made to state that multiple fixations on a word is an indicator of comprehension difficulty. That might be the case, of course, but it also could be that the respondent is absorbing the content of the major concepts of the question without difficulty. Furthermore, if the objective of this gold standard is to spot problems of the five types that QUAID can reliably detect, it is not clear how the eye tracking methodology can achieve this. I think some demonstration of the validity of this criterion measure is necessary before it is used in this way.

My view is that, since the stated objective of QUAID is to be "a computer tool that assists survey methodologists who want to improve the wording, syntax, and semantics of questions on surveys and questionnaires," the perspective of the survey methodologist is a logical place to start in assessing how well the computer tool works in terms of meeting its objective. QUAID would be useful if it could provide an easy, automated means for providing the same types of information about questionnaire problems that can already be obtained with a lot more effort through other means. There are probably other ways to look at the issue, and I would be open to other standards if they could improve on the information that is already available, but from my perspective that is the minimum standard that would make QUAID a viable method for testing survey questions.

In conclusion, I would encourage the authors to continue their development of the QUAID program and make it into a useful tool for questionnaire designers.

References

Graesser, A.C., Kennedy, T., Wiemer-Hastings, P., and Ottati, V. (1999) The use of cognitive models to improve questions on surveys and questionnaires. In M.G. Sirken, D.J. Hermann, S. Schechter, N. Schwarz, J.M. Tanur, and R. Tourangeau (Eds.), Cognition and Survey Research (pp. 199-216). New York: Wiley.

Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., and Kreuz, R. (2000) "The gold standard of question quality on surveys: Experts, computer tools, versus statistical indices," Proceedings of the American Statistical Association (Survey Research Methods Section), Alexandria, VA: American Statistical Association, forthcoming.

Lessler, J.T., and Forsyth, B. H. (1996) A coding system for appraising questionnaires. In N. Schwarz and S. Sudman (Eds.), Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research (pp. 389-402). San Francisco: Jossey-Bass.

Presser, S. and Blair, J. (1994) Survey pretesting: Do different methods produce different results? In P.V. Marsden (ed.), Sociological Methodology, Vol. 24, Washington, DC: American Sociological Association, pp. 73-104.

Redline, C.D. and Lankford, C.P. (2001) "Eye movement analysis: A new tool for evaluating the design of visually administered instruments (paper and web), paper prepared for presentation at the Annual Meetings of the American Association for Public Opinion Research, Montreal, Canada, May 2001.

Rothgeb, J.R., Willis, G.B., and Forsyth, B. (2001) "Questionnaire pretesting methods: Do different techniques and different organizations produce similar results?", paper prepared for presentation at the Annual Meetings of the American Association for Public Opinion Research, Montreal, Canada, May 2001.

Willis, G.B. and Lessler, J.T. (1999) The question appraisal system: A guide for systematically evaluating survey question wording. Final Report submitted to the Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion. Research Triangle Institute.

Session 4

Cognitive Issues in Designing Web Surveys

Social Presence in Web Surveys

Mick P. Couper
University of Michigan
P.O. Box 1248, Ann Arbor,
MI 48106 USA
+1 734 647 3577

Roger Tourangeau
University of Michigan
P.O. Box 1248, Ann Arbor,
MI 48106 USA
+1 301 314 7984

Darby M. Steiger
The Gallup Organization
One Church Street, Suite 900
Rockville, MD 20850
+1 301 309 9439

ABSTRACT

Social interface theory has widespread influence in the field of human-computer interaction. The basic thesis is that humanizing cues in a computer interface can engender responses from users similar to human-human interaction. In contrast, the survey interviewing literature suggests that computer administration of surveys on highly sensitive topics reduces or eliminates social desirability effects, even when such humanizing features as voice are used.

In attempting to reconcile these apparently contradictory findings, we varied features of the interface in a Web survey (n=3047). In one treatment, we presented an image of 1) a male researcher, 2) a female researcher, or 3) the study logo at several points. In another, we varied the extent of personal feedback provided. We find little support for the social interface hypothesis. We describe our study and discuss possible reasons for the contradictory evidence on social interfaces.

Keywords
Social interfaces, Web surveys, social desirability

INTRODUCTION

Social interface theory [8][11][21] appears to be generating much interest in the world of human-computer interaction. Much of the support for this perspective comes from laboratory-based studies.

A growing number of laboratory experiments suggest that relatively subtle cues (such as "gendered" text or simple inanimate line drawings of a face) in a computer interface can evoke reactions similar to those produced by a human, including social desirability effects. Nass, Moon and Green [17], for example, conclude that the tendency to stereotype by gender can be triggered by such minimal cues as the voice on a computer. Based on the results of a series of experiments that varied a number of cues in computer tutoring and other tasks, Nass and colleagues [9][16][17][18] argue that computer interfaces (even the words used in a text-based tutoring task) can engender reactions from subjects similar to those evoked by interactions with other people. Their central thesis is that people treat computers as social actors not as inanimate tools (see also [3]).

Additional support for the hypothesis that a computer interface can function as a virtual human presence comes from a study by Walker, Sproull, and Subramani [27]. They administered questionnaires to people using either a text display or one of two talking-face displays to ask the questions. Those interacting with a talking-face display spent more time, made fewer mistakes, and wrote more comments than did people interacting with the text display. However, people who interacted with the more expressive face liked the face and the experience less than those who interacted with the less expressive face. In a subsequent experiment, Sproull and colleagues [23] varied the expression of a talking face on a computer-administered career counseling interview; one face was stern, the other more pleasant. The faces were computer-generated images with animated mouths. They found that: "People respond to a talking-face display differently than to a text display. They attribute some personality attributes to the faces differently than to a text display. They report themselves to be more aroused (less relaxed, less confident). They present themselves in a more positive light to the talking-face displays." (p. 116) (see also [20]).

If the social interface theory is correct, it has important implications for the survey research industry for several reasons: 1) There is an increasing trend toward the use of computer-assisted interviewing, and especially the use of the World Wide Web, for administration of surveys [4][5]. 2) More and more surveys include sensitive questions (on sexual behavior, drug use, etc.), raising concerns about social desirability effects and interviewer influences. 3) Concomitant with the above, there is an increasing move towards the using of computer-assisted self-interviewing (CASI) methods, whereby the respondent interacts directly with the computer to answer questions. The most recent manifestation of this trend is the development of audio-CASI, in which the respondent listens to the questions read over headphones using a digitized voice, and enters the responses into the computer. A number of studies have compared CASI and audio-CASI to alternative approaches in field-based experiments. The general finding is that CASI methods (including audio-CASI) reduce social desirability distortions (i.e., increase reporting of sensitive information) over both interviewer-administered and paper-based self-administered methods [24]. Some have gone so far as to argue that voice does not matter when asking questions about sexual behavior (e.g., [25][26]), although these claims have not been empirically verified.

These results appear to contradict the findings of the social interface researchers. If subtle humanizing cues do indeed influence the behavior of computer users, we would fully expect the gender of the voice to affect the answers given to survey questions on topics such as gender attitudes and sexual behavior. Given the increasing use of multimedia tools on the Web, the addition of a variety of humanizing visual and/or aural cues, as is possible in Web surveys, may negate or at least mitigate the beneficial effects of self-administration, especially for items of a sensitive nature. It is thus important to explore the apparent contradiction between the social interface and survey methods work, and attempt to bring these two strands of research together.

There are several differences between the two literatures that could account for the discrepant results. For one, virtually all of the social interface research has been conducted in laboratory settings with students as volunteer subjects. In contrast, the survey-based findings are from probability samples of broader populations (e.g., teenage males, women 15-44, adult U.S. population). In the former, the number of subjects is typically measured in tens or scores while, in the latter, sample sizes go up to the thousands. The measurement settings also differ considerably. The social interface work is typically done in a laboratory setting, free from distractions and with privacy ensured. Most of the CASI surveys are conducted in the respondent's home with an interviewer present, and sometimes with other family members home at the time. The perceived threat from disclosure varies greatly across the two settings. The more sterile, controlled environment of the laboratory may well focus subjects' attention on the experimental manipulation more than in an uncontrolled real-world setting with many potential distractors and less expectation of experimental manipulation. Furthermore, the measurement devices differ considerably between the two approaches. The social interface experiments often use subjects' performance on a computer task as the dependent measure. When questionnaire measures are used, they are typically self-reports of social desirability or impression management. The findings from the survey world are based on overt measures of highly sensitive behaviors (e.g., abortions, number of sex partners, engagement in high risk sexual behaviors, illicit drug use, etc.).

We obviously cannot address all these issues and resolve the controversy in a single study. We are engaged in a program of research to explore the issue of the effect of interface design and social interface features on survey responses. Work currently underway involves experiments on the effect of virtual interviewers (talking heads) on racial attitudes, manipulation of voice (male/female) in audio-CASI surveys, manipulation of privacy effects on self-disclosure in text-CASI versus audio-CASI surveys, and the effect of interface features on social desirability distortions in Web and interactive voice response (IVR) surveys. In this paper we report on the Web survey experiment we conducted as part of this broader research agenda.

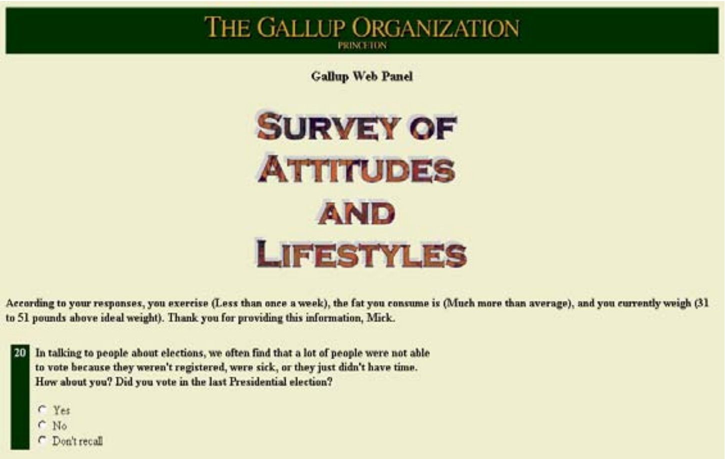
METHODS

We carried out two studies that examined the impact of characteristics of the interface on the responses obtained in a Web survey. Our first study compared six versions of a Web survey administered to 202 participants in a Web panel maintained by the Gallup Organization. The second study compared the same six versions of the survey in a much larger sample of Web users purchased from a commercial vendor, Survey Sampling, Inc. (SSI). Given that the design of the survey was identical across versions, and the findings were very similar, we focus on the larger sample from SSI here.

Experimental Manipulation

The different versions of the Web questionnaire differed along two dimensions--the degree that the program presented personalizing cues and the degree that it seemed to interact with the respondent. At several points in the questionnaire, the personalized versions of the questionnaire displayed a picture of one of the male researchers, or one of the female researchers. A comparison version of the questionnaire presented the logo for the study, instead of the investigators' picture. Along with the pictures, the program displayed relevant statements from the investigator: "Hi! My name is Roger Tourangeau. I'm one of the investigators on this project. Thanks for taking part in my study." The high interaction versions of the questionnaire used the first person in introductions and transitional phrases (e.g., "Thanks, [name]. Now I'd like to ask you a few questions about the roles of men and women") and occasionally echoed back to the respondents their earlier answers ("According to your responses, you exercise once daily ..."). The low interaction versions used more impersonal language ("The next series of questions is about the roles of men and women") and gave less tailored feedback ("Thank you for this information"). Examples of these designs are shown in Figures 1-3 below.

This resulted in a 3×2 experiment, fully crossing the two dimensions of social presence we manipulated. We randomly assigned respondents to one of the six cells in the design, as shown in Table 1.



THE GALLUP ORGANIZATION
PRINCETON

Gallup Web Panel

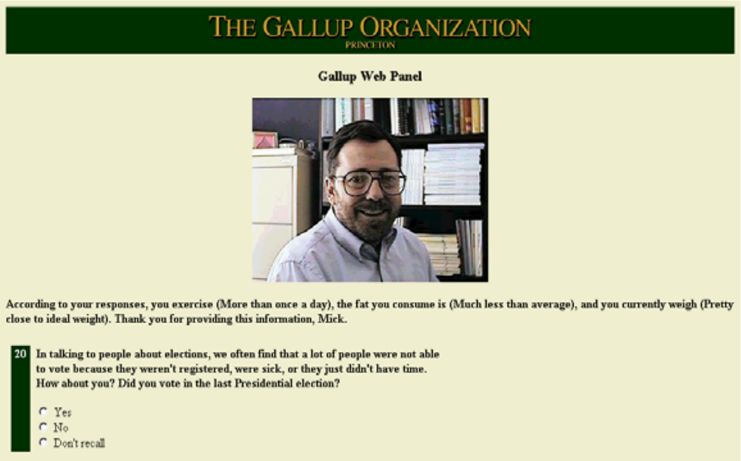
SURVEY OF ATTITUDES AND LIFESTYLES

According to your responses, you exercise (Less than once a week), the fat you consume is (Much more than average), and you currently weigh (31 to 51 pounds above ideal weight). Thank you for providing this information, Mick.

20 In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, were sick, or they just didn't have time. How about you? Did you vote in the last Presidential election?

○ Yes
○ No
○ Don't recall

Figure 1: Logo and Personal Feedback



THE GALLUP ORGANIZATION
PRINCETON

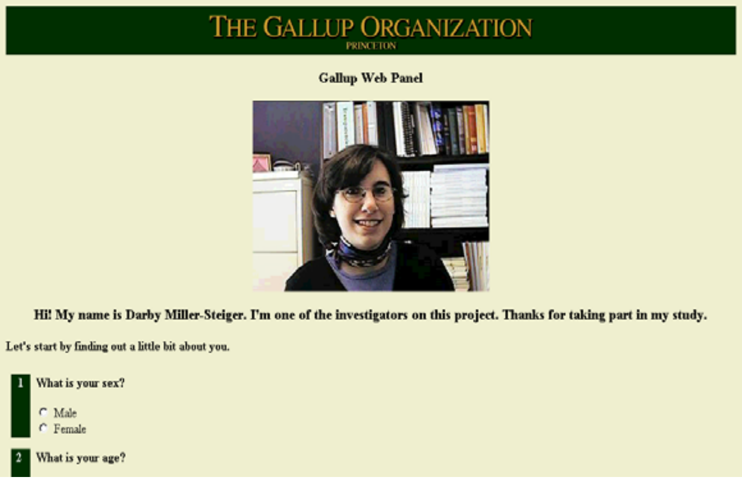
Gallup Web Panel

According to your responses, you exercise (More than once a day), the fat you consume is (Much less than average), and you currently weigh (Pretty close to ideal weight). Thank you for providing this information, Mick.

20 In talking to people about elections, we often find that a lot of people were not able to vote because they weren't registered, were sick, or they just didn't have time. How about you? Did you vote in the last Presidential election?

○ Yes
○ No
○ Don't recall

Figure 2: "Male" Interface and Personal Feedback



THE GALLUP ORGANIZATION
PRINCETON

Gallup Web Panel

Hi! My name is Darby Miller-Steiger. I'm one of the investigators on this project. Thanks for taking part in my study.

Let's start by finding out a little bit about you.

1 What is your sex?

○ Male
○ Female

2 What is your age?

Figure 3: "Female" Interface

Questionnaire

The survey questionnaire contained the following types of items:

- Gender attitudes: 8 items from Kane & Macauley's [10] study regarding the roles of men and women (e.g., *Thinking about men as a group, do you think men have too much influence, about the right amount of influence, or too little influence in society?*).
- Socially undesirable behaviors: 5 items on drinking and illicit drug use, 3 less-sensitive items on diet and exercise.
- Socially desirable behaviors: items on voting and church attendance.
- Self-reported social desirability: 16 items from the Marlowe-Crowne Social Desirability (SD) Scale [6] and the 20-item Impression Management (IM) scale from the Balanced Inventory of Desirable Responding (BIDR)[20].
- Trust: 3 items on trust (e.g., *Most people can be trusted*).
- Debriefing questions: 9 items to assess social presence and evaluate the interview experience (e.g., *How much was this interview like an ordinary conversation? How much was it like dealing with a machine?*).
- Demographic questions.

We included the gender attitude items to see whether our attempt to personalize the interface produced "deference" effects paralleling the gender-of-interviewer effects with actual interviewers – that is, more pro-feminist responses with the "female" than with the "male" interface. The items on diet, exercise, drinking, drug use, voting, and attendance at church were all included to test the hypothesis that humanizing the interface (both by personalizing it and by making it more interactive) would increase the number of socially desirable responses and decrease the number of socially undesirable responses. The SD and IM items have been used for similar purposes (to measure socially desirable responding) in the work by Nass and colleagues, and we included them in our studies for the sake of comparability. We included the trust items to see whether the impact of the experimental variables was greater among those low in trust (as found by Aquilino and LoSciuto [1]). The demographic items were included as a check on the randomization and to assess subgroup differences. On average, the questionnaire took about 15 minutes to complete.

Hypotheses

Consistent with the social interface theory, our hypotheses were that increasing the social nature of the Web survey interaction, whether by personalization or interaction, would yield: 1) higher self-reports of social desirability and impression management, and 2) lower reports of socially undesirable behaviors (drug use, drinking, fat consumption) and higher reports for socially desirable behaviors (church attendance, voting, exercise). We also hypothesized that the "male" interface would elicit less positive attitudes toward women, while the "female" interface would yield more positive attitudes, with the neutral logo occupying a middle position.

Sample Design and Implementation

The frame for the SSI sample consists of more than seven million e-mail addresses of Web users. SSI has compiled this list from various sources; in each case, visitors to specific Web sites agreed to receive messages on a topic of interest. SSI selected a sample of 15,000 e-mail addresses and sent out an initial e-mail invitation to take part in “a study of attitudes and lifestyles.” The e-mail invitation included the URL of the Web site where our survey resided and a PIN number (which prevented respondents from completing the questionnaire more than once). After ten days, SSI sent a second reminder e-mail to sample persons who had not yet completed the survey. A total of 3,047 sample members completed the questionnaire, for a response rate of approximately 20%. (Less than 1% of the e-mails bounced back as invalid addresses.) Another 434 persons (3% of the sample) began the survey but broke off without finishing it. We focus here on the respondents who completed the survey. The number of completed cases per cell is shown in Table 1.

Table 1. Number of Subjects per Cell

Personalizing Cues	**Interaction**		
	Low	High	Total
Logo	492	502	994
Male picture	529	529	1058
Female picture	501	492	993
Total	1522	1523	3047

The number of cases we obtained far exceed that for most of the experimental studies on social interfaces (typically 10-20 subjects per cell). Statistical power to detect effects of the manipulations should not be a problem in our study. Furthermore, the respondents to our survey represent a much more diverse group than is typically found in laboratory-based experiments.

ANALYSIS AND RESULTS

We created a number of scales for the key measures in our study. For the social desirability scale we assigned a score of 1 to every answer that represented socially desirable responding, and a 0 to every response that did not. This yielded a scale with a range of 0 to 16, with a high score indicating a greater tendency towards socially desirable responding. We used the same strategy for the impression management scale, creating a summary score ranging from 0 to 20, again with a high score indicating greater impression management. For the gender attitude items, we created a scale that combined responses across the eight items, by scoring responses to each item in a consistent direction and then summing across the items. The resultant scale ranged from 8 to 24, with a high score indicating pro-feminist or more egalitarian attitudes. Similarly, we created an index to combine answers to a number of the sensitive questions. Our index was the number of embarrassing answers given in response to those questions; the index varied from 0 to 7. Respondents got a point each if they reported they consumed more dietary fat than the average person, were 20 pounds or more over their ideal weight, drank alcohol almost every day (or more often), had smoked marijuana, had used cocaine, did not vote in the last election, and did not attend church in the last week.

The results for each of these scales by each of the two experimental conditions are presented in Table 2. None of the effects reach statistical significance (p>.10) with the exception of the effect of personalization on gender attitudes, to which we return later. To perform a stronger test of the social interface hypothesis, we combined the two experimental conditions, and contrasted the high social interface group (high interaction, and male/female picture) with the low social interface group (low interaction, logo). The differences in means again do not approach statistical significance. We tried a variety of alternative specifications, including control variables, and interaction terms, but the findings essentially remain the same.

Table 2. Scale Means by Condition (Standard Errors in Parentheses)

	Social Desirability	Impression Management	Gender Attitudes	Sensitive Admissions
Interaction	n.s.	n.s.	n.s.	n.s.
Low interaction	7.87 (0.14)	8.84 (0.19)	18.25 (0.16)	3.27 (0.07)
High interaction	7.83 (0.10)	8.91 (0.13)	17.98 (0.11)	3.30 (0.05)
Personalization	n.s.	n.s.	p<.05	n.s.
Logo	7.95 (0.10)	8.87 (0.13)	18.09 (0.12)	3.27 (0.05)
Male Picture	7.77 (0.09)	8.73 (0.13)	17.77 (0.11)	3.21 (0.05)
Female Picture	7.85 (0.09)	8.84 (0.13)	18.19 (0.11)	3.31 (0.05)

Table 3. Percentages on Behavior Variables by Condition

	% Used Cocaine in Lifetime	% Smoked Marijuana in Last Year	% Drink Daily or Almost Daily	% Attended Church Last Week	% Voted in Last Election
Interaction	n.s.	n.s.	n.s.	n.s.	n.s.
Low interaction	14.2	10.7	7.8	23.3	53.2
High interaction	15.3	10.2	7.7	25.7	52.2
Personalization	n.s.	n.s.	n.s.	n.s.	p<.05
Logo	15.4	10.8	7.4	23.2	52.8
Male Picture	14.7	9.9	8.0	24.3	55.3
Female Picture	14.2	10.5	7.7	26.1	49.7

There were a few scattered findings for some of the individual sensitive items. We include a few examples of both socially undesirable and socially desirable behaviors in Table 3. For reports about voting, the personalization variable had a significant impact (X^2=6.35, df=2, *p* <.05). Contrary to expectation, the respondents who got the female picture were least likely to say they had voted in the most recent election, while those who got the male picture were most likely to say they had voted. In general, though, neither the level of personalization nor the level of interaction had much effect on reports about sensitive topics.

The only expected effect that found support in our data was related to gender attitudes (see Table 2). We expected respondents of both sexes to report the most pro-feminist attitudes when the program displayed pictures and messages from the female investigator and the least pro-feminist attitudes when the program displayed the pictures and messages from the male investigator. We expected the group who got the survey logo to fall in between the other two. This pattern was apparent, and reached statistical significance (*F*=5.52, df=1,3028, *p*<.05).

One explanation for the significant gender effect could relate to the "mere presence" hypothesis from studies of prejudice. Research on race-of-interviewer effects [7][9] has found that racial stereotypes can be "primed" simply by presenting an image of the target group. This view is an alternative to the "racial deference" or "polite stranger" hypotheses [2][22] which suggest that people avoid articulating negative stereotypes in the presence of another person, particularly a member of the target group, out of politeness. This latter view is more akin to the social presence model. The fact that the female picture elicits the most pro-feminist attitudes, and the male picture the least, with the logo occupying a middle position, may suggest support for the "mere presence" theory of stereotypes, rather than for a social presence interpretation. This obviously deserves further research attention.

DISCUSSION AND CONCLUSIONS

Our results were much weaker than the ones reported by Nass, Sproull, or their colleagues. We were puzzled by the discrepancy. We included some of the same measures used in the past work (e.g., the BIDR), and our sample sizes were much larger than in the earlier studies. Several explanations may account for the discrepancy. One could argue that our experimental manipulations were not sufficiently blatant to generate this hypothesized effect. We believe our manipulations to be at least as obvious as many of the social interface research studies which often use very subtle cues, such as a label on a computer monitor [14] or the shape of a mouth on a computer-displayed face [23] (see also [11][19]). Another explanation may relate to the use of college students in the experimental studies. In our study we had sufficient sample size to control for several variables--whether the respondent was currently a student, age, prior survey experience, and level of trust--that we though might interact with the experimental variables and explain why our results differ from those of the earlier studies. For example, we tested the hypotheses that students are more sensitive to the characteristics of the interface and that respondents with prior experience with Web surveys would be less sensitive to them. None of these hypotheses received much support--we did not find any significant interactions between these individual differences variables and the experimental variables on the reporting of sensitive information or gender attitudes.

Another possible explanation, which we could not test, is that the demand characteristics of laboratory-based experimented yield results that are not replicated in distraction-filled field-based surveys. In the experimental work, undergraduate students (often in psychology classes) typically are recruited for an experiment. They are aware of being in an experiment, and may be alert to any cues that might help them figure out the experimental manipulation. In contrast, survey respondents are typically unaware of being in an experiment, and believe the ostensible reason for the survey is to elicit their views on particular topics. These differences may account for the failure of the social interface theory to replicate beyond the laboratory.

Given the influence of the social interface perspective in human-computer interaction (HCI) research and interface design, it is important to understand whether and how the findings from this work translate to the real-world experiences of those who interact with computers. In one such application (a Web survey) we appear to find little support for the social interface hypothesis.

ACKNOWLEDGMENTS

This research is supported by the National Science Foundation, Award # SES-9910882. We thank Robert Tortora and Margrethe Montgomery of the Gallup Organization for their assistance with in planning and carrying out this project. This paper was originally presented at the CHI '01 Conference on Human Factors in Computing Systems in Seattle, Washington in April, 2001.

REFERENCES

1. Aquilino, W., & LoSciuto, L. Effect of Interview Mode on Self-Reported Drug Use. Public Opinion Quarterly, 54 (1990), 362-395.
2. Athey, K.R., Coleman, J.E., Reitman, A.P., & Tang, J. Two Experiments Showing the Effects of the Interviewer's Racial Background on Responses to Questionnaires Concerning Racial Issues. Journal of Applied Psychology, 44 (1960), 562-566.
3. Couper, M.P. Review of Byron Reeves and Clifford Nass, 'The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places'. Journal of Official Statistics, 13 (1998), 441-443.
4. Couper, M.P. Web Surveys: A Review of Issues and Approaches. Public Opinion Quarterly, 64 (2000), forthcoming.
5. Couper, M.P., & Nicholls II, W.L. The History and Development of Computer Assisted Survey Information Collection. In M.P. Couper, et al. (eds.), Computer Assisted Survey Information Collection. New York: Wiley, 1998.
6. Crowne, D., & Marlowe, D. The Approval Motive. New York: John Wiley, 1964.
7. Devine, P.G. Stereotypes and Prejudice: Their Automatic and Controlled Components. Journal of Personality and Social Psychology, 56 (1989), 5-18.
8. Dryer, D.C. Getting Personal with Computers: How to Design Personalities for Agents. Applied Artificial Intelligence, 13 (1999), 273-295.
9. Fogg, B.J., & Nass, C. Silicon Sycophants: The Effects of Computers That Flatter. International Journal of Human-Computer Studies, 46 (1997), 551-561.
10. Kane, E.W., & Macauley, L.J. Interviewer Gender and Gender Attitudes. Public Opinion Quarterly, 57 (1993), 1-28.
11. Kiesler, S. & Sproull, L. 'Social' Human-Computer Interaction. In B. Friedman (ed.) Human Values and the Design of Computer Technology. Stanford, CA: CSLI Press, 1997.

12. Kiesler, S., Sieff, E., & Geary, C. The Illusion of Privacy in Human-Computer Interaction. Carnegie Mellon University: Unpublished paper, 1992.

13. Kim, J., & Moon, J.Y. Designing Towards Emotional Usability in Customer Interfaces– Trustworthiness of Cyber-Banking System Interfaces. Interacting with Computers, 10 (1998), 1-29.

14. Moon, Y. Impression Management in Computer-Based Interviews: The Effects of Input Modality, Output Modality, and Distance. Public Opinion Quarterly, 62 (1998), 610-622.

15. Moon, Y. Intimate Exchanges: Using Computers to Elicit Self-Disclosure from Consumers. Journal of Consumer Research, 26 (2000), 323-339.

16. Nass, C., Fogg, B.J., & Moon, Y. Can Computers be Teammates? International Journal of Human-Computer Studies, 45 (1996), 669-678.

17. Nass, C., Moon, Y., & Carney, P. Are People Polite to Computers? Responses to Computer-Based Interviewing Systems. Journal of Applied Social Psychology, 29 (1999), 1093-1110.

18. Nass, C., Moon, Y., & Green, N. Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers with Voices. Journal of Applied Social Psychology, 27 (1997), 864-876.

19. Parise, S., Kiesler, S., Sproull, L. & Waters, K. Cooperating with Life-Like Interface Agents. Computers in Human Behavior, 15: (1999), 123-142.

20. Paulhus, D.L. Two-Component Models of Socially Desirable Responding. Journal of Personality and Social Psychology, 46 (1984), 598-609.

21. Reeves, B., & Nass, C. The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places. Cambridge: CSLI and Cambridge University Press, 1997.

22. Schuman, H.. & Converse, J. The Effects of Black and White Interviewers on Black Responses in 1968. Public Opinion Quarterly, 35 (1971), 44-68.

23. Sproull, L., Subramani, M., Kiesler, S., Walker, J.H., & Water, K. When the Interface Is a Face. Human-Computer Interaction, 11, (1996), 97-124.

24. Tourangeau, R. & Smith, T. Collecting Sensitive Information with Different Modes of Data Collection. In M.P. Couper, et al. (eds.), Computer Assisted Survey Information Collection. New York: Wiley, 1998.

25. Turner, C.F., Forsyth, B.H., O'Reilly, J.M., Cooley, P.C., Smith, T.K., Rogers, S.M., & Miller, H.G. Automated Self-Interviewing and the Survey Measurement of Sensitive Behaviors." In M.P. Couper, et al., (eds.), Computer Assisted Survey Information Collection. New York: Wiley, 1998.

26. Turner, C.F., Ku, L., Rogers, S.M., Lindberg, L.D., Pleck, J.H., & Sonenstein, F.L. Adolescent Sexual Behavior, Drug Use and Violence: Increased Reporting with Computer Survey Technology. Science, 280 (May, 1998): 867-873.

27. Walker, J., Sproull, L., & Subramani, R. Using a Human Face in an Interface. Proceedings of the Conference on Human Factors in Computers. Boston (1994), ACM Press, pp. 85-91.

Comments on “Social Presence in Web Surveys” by Mick Couper, Roger Tourangeau & Darby Steiger.

Frederick Conrad
Bureau of Labor Statistics

The paper by Mick Couper, Roger Tourangeau, and Darby Steiger presents exactly the kind of foundational work required to understand the consequences of moving to web surveys. Although this move is inevitable – web-based questionnaires will eventually be a common, if not the predominant, method for collecting Federal survey data – the differences between the web and other media are only just beginning to be identified and studied.

The web is actually similar to many other media but not the same as any of them. For example, it is similar to paper. The primary content is (still) presented in documents – hence the page metaphor. But it is different from paper in several ways, the most obvious being that the content in a web document is hyperlinked, making the document interactive. Web pages are similar to software in that users interact with both by clicking and typing; yet they are different in the sense that web pages are essentially static files (embedded JavaScript not withstanding). The web resembles television: the content is sometimes partitioned into channels but the two media are different in the sense that web content is mostly text and graphics but not video and is available on demand while TV content is made available at prescribed times. And so on.

The point is that it is natural to apply what is already known about a similar seeming medium when adapting a task to web administration. Web surveys are a case in point. The 1999 book by Dillman, *Mail and Internet Surveys* is an update of his 1978 book *Mail and Telephone Surveys*, reflecting the fundamental similarities between paper and web-based questionnaires. In contrast, the Couper et al. study explores one of the novel aspects of the medium – for which I applaud the authors – namely, the interactive character of web surveys. In particular, the authors ask whether the interactivity of web surveys produces social presence effects – the tendency for respondents to behave as if the survey instrument was animate or administered by an interviewer. Social presence effects have been demonstrated in other interactive media by researchers such as Reeves and Nass and their colleagues. If web surveys do produce social presence effects then these must be reconciled with the increased sense of privacy that is apparently produced by self-administration – including self-administration of computerized survey instruments such as CASI and ACASI. In general, self-administration leads to what seems to be more honest reporting of sensitive behaviors (e.g. Turner, Forsyth, O’Reilly, Cooley, Smith, Rogers, & Miller, 1998; Tourangeau & Smith,1996; Schaeffer, 2000).

In fact, Couper, et al. report no social presence effects in their two experiments on self-administered web-based surveys. This is potentially good news for researchers hoping to reap the benefits of self-administration on the web. After all, if the celebrated advantages of CASI and ACASI were to disappear when a remote computer is involved, then the move to web surveys would be a giant step backward, at least for collecting sensitive information. The problem is that Couper et al. report *no effect*, not *the disappearance of an effect* under a particular experimental manipulation. So it is hard to interpret what is essentially a null result (albeit with a good sample and adequate power[1]). It could be that a different experimental manipulation would produce evidence of social presence. Alternatively, it could be that the experimental groups actually exhibit evidence of social presence but this is undetectable without a comparison group. Yet another possibility is that Couper et al. found no evidence of social presence in their experiments because such effects are confined to laboratory studies involving special tasks. I will take up each of these possibilities in turn.

Social presence effects may be restricted to situations in which the user (or respondent) is particularly aware of the agent-like character of the computer (or other medium) and it could be that Couper et al. did not sufficiently create this awareness among their respondents. The authors dismiss this kind of explanation because Reeves & Nass (1996) report many effects based on subtle manipulations such as those involving stick figures or gender of computer voices. But I think this kind of explanation may still apply because Reeves & Nass and their colleagues sometimes go to great lengths to make these manipulations effective. In a well known study by Nass, Moon & Carney (1999) (reported in the Reeves & Nass, 1996 book) users rated a computer's performance on a tutoring task more favorably and homogeneously when they registered their ratings using the same computer they were evaluating than when they used another computer or a paper questionnaire – as if they were being polite to the computer while interacting with it. During the tutoring task, the computer presented a series of facts to users; after reading each fact, users rated how much they knew about that fact. Users were led to believe that the more they knew about a fact, the fewer related facts they would be presented (in fact all users were presented the same facts). According to Nass et al. (1999, p. 1098) the goal was to "ensure that subjects felt they were interacting with the computer rather than simply being passive readers."

It could be that what Couper et al. did to increase the sense of interactivity in their experiments, namely to fill the user's name and content of earlier responses into the questions, did not give individual respondents the sense that the computer was designing its interaction specifically for them. Computerized questionnaires are, in fact, highly interactive in the sense that the particular set of questions asked of any one respondent depends on previous answers and may be unique. Perhaps if this tailoring of question choice and sequence were made more explicit it would lead to more evidence of social presence.

Although Couper et al. did not detect differences in socially desirable responding among the various groups in their experiments, this does not necessarily mean there were no such effects. It just means that any effects were the same for all four groups. The fact that all of the groups completed the questionnaire on a remote computer could increase

[1] In fact, the scientifically constructed samples in the Couper et al. paper are a major advance over the convenience samples used in laboratory studies which report social presence effects, such as those of Reeves & Nass, 1996 and others.

socially desirable responding for all of them. Moon (1998) found more impression management and socially desirable responding (using the same scales[2] as Couper at al.) when the host computer appeared to very remote (3000 miles away) than when it appeared to be nearby (a few miles away) or a standalone machine. Respondents in the Couper at al. study surely perceived the server to be remote which may have led to greater impression management and socially desirable responding in all versions of the questionnaire than would have been observed in corresponding instruments administered on standalone computers. Thus an additional control condition in which it is clear to respondents that there is no network connectivity could help tease this apart. If the "unwired" groups scored lower on Impression management and socially desirable responding than their "wired' counterparts, this explanation would seem to hold.

Of course, it is possible that Couper et al. detected no evidence of social presence because there really is none. By this view, the kind of effects that fill the Reeves & Nass (1996) book are confined to laboratory studies in which convenience samples of subjects carry out special set-up tasks unlikely to occur under ordinary conditions of survey administration. I think this is partly right but that there is something to the Reeves & Nass (1996) kind of finding. One way to reconcile social presence effects in the laboratory with their absence in the current research is to acknowledge that people are ordinarily sensitive to the agent-like character of computers and in some ways treat them like people (e.g. pleading with computers not to crash before a save command is completed). But people know the difference between computers and people and under circumstances where this difference matters – such as reporting sensitive material – the inanimacy and privacy of the medium outweighs its social character. When web survey respondents report about sensitive topics they suspend the perception of the computer as a social agent.

Clearly this work has important implications for collecting information on sensitive topics but does that limit its usefulness for Federal statistical agencies? My sense was that Federal surveys overwhelmingly concern mundane facts and behaviors about which respondents are unlikely to be sensitive. However, if one steps through the "A to Z" topic index on the FedStats web site, many of these topics are potentially sensitive (see Table 1). Furthermore, self-presentation concerns may be more relevant for reporting mundane behaviors than is typically assumed. In a recent study we (Conrad & Schober, 1999) report that for mundane concepts like "more than one job," "household furniture," and "live in this house," respondents were more likely to request clarification from a computer than from an interviewer. In part this may be because it is easier to click a mouse on highlighted text than to formulate and utter a question. But it may also reflect less shame in indicating confusion about everyday concepts to a computer than to an

[2] Self-reports of impression management and socially desirable responding are tricky. If someone tells you they are unlikely to be honest under the very circumstances in which they are telling this to you, it's hard to know if they are currently being honest. For that matter, if they tell you they are likely to be honest under particular circumstances, they could be indicating this for reasons of impression management or social desirability. For current purposes, I accept the validity of these measures, but believe their use deserves additional scrutiny in the future.

interviewer. This was a small scale laboratory study and it's not clear whether the results will scale up to large web-based surveys. But the lack of social presence effects in the Couper at al. research bodes well for the benefit of computerized self-administration of both mundane and sensitive questions when asked in web surveys.

Table 1. Potentially sensitive topics and Federal agencies that collect information on those topics.

Topic	*Agency*
Abortion	NCHS
AIDS and STDs	NCHS
Crime victimization	BJS, OJJDP
Criminal offenses	BJS, OJJDP
Divorce	NCHS
Drug abuse	SAMHSA
Educational Assessment	NCES
Immigration status	INS
Income	BLS, Census
Job loss	BLS

References

Conrad, F.G. & Schober, M.F. (1999). A conversational approach to text-based computer-administered questionnaires. In *Proceedings of the Third International ASC Conference*. Chesham, UK: Association for Survey Computing, pp. 91-101.

Moon, Y. (1998). Impression management in computer-based interviews: The effects of input modality, output modality and distance. *Public Opinion Quarterly*, **62**, 610-622.

Nass, B. & Reeves, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Palo Alto, CA and Cambridge, UK: CSLI Publications and Cambridge University Press.

Reeves, C., Moon, Y. & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology*, **29**, 1093-1110.

Schaeffer, N.C. (2000). Asking questions about threatening topics: A selective overview. In A. A. Stone, J. S. Turkkan, A. A. Bachrach, J. J. Jobe, H. S. Kurtzman & V. S. Cain (Eds.), *The Science of Self –Report*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, pp. 105-121.

Tourangeau, R. & Smith, T.W. (1996) Asking sensitive questions: The impact of data collection mode, question format and question context. *Public Opinion Quarterly*, **60**, 274-304.

Turner, C. F., Forsyth, B. H., O'Reilly, J.M., Cooley, P.C., Smith, T.K., Rogers, S.M., & Miller, H.G. (1998). Automated self interviewing and the survey measurement of sensitive behaviors. In M. Couper, R. Baker, J. Bethlehem, C. Clark, J. Martin, W. Nicholls, & J. O'Reilly (Eds.)*, Computer Assisted Survey Information Collection*. New York: John Wiley & Sons, pp. 455-473.

Discussion of paper: Social Presence in Web Surveys
by Roger Tourangeau, Mick P. Cooper, and Darby M. Steiger

David Mingay, Statistical Research Division, U.S. Census Bureau,
Washington, DC 20233-9100. david.j.mingay@census.gov

BACKGROUND

The research presented in this paper is just a small part of an important and innovative series of experiments that Tourangeau and Couper are conducting under their NSF grant. These projects should significantly expand our knowledge of how interface design affects survey responding. Knowing which design features have positive and negative effects on data quality should help Web designers to design survey instruments that collect high quality data and help those using the data to assess its quality. The paper follows the Cognitive Aspects of Survey Methods (CASM) tradition of taking a theory that was developed in a different domain and testing predictions derived from it when applied to a survey setting.

Web technology provides a substantial opportunity to add a variety of visual features to the survey instrument, including features that make the survey feel more personal. We know little about which features are helpful, harmful, or neutral in their effect on data quality. On the one hand, adding certain visual features to the Web questionnaire, including personalization features, may be helpful. Respondents may perceive the survey as being more interesting, make greater effort to answer the questions accurately, and be less likely to break off the interview. Those who do not complete the interview in a single setting may be more likely to complete it later.

But there can be disadvantages with adding visual features. For example, visual images take time to download, and the added time may frustrate the respondent—particularly if they have a 14K, 28K, or even 56K modem. Interactive features may require the respondent to have Flash or other software installed on their computer, which many people do not have. And adding features often introduces visual clutter and may distract the respondent from answering the questions in a considered manner. Most importantly, some features can affect the interpretation of questions or cause other response problems (Couper, in press).

A recent study of Web TV panel members conducted by the authors illustrates the last point (Kenyon, Couper, and Tourangeau, 2001). They looked at a feature that is being used in some commercial Internet surveys, namely pictures that illustrate the topic of the questions. Questions about the frequency with which panel members engaged in six activities (e.g., overnight trips in the last year) were accompanied by either no picture, a picture showing an example of one way of doing the activity (e.g., a businessman at an airport), a picture showing an example of a different way (e.g., a family in a station wagon), or both pictures. Nineteen of the 54 two-way comparisons showed statistically significant differences ($p<.10$). Most notably, on all six comparisons between the two pictures illustrating an activity, one picture resulted in a significantly higher reported frequency for the activity than the other picture. In addition, asking the same question with no picture versus one picture produced statistically significant mean scores in six of the twelve comparisons. These results suggest that the visual material accompanying a question can elicit different responses to that question.

In the paper presented in this volume, Tourangeau and Couper looked at two other types of features that are commonly found on questionnaires administered on the World Wide Web to see if they resulted in a more socially desirable response to questions. The features selected allowed the testing of predictions derived from social interface theory. This theory suggests that when the interaction with the computer has some of the features of an interaction with a human, responses similar to those elicited in a human-to-human interaction are obtained (Kiesler and Sproull, 1977). In contrast, the survey research literature suggests that computer-administration of surveys on highly sensitive topics reduces or eliminates the social desirability effects found with human administration of questions even when such humanizing features as a voice are used. This study explored the apparent contradiction between the social interface and survey methods findings.

Two characteristics of interactions were manipulated, personalization and interaction. The level of personalization was varied by presenting a study logo or an image of a male or female researcher along with text of relevant statements from the researcher. The level of interactivity was varied by changing the language used. For example, the high interaction version used the first person in introductions and transitional phrases and occasionally echoed back one of the respondent's earlier answers. Two Web surveys with identical designs were conducted. One survey was completed by 202 members of a Gallup Organization Web panel and the other by 3047 members of a sample of Web users purchased from Survey Sampling, Inc. (SSI).

RESULTS

Noting that the smaller study had very similar findings, the authors only report the results of the larger study. Only one finding that offers some limited support for social interface theory is reported. Respondents of both sexes showed the most pro-feminist attitudes when the program displayed pictures and messages from the female investigator and the least pro-feminist attitudes when the pictures and messages were from the male investigator, with the attitudes reported by the group getting the survey logo falling between the two. Very few of the numerous other predicted interactions were observed (Tables 2 and 3 shows 16 of the results) and one—the effect of personalization on reports of having voted in the last election—was in the opposite direction to that predicted by social interface theory. Thus, neither level of personalization nor level of interaction demonstrably affected reports about sensitive topics. This is consistent with the survey literature, but not with social interface theory.

DISCUSSION

Why were the results predicted by social interface theory not found in the two surveys? The authors appear to favor the idea that the sterile, controlled environment of the laboratory in social interface studies focuses subjects' attention on the experimental manipulation. In contrast, their two Web surveys involved a much more uncontrolled real world setting with many potential distracters.

I would suggest that respondents' lack of attention to the experimental manipulation might also be due to low involvement in the surveys. Tentative evidence for low involvement is provided by the fairly poor response rate: only 20% of people in the SSI panel who received an email invitation to participate and one email reminder did so. In addition, 12.5% of those starting the interview failed to complete it.

Uncontrolled real world settings with many potential distracters are intrinsic to Web surveys (Couper, in press) and respondent uninvolvement seems likely to be true for most, but not all, Web surveys. This suggests that these results may apply to other Web surveys as well. Several staff at commercial Web survey organizations have told me that their panel members often sign up to be on several other panels and receive numerous invitations to take surveys. They often chose to answer a questionnaire to be entered in a drawing for a prize or to get a small fee and want to finish as quickly as possible. In the terms of the Krosnick and Alwin (1980) model of survey responding, most respondents probably show considerable satisficing behavior. In addition, the experience of taking numerous prior surveys may affect how panel members respond to survey questions—perhaps making them less attentive to subtle cues, for example.

Of course, interactivity and personalization may commonly affect reporting, but the design of this study prevented the effects from being observed. The authors point out that with 3047 people completing the questionnaire in the second study, there was ample statistical power. However, the use of the first person and transitional phrases seems a relatively weak way to increase interactivity (although many of the manipulations used successfully in social interface research appear even more subtle). Other procedures might have resulted in more socially desirable responding in the high interactivity and/or personalization conditions.

I have some concern about using the method of echoing back certain answers (e.g., "According to your responses, you exercise once daily") to increase the level of interactivity. While that does increase the interactivity of the interview, it also conveys other information to the respondent--for example, that the answers are particularly important. Respondents may then feel more pressure to answer accurately and may show fewer tendencies to bias their answers in a socially desirable manner. Thus it is possible that providing feedback as one of the interactive features may have somewhat reduced the effectiveness of the high interactive version of the interview for increasing social desirability biases.

There is also some evidence that the results of this research may not be entirely reliable or generalizable. In a recently completed third study that was mentioned in the talk, high interactivity and personalization was associated with more socially desirable reporting in a number of instances.

However, the findings are consistent with other evidence that computer-administration reduces social desirability effects even when there is quite a strong degree of personalization. For example, methodological research by RTI using telephone audio computer-assisted self-interviewing (T-ACASI) and interactive voice response (IVR) administration methods for federal drug surveys found high levels of reporting of sensitive information despite the significant level of personalization associated with using a voice to administer the questions (Turner et al., 1998).

CONCLUSIONS

This well-designed study addresses the important issue of whether greater personalization and interactivity in a Web survey increases social desirability effects in reporting. There was little evidence for this although an additional study that was mentioned in the talk provides somewhat contradictory findings. Several other studies being conducted as part of this NSF-funded project address related issues. Taken together, these studies should provide considerable insight into whether, as is postulated by social interface theory, humanizing cues in a computer interface can evoke reactions similar to those produced by a human, such as social desirability response effects, and, perhaps, under what conditions the effects are produced.

A particularly noteworthy feature of the research is that it investigated two important dimensions of surveys that have been neglected by researchers. Web survey designers often use personalization, interactivity, and other features in their surveys, believing that they may increase the respondent's involvement in the questionnaire, and thus improve reporting accuracy and reduce breakoffs. Research of this type is very important for determining whether introducing such features improves, harms, or does not affect the quality of the survey data.

REFERENCES

Couper, M. (in press). The promises and perils of Web surveys. In A. Westlake (ed.). The challenge of the Internet. Association for Survey Computing.

Kenyon, K., Couper, M., and Tourangeau, R. (2001). *Experiments on visual effects in Web surveys*. Paper presented at the meeting of the American Association for Public Opinion Research, May 17-20, Montreal, Canada.

Kiesler, S., and Sproull, L. (1997). 'Social' human-computer interaction. In B. Friedman (ed.). *Human values and the design of computer technology*. Stanford, CA: CSLI Press.

Krosnick, J., and Alwin, D.F. (1987). *An evaluation of a cognitive theory of response order effects in attitude measurement*. Public Opinion Quarterly, 51, 201-219.

Turner, C.F., B.H. Forsyth, J. O'Reilly, et al. (1998). Automated self-interviewing and the survey measurement of sensitive behaviors. In M. Couper (ed.). *Computer-Assisted Survey Interviewing Collection.* New York: Wiley.

Session 5

Application of Jackknife Theory in Small Area Estimation

Jackknifing in The Fay-Herriot Model with An Example

Jiming Jiang, P. Lahiri, Shu-Mei Wan, Chien-Hua Wu

Abstract

The paper reviews empirical best linear prediction (EBLUP) and the associated jackknife MSE estimator of EBLUP. The bias of jackknife MSE estimator is of order $o(m^{-1})$, where m is the number of small areas. The jackknife works well both for normal and nonnormal Fay-Herriot models. The proposed methodology is illustrated using a real life example from the National Health and Interview Survey.

1 Introduction

Fay and Herriot (1979) put forward an empirical Bayes method to estimate per-cpita income of small-places (population less than 1000) using a Bayesian model that combines Current Population Survey data in conjunction with relevant administrative and census data. Their empirical results demonstrate that their empirical Bayes estimator performed better that both direct survey estimator and a synthetic estimator which is a direct estimator for the corresponding county. The Fay-Herriot method is a popular small-area method because of its simplicity and its demonstrated good empirical performances. It also produces design consistent estimator, a desirable property which brings a model-based estimator closer to direct estimator for large sample, irrespective of the true model.

Prasad and Rao (1990) developed a delta method for estimating mean square error (MSE)

[0] J. Jiang, Assoc. Prof., Univ. of Calif.-Davis, U.S.A; P. Lahiri, Milton Mohr Distinguished Prof., Univ. of Nebraska-Lincoln, U.S.A.; S. Wan, Assist. Prof., Lung-Hua Inst. Tech., Taiwan; C. Wu, Statistical Reviewer, Center for Drug Evaluation, Taiwan.

of empirical best linear unbiased predictor (EBLUP) of a general mixed effect in the context of a mixed *linear normal* model which covers the Fay-Herriot model. Lahiri and Rao (1995) robustified the Prasad-Rao method by allowing *nonnormal* random effects in the Fay-Herriot model. However, both the papers are vaild only for ANOVA method of estimating the model parameters. Datta and Lahiri (2000) considered the mixed linear normal model considered by Prasad and Rao (1990) but generalized the Prasad-Rao's method to include ML and REML variance component estimators. More recently, Jiang *et al.* (2001) proposed a jackknife method to estimate the MSE of empirical best predictor (EBP) for *nonnormal* and *nonlinear mixed* models and for general M-estimators of model parameters. Their MSE estimator enjoys the desirable property that the bias is of order $o(m^{-1})$.

The main purpose of this paper is to spell out the jackknife method for the Fay-Herriot model. For a very special case of the Fay-Herriot model, Lahiri (1995) noted that the jackknife MSE estimator of an EBLUP involves estimated skewness and kurtosis terms. The jackknife MSE estimtor is also asymptotically equivalent to Morris' (see Morris 1983) formula which was obtained as an approximation to the posterior variance under a uniform improper prior distribution on the model parameters. Thus, the jackknife is very similar to a Bayesian procedure, at least for this special case.

As for an illustration of our methodology, we carry out a data analysis to estimate the proportion of people who did not visit a doctor's office during the last twelve months for each state and the District of Columbia (small areas). The Fay-Herriot model cannot be applied directly to the survey estimates since one would expect its true sampling variances to be related to the corresponding true small-area proportions. Note that the Fay-Herriot model assumes that the sampling variances to be known. Thus, we first make a suitable transformation of the direct survey estimates to stabilize their sampling variances and then assumed known design effects. We then simply apply the Fay-Herriot model on the transformed survey estimates in order to combine information from various relevant census and administrative data. The performances of the Fay-Herriot type of estimator and the associated jackknife MSE estimator seem reasonable.

2 Estimation in a non-normal Fay-Herriot Model

We assume a non-normal version of the small area model considered by Fay and Herriot (1979). According to the model, $y_i = \theta_i + e_i; \theta_i = x_i'\beta + v_i$, where e_i and v_i are all uncorrelated with $E(e_i) = E(v_i) = 0$ and $Var(e_i) = D_i,\ Var(v_i) = A\ (i = 1, \cdots, m)$. In the model, D_i's $(i = 1, \cdots, m)$ are assumed to be known. Let p be the dimension of x_i and $\phi = (\beta, A)$.

When ϕ is known the best predictor (BP) of θ_i is simply the conditional mean of θ_i given y_i and is given by $\hat{\theta}_i(y_i;\phi) = (1-\gamma_i)x_i'\beta + \gamma_i y_i$, where $\gamma_i = A/(A + D_i)$. Note that the above BP can be also interpreted as a Bayes estimator When β is unknown but A is known, one can estimate β by the generalized least square estimator of β, given by $\hat{\beta}(A) = \left(\sum_{i=1}^m (A+D_i)^{-1} x_i x_i'\right)^{-1} \sum_{i=1}^m (A+D_i)^{-1} x_i y_i$.

An EBP [or empirical Bayes (EB)] of θ_i is then obtained by replacing β in the BP by $\hat{\beta}(A)$. Note that this is also the best linear unbiased predictor (BLUP), see Prasad and Rao (1990).

In practice A is rarely known and so it needs to be estimated from the data. Prasad and Rao (1990) used a method of moments (MOM) estimator of A, defined by $\hat{A} = max[0, \tilde{A}]$ with $\tilde{A} = (m-p)^{-1} \sum_{i=1}^m \{(y_i - x_i'\hat{\beta}_{OLS})^2 - (1-h_i)D_i\}$, where $h_i = x_i'(\sum_{i=1}^m x_i x_i')^{-1} x_i$, $\hat{\beta}_{OLS} =$

$(\sum_{i=1}^{m} x_i x_i')^{-1} \sum_{i=1}^{m} x_i y_i$.

Researchers have used other methods of estimating A (see, *e.g.* Fay and Herriot 1979, Jiang *et al.* 2001, among others). Plugging in an estimator of A in the BLUP yields EBLUP $\hat{\theta}_i = \hat{\theta}_i(y_i; \hat{\phi}) = (1 - \hat{\gamma}_i)x_i'\hat{\beta}(\hat{A}) + \hat{\gamma}_i y_i$, where $\hat{\gamma}_i = \hat{A}/(\hat{A} + D_i)$ and

$$\hat{\beta}(\hat{A}) = (\sum_{i=1}^{m} (\hat{A} + D_i)^{-1} x_i x_i')^{-1} \sum_{i=1}^{m} (\hat{A} + D_i)^{-1} x_i y_i.$$

Note that it can be also interpreted as an EB estimator.

In order to understand if EBLUP method is effective, we now develop a method of constructing confidence intervals of γ_i based on the point estimates $\hat{\gamma}_i$ $(i = 1, \cdots, m)$. Applying Taylor series method, we obtain an estimator of $Var(\hat{\gamma}_i)$ as $v_J(\hat{\gamma}_i) = \frac{D_i^2}{(\hat{A}+D_i)^4} v_J(\hat{A})$, where $v_J(\hat{A}) = \frac{m-1}{m} \sum_{u=1}^{m} (\hat{A}_{-u} - \hat{A})^2$ is a jackknife estimator of $Var(\hat{A})$. Here $\hat{A}_{-u}$ is calculated exactly in the same way as A except that the data for the uth small area is deleted $(u = 1, \cdots, m)$ We can construct the confidence intervals of γ_i as $\{\hat{\gamma}_i - 2\sqrt{v_J(\hat{\gamma}_i)}, \hat{\gamma}_i + 2\sqrt{v_J(\hat{\gamma}_i)}\}$, $i = 1, \cdots, m$. If EBLUP is effective, the confidence intervals of γ_i for most of the states will not contain 1 or 0.

Next, we discuss the important assumption of known sampling variances D_i's. In Fay and Herriot (1979), the following justification was given. They assumed that the coefficient of variation for the i th small area direct estimate is approximately $3/\sqrt{N_i}$, where N_i denotes 20 percent sample count. This approximation was made based on an empirical study which found that above approximation works well for eight states. A log transformation was then taken to stabilize the variance and the Fay-Herriot model was applied on the transformed direct estimates with $D_i = 9/N_i$.

In view of the above discussion, we are often encountered with the problem of estimation of $h(\theta_i)$, a function, possibly nonlinear, in θ_i. A simple estimator (*e.g.,* Fay and Herriot 1979) is $h^{-1}(\hat{\theta}_i)$, where $h^{-1}(.)$ is the inverse transformation of $h(.)$. This is not an EBP but should work fine as long as sample size for the small areas are not very small. We, however, note that it is possible to come up to the BP of $h(\theta_i)$ and hence EBP (see, Lahiri 1999).

3 Jackknifing in the Fay-Herriot Model

In this section, we spell out the jackknife MSE estimator of the Fay-Herriot type estimator based on Jiang *et al* (2001). The MSE of $\hat{\theta}_i$ is defined as $MSE(\hat{\theta}_i) = E(\hat{\theta}_i - \theta_i)^2$, where E is with respect to the Fay-Herriot mixed model. Note that the MSE of the BP is given by $g_{1i}(A) = A(1-\gamma_i)$. One can then naively propose a MSE estimator of EBLUP as $g_{1i}(\hat{A})$. The problem with this naive estimator is that it does not incorporate the extra variabilities due to the estimation of ϕ and so underestimates the true MSE. Several researchers addressed this important issue and came up with improved MSE estimators which account for these extra variabilities (see, *e.g.,* Prasad and Rao 1990, Lahiri and Rao 1995, Datta and Lahiri 2000, among others). But they are all valid for mixed linear normal model.

Recently, Jiang *et al.* (2001) proposed a jackknife method which takes into account uncertainties due to the estimation of ϕ. This method is valid for nonnormal and nonlinear mixed models and for general M-estimators of model parameters. For the Fay-Herriot model, the jackknife MSE estimator of EBLUP is given by

$$mse_J(\hat{\theta}_i) = g_{1i}(\hat{A}) - \frac{m-1}{m}\sum_{u=1}^{m}[g_{1i}(\hat{A}_{-u}) - g_{1i}(\hat{A})] + \frac{m-1}{m}\sum_{u=1}^{m}(\hat{\theta}_{i,-u} - \hat{\theta}_i)^2,$$

where $\hat{A}_{-u}(\hat{\beta}_{-u})$ is the estimator of A (β) after deleting the u^{th} small-area data,

$$\hat{\theta}_{i,-u} \quad = \quad \hat{\gamma}_{i,-u}y_i + (1-\hat{\gamma}_{i,-u})x_i'\hat{\beta}_{-u},$$

$$\begin{aligned}\hat{\gamma}_{i,-u} &= \frac{\hat{A}_{-u}}{\hat{A}_{-u}+D_i},\\ g_{1i}(\hat{A}_{-u}) &= \hat{A}_{-u}(1-\hat{\gamma}_{i,-u}).\end{aligned}$$

Lahiri (1995) examined the above jackknife MSE estimator for a very special case of the Fay-Herriot model with $D_i = D$ and $x_i'\beta = \mu$ $(i = 1, \cdots, m)$. He showed that

$$\begin{aligned}\widehat{MSE}(\hat{\theta}_i) &= g_1(\hat{A}) + g_2(\hat{A}) + \frac{D^2}{m(\hat{A}+D)}(b_2-1)\\ &\quad + \frac{D^2}{m(\hat{A}+D)^2}(b_2-1)(y_i-\bar{y})^2 - \frac{2D^2}{m(\hat{A}+D)^{3/2}}\sqrt{b_1}(y_i-\bar{y}),\end{aligned}$$

where $b_1 = m_3^2/(\hat{A}+D)^3$, $b_2 = m_4/(\hat{A}+D)^2$ and $g_2(A) = \frac{D_i^2}{(A+D_i)^2}x_i'(\sum_{u=1}^m \frac{x_u x_u'}{A+D_u})^{-1}x_i$. Thus, unlike the normality-based MSE estimators, the jackknife MSE estimator involves estimated skewness and kurtosis terms. Lahiri (1995) also compared jackknife with the two normality-based MSE estimators of the following EBP of θ_i. $\hat{\theta}_i = \bar{y} + (1-\hat{B}_1)(y_i-\bar{y})$, where $\bar{y} = m^{-1}\sum_{i=1}^m y_i$ and $\hat{B} = \frac{D(m-3)}{\sum_{i=1}^m (y_i-\bar{y})^2}$.

We present three formulae below for comparison:

Morris (1983):

$$(1-\hat{B}_1) + \frac{D\hat{B}_1}{m} + \frac{2\hat{B_1}^2}{m-3}(y_i-\bar{y})^2,$$

Prasad and Rao (1990):

$$(1-\hat{B}_1)D + \frac{D\hat{B}_1}{m} + \frac{2D\hat{B}_1}{m},$$

The jackknife Formula:

$$(1-\hat{B}_1)D + \frac{D\hat{B}_1}{m} + \frac{2\hat{B}_1^2}{m}(y_i-\bar{y})^2.$$

The jackknife estimator is equivalent to the Morris' formula which was obtained as an approximation to the posterior variance formula under uniform improper prior on the model parameters: μ and A. Also, the bias of our jackknife MSE estimator is of the order $o(m^{-1})$. Thus, our jackknife MSE estimator enjoys both good frequentist and Bayesian properties. It is interesting to note that the Prasad-Rao MSE estimator, unlike Morris' and ours, is the same for all the areas in this balanced case.

The above results are for the transformed scale. We need to provide results in the original scale. We approximate the MSE of $h^{-1}(\hat{\theta}_i)$ by $mse[h^{-1}(\hat{\theta}_i)] = [h^{-1\prime}(\hat{\theta}_i)]mse(\hat{\theta}_i)$, where $h^{-1\prime}(x)$ denotes the derivative of $h^{-1}(x)$ with respect to x and $mse(\hat{\theta}_i)$ is an estimate of MSE obtained by jackknife method as described above.

4 Data Analysis

In this section, we demonstrate our methodology to estimate the proportion of individuals who did not visit doctor's office during the last twelve months for all the fifty states and the District of Columbia (small areas) using the National Health Interview Survey (NHIS) data in conjunction with relevant administrative and census data. Earlier Malec *et al.* (1997) proposed a hierarchical Bayes method to address the same estimation problem. Unlike our modeling, they used an individual level model and their method does not produce design consistent small area estimators. Also, they did not use auxiliary data at the small area level. Our method can be viewed as a first step in getting a simple minded design consistent small area estimators. It has also a huge computational advantage over their procedure.

Let n_i be the sample size for the ith state and w_{ij} be the sampling weight for the jth individual in the ith state $(i = 1, \cdots, m = 51; j = 1, \cdots, n_i)$. For the jth individual in the ith state, we observe a binary response z_{ij} which takes on the value 1 if the individual did not visit a doctor's office during the last 12 months and 0 otherwise ($i = 1, \cdots, m; j = 1, \cdots, n_i$). Then, $z_i = \sum_{j=1}^{n_i} w_{ij} z_{ij} / \sum_{j=1}^{n_i} w_{ij}$ is the direct survey estimate of π_i, the true proportion of individuals who did not visit a doctor's office during the last 12 months for the ith state ($i = 1, \cdots, m$). Using SUDDAN, sample survey software, the NCHS has provided data on z_i and its sampling variance V_i for $i = 1, \cdots, m$.

Note that $E(z_i|\pi_i) \approx \pi_i$, and

$$V(z_i|\pi_i) = D_{i0}^{\star} \cdot \frac{\pi_i(1-\pi_i)}{n_i}$$

where

$$D_{i0}^{\star} = \frac{V_i^*}{\frac{\pi_i(1-\pi_i)}{n_i}} \approx \frac{V_i}{\frac{z_i(1-z_i)}{n_i}} D_{i0}.$$

The factor $D_{i0}^{\star}$ is known as a design effect and adjusts the simple random sampling formula by incorporating effects due to clustering and unequal probability selections. Consider the transformation: $y_i = \sin^{-1}\sqrt{z_i}$. Thus, for this example, $h(.) = \sin^{-1}(.)$. By Taylor series argument, we have

$$E(y_i|\theta_i) \approx \sin^{-1}\sqrt{\pi_i} = \theta_i,$$

and

$$\begin{aligned} V(y_i|\theta_i) &= V[\sin^{-1}\sqrt{z_i}|\theta_i] \\ &\approx V[\sin^{-1}\sqrt{\pi_i} + (z_i - \pi_i) \cdot \frac{1}{2\sqrt{\pi_i(1-\pi_i)}}] \\ &= \frac{1}{4\pi_i(1-\pi_i)} \cdot D_{i0}^{\star} \cdot \frac{\pi_i(1-\pi_i)}{n_i} \\ &= \frac{D_{i0}^{\star}}{4n_i}. \end{aligned}$$

We will assume that $D_i = \frac{D_{i0}}{4n_i}$ is the estimated sampling variance of y_i $(i = 1, \cdots, m)$.

In addition to z_i and V_i provided to us by the NCHS, we collected data on 1990 urban population (X_1), 1995 Bachelor's degree completion for 25+ population (X_2), 1995 high school completion for the 25+ population (X_3), 1995 health insurance coverage (X_4), and 1990 physician population (X_5) for each of the 50 states and the District of Columbia. The covariate information was obtained from the Census Bureau web site.

We first consider the issue of covariate selection. For this purpose, we considered the largest 15 states (in terms of sample size) and used SAS to produce Tables 1 and 2. For these states sampling variabilities are very low and so the usual SAS procedures are justified. Note that correlations between Y and each of the three covariates X_2, X_3, and X_4 are significant (at 0.1 level). While X_2 is significantly correlated with X_3, it is not significantly correlated with X_4. Likewise, X_4 is significantly correlated with X_3, but not with X_2. Thus, keeping the aspect of multicollinearity in mind, Table 1 suggests to consider X_2 and X_4 in the model. The selection of these two covariates is confirmed by Table 2. Both R^2 and adjusted R^2 are the highest when we include X_2 and X_4 in the model. We would also select these two covariates when we apply the C_p criterion.

The estimates of γ ranges between .09 (South Dakota) and .95 (California), depending on the sampling variability of the corresponding state NCHS estimate. The γ values are low for small states (indicating that the NCHS estimates are highly unreliable) and large for large states (indicating the NCHS estimates are reliable). None of the confidence intervals [(L.L.,U.L)] includes 0 or 1 suggesting the use of EBLUP. See Table 3 and Figure 3.

The estimator of π_i is then given by $\hat{\pi}_i = \sin^2(\hat{\theta}_i)$. Thus, here $h^{-1}(.) = \sin^2(.)$. The MSE of $\hat{\pi}_i$ is given by $4\hat{\pi}_i(1-\hat{\pi}_i)mse(\hat{\theta}_i)$. A synthetic estimator of π_i is given by $\hat{\pi}_i^{syn} = \sin^2(x_i'\hat{\beta})$.

Table 4 reports the NCHS estimates (z), proposed composite estimates ($\hat{\pi}$), and synthetic estimates ($\hat{\pi}^{syn}$). For large states (e.g., California , Texas, etc.), our proposed composite estimates are similar to the NCHS estimates. Figure 1 plots these estimates.

Finally, Tables 5 provides standard errors of the NCHS estimates ($se(z) = \sqrt{V_i}$), the jackknife MSE estimates of our proposed composite estimates ($se(\hat{\pi}) = \sqrt{mse_J(\hat{\pi})}$), and percent improvement defined by $PCTIMP = 100 \times \frac{se(z)-se(\hat{\pi})}{se(z)}$. A corresponding plot of $PCTIMP$ is given in Figure 3. For small states (e.g., South Dakota, Vermont, etc.), improvement is quite substantial.

Acknowledgements

The research was supported, in part, by the National Science Foundation Grants and a contract from the National Center for Health Statistics.

REFERENCES

Chattopadhyay, M., Lahiri, P., Larsen, M., and Reimnitz, J. (1999), Composite estimation of drug prevalences for sub-state areas, *Survey Methodology*, 25, 81-86.

Datta, G.S. and Lahiri, P. (2000), A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems *Statist. Sinica.*

Jiang, J., Lahiri, P., and Wan, S. (2001), A unified jackknife theory, under revision for *Ann. Statist.*

Fay, R. E., and Herriot, R. A. (1979), Estimates of income for small places: an application of James-Stein procedures to census data, *J. Amer. Statist. Assoc.* 74, 269-277.

Lahiri, P. (1995), A jackknife measure of uncertainty of linear empirical Bayes estimators, unpublished manuscript.

Lahiri, P. (1999), Discussion on *Current Trends in Sample Survey* by J.N.K. Rao, *Sankhya.*

Lahiri, P., and Rao, J.N.K. (1995), Robust estimation of mean squared error of small area estimators, *J. Amer. Statist. Asso.*, 90, 758-766.

Malec, D., Sedransk, J., Moriarity, C.L., and LeClere, F.B. (1997), Small area inference for binary variables in the National Health Interview Survey, *J. Amer. Statist. Asso.* 92, 815-826.

Morris, C. (1983), Parametric empirical Bayes inference: theory and application, *J. Amer. Statist. Asso.,* 78, 47-59.

Prasad, N.G.N., and Rao, J.N.K. (1990), The estimation of mean squared errors of small area estimators, *J. Amer. Statist. Asso.*, 85, 163-171.

Table 1: Pearson Correlation Coefficients

	Y	X_1	X_2	X_3	X_4	X_5
Y	1.00000	0.10733	0.59754	0.55709	0.51929	0.26759
	0.0	0.7034	0.0187	0.0310	0.0473	0.3349
X_1	0.10733	1.00000	0.56726	0.44259	-0.25401	-0.64715
	0.7034	0.0	0.0274	0.0985	0.3610	0.0091
X_2	0.59754	0.56726	1.00000	0.45595	0.02605	0.05537
	0.0187	0.0274	0.0	0.0876	0.9266	0.8446
X_3	0.55709	0.44259	0.45595	1.00000	0.62160	-0.09249
	0.0310	0.0985	0.0876	0.0	0.0134	0.7430
X_4	0.51929	-0.25401	0.02605	0.62160	1.00000	0.26579
	0.0473	0.3610	0.9266	0.0134	0.0	0.3383
X_5	0.26759	-0.64715	0.05537	-0.09249	0.26579	1.00000
	0.3349	0.0091	0.8446	0.7430	0.3383	0.0

Table 2: Values of C_p statistic, R^2, and Adjusted R^2 for different possible models

Model	Variables	C_p	R^2	Adjusted R^2
1	X_2	7.23672	0.35705587	0.3076
2	X_3	8.56150	0.31035017	0.2573
3	X_4	9.71567	0.26965933	0.2135
4	X_2, X_3	6.33543	0.45934214	0.3692
5	X_2, X_4	2.03477	0.61096386	0.5461
6	X_3, X_4	9.17802	0.35912540	0.2523
7	X_2, X_3, X_4	4.00000	0.61218982	0.5064

Table 3: Direct survey estimates (y), synthetic estimates ($x'\hat{\beta}$), EBLUP's ($\hat{\theta}$), and confidence Intervals of γ's

	STATE	y	$x'\hat{\beta}$	$\hat{\gamma}$	$\hat{\theta}$	L.L.	U.L.
1	Alabama	1.08078	1.04874	0.85047	1.07599	0.72554	0.97539
2	Alaska	0.99007	1.08821	0.31453	1.05734	0.10274	0.52632
3	Arizona	1.05567	1.02178	0.73862	1.04681	0.54898	0.92827
4	Arkansas	1.05055	1.01292	0.72774	1.04031	0.53310	0.92237
5	California	1.04064	1.04298	0.95314	1.04075	0.90926	0.99701
6	Colorado	1.09146	1.11180	0.63604	1.09886	0.40864	0.86344
7	Connecticut	1.16463	1.13980	0.67252	1.15649	0.45618	0.88887
8	Delaware	1.12824	1.06208	0.29641	1.08169	0.09154	0.50127
9	D.C.	1.19544	1.12036	0.28924	1.14207	0.08729	0.49118
10	Florida	1.00245	1.04526	0.82933	1.00976	0.69029	0.96837
11	Georgia	1.06529	1.04995	0.70721	1.06080	0.50380	0.91061
12	Hawaii	1.10652	1.09525	0.45527	1.10038	0.21166	0.69889
13	Idaho	1.03265	1.06746	0.37944	1.05425	0.14814	0.61074
14	Illinois	1.03504	1.09350	0.84473	1.04412	0.71589	0.97357
15	Indiana	1.00212	1.05179	0.61762	1.02111	0.38563	0.84961
16	Iowa	1.13109	1.07058	0.47988	1.09962	0.23469	0.72506
17	Kansas	1.07790	1.09118	0.62394	1.08289	0.39345	0.85443
18	Kentucky	1.04731	1.05219	0.72660	1.04865	0.53145	0.92174
19	Louisiana	0.99794	1.02531	0.69639	1.00625	0.48870	0.90409
20	Maine	1.06623	1.06694	0.69604	1.06645	0.48821	0.90387
21	Maryland	1.11723	1.07953	0.76824	1.10849	0.59334	0.94314
22	Massachusetts	1.15534	1.12785	0.83892	1.15091	0.70618	0.97166
23	Michigan	1.06529	1.08287	0.80384	1.06874	0.64894	0.95873
24	Minnesota	1.08561	1.11693	0.85396	1.09019	0.73145	0.97647
25	Mississippi	0.99608	1.01842	0.71197	1.00251	0.51053	0.91342

Table 3 continued

	STATE	y	$x'\hat{\beta}$	γ	$\hat{\theta}$	L.L.	U.L.
26	Missouri	1.05847	1.06318	0.80648	1.05938	0.65316	0.95979
27	Montana	1.05871	1.07378	0.52071	1.06593	0.27555	0.76587
28	Nebraska	1.07682	1.10131	0.50497	1.08894	0.25941	0.75053
29	Nevada	1.03356	1.02295	0.31109	1.02625	0.10056	0.52161
30	New Hampshire	1.13083	1.10371	0.52596	1.11797	0.28104	0.77088
31	New Jersey	1.07383	1.09149	0.80306	1.07731	0.64769	0.95842
32	New Mexico	0.92677	1.00499	0.45629	0.96930	0.21258	0.69999
33	New York	1.09796	1.07945	0.92424	1.09656	0.85545	0.99302
34	North Carolina	1.03038	1.05896	0.79596	1.03621	0.63642	0.95550
35	North Dakota	1.08139	1.08619	0.22328	1.08512	0.05292	0.39364
36	Ohio	1.10328	1.06747	0.84758	1.09783	0.72067	0.97448
37	Oklahoma	1.05532	1.02779	0.76275	1.04879	0.58499	0.94052
38	Oregon	1.04041	1.07866	0.62788	1.05464	0.39836	0.85740
39	Pennsylvania	1.10080	1.08103	0.82972	1.09744	0.69093	0.96851
40	Rhode Island	1.19764	1.09811	0.42828	1.14074	0.18775	0.66881
41	South Carolina	1.05392	1.04741	0.71028	1.05203	0.50814	0.91243
42	South Dakota	1.03038	1.07693	0.09482	1.07252	0.01051	0.17913
43	Tennessee	1.09024	1.04420	0.79321	1.08071	0.63208	0.95434
44	Texas	1.00598	1.01310	0.94163	1.00640	0.88764	0.99562
45	Utah	1.03243	1.08697	0.47525	1.06105	0.23027	0.72023
46	Vermont	1.06659	1.10686	0.19413	1.09904	0.04045	0.34781
47	Virginia	1.09292	1.08670	0.81669	1.09178	0.66964	0.96375
48	Washington	1.08513	1.09427	0.78455	1.08710	0.61850	0.95059
49	West Virginia	1.03618	1.01954	0.32898	1.02501	0.11213	0.54583
50	Wisconsin	1.08380	1.09456	0.81777	1.08576	0.67138	0.96416
51	Wyoming	1.08030	1.05342	0.52217	1.06746	0.27707	0.76727

Table 4: Direct estimates (z), composite estimates ($\hat{\pi}$), and synthetic estimates ($\hat{\pi}^{syn}$)

	State	z	$\hat{\pi}$	$\hat{\pi}^{syn}$
1	Alabama	0.7785	0.77451	0.75134
2	Alaska	0.6990	0.75873	0.78464
3	Arizona	0.7573	0.74967	0.72768
4	Arkansas	0.7529	0.74401	0.71975
5	California	0.7443	0.74440	0.74634
6	Colorado	0.7873	0.79333	0.80371
7	Connecticut	0.8439	0.83795	0.82546
8	Delaware	0.8166	0.77925	0.76277
9	District of Columbia	0.8656	0.82719	0.81046
10	Florida	0.7103	0.71691	0.74832
11	Georgia	0.7655	0.76168	0.75238
12	Hawaii	0.7995	0.79456	0.79040
13	Idaho	0.7373	0.75609	0.76734
14	Illinois	0.7394	0.74733	0.78897
15	Indiana	0.7100	0.72708	0.75397
16	Iowa	0.8188	0.79394	0.76997
17	Kansas	0.7761	0.78025	0.78707
18	Kentucky	0.7501	0.75125	0.75431
19	Louisiana	0.7062	0.71374	0.73082
20	Maine	0.7663	0.76648	0.76690
21	Maryland	0.8080	0.80107	0.77746
22	Massachusetts	0.8371	0.83382	0.81630
23	Michigan	0.7655	0.76842	0.78023
24	Minnesota	0.7825	0.78626	0.80777
25	Mississippi	0.7045	0.71036	0.72468

Table 4 continued

	State	z	$\hat{\pi}$	$\hat{\pi}^{syn}$
26	Missouri	0.7597	0.76048	0.76371
27	Montana	0.7599	0.76604	0.77265
28	Nebraska	0.7752	0.78524	0.79531
29	Nevada	0.7381	0.73165	0.72872
30	New Hampshire	0.8186	0.80859	0.79724
31	New Jersey	0.7727	0.77561	0.78733
32	New Mexico	0.6395	0.67978	0.71260
33	New York	0.7926	0.79146	0.77739
34	North Carolina	0.7353	0.74043	0.76012
35	North Dakota	0.7790	0.78209	0.78297
36	Ohio	0.7969	0.79249	0.76735
37	Oklahoma	0.7570	0.75138	0.73301
38	Oregon	0.7441	0.75642	0.77673
39	Pennsylvania	0.7949	0.79217	0.77871
40	Rhode Island	0.8671	0.82618	0.79272
41	South Carolina	0.7558	0.75418	0.75018
42	South Dakota	0.7353	0.77160	0.77529
43	Tennessee	0.7863	0.77844	0.74740
44	Texas	0.7135	0.71388	0.71991
45	Utah	0.7371	0.76190	0.78362
46	Vermont	0.7666	0.79348	0.79977
47	Virginia	0.7885	0.78757	0.78339
48	Washington	0.7821	0.78372	0.78960
49	West Virginia	0.7404	0.73055	0.72568
50	Wisconsin	0.7810	0.78262	0.78983
51	Wyoming	0.7781	0.76733	0.75537

Table 5: Design effects (D_0), standard errors of direct estimates [$se(z)$], and jackknife standard errors of composite estimates [$se(\hat{\pi})$], and the percent improvement ($PCTIMP$)

STATE	D_0	$se(z)$	$se(\hat{\pi})$	$CV(z)$	$CV(\hat{\pi})$	$PCTIMP$
1	0.94237	0.0099	0.009678	0.01272	0.012495	2.2442
2	1.00743	0.0385	0.027225	0.05508	0.035889	29.2862
3	1.77080	0.0145	0.013842	0.01915	0.018464	4.5357
4	0.97357	0.0150	0.014295	0.01992	0.019212	4.7026
5	2.33951	0.0055	0.005454	0.00739	0.007327	0.8372
6	2.28631	0.0176	0.015935	0.02235	0.020087	9.4589
7	2.15494	0.0144	0.013541	0.01706	0.016159	5.9646
8	1.89534	0.0339	0.027026	0.04151	0.034678	20.2782
9	1.86680	0.0304	0.024869	0.03512	0.030061	18.1926
10	3.67215	0.0117	0.011234	0.01647	0.015671	3.9842
11	3.32450	0.0155	0.014593	0.02025	0.019159	5.8500
12	2.07315	0.0249	0.021185	0.03114	0.026662	14.9180
13	2.14117	0.0320	0.025036	0.04340	0.033115	21.7635
14	2.47474	0.0107	0.010277	0.01447	0.013752	3.9540
15	4.29302	0.0203	0.018116	0.02859	0.024918	10.7605
16	2.97118	0.0228	0.020464	0.02785	0.025772	10.2464
17	1.82169	0.0184	0.016652	0.02371	0.021342	9.5004
18	1.76132	0.0151	0.014196	0.02013	0.018897	5.9837
19	2.16050	0.0171	0.015837	0.02421	0.022190	7.3858
20	0.61690	0.0159	0.014833	0.02075	0.019352	6.7100
21	1.63347	0.0123	0.011871	0.01522	0.014819	3.4840
22	1.38477	0.0092	0.008985	0.01099	0.010775	2.3375
23	2.77288	0.0119	0.011386	0.01555	0.014818	4.3202
24	0.93043	0.0097	0.009371	0.01240	0.011918	3.3941
25	1.11029	0.0165	0.015381	0.02342	0.021653	6.7832

Table 5 continued

STATE	D_0	$se(z)$	$se(\hat{\pi})$	$CV(z)$	$CV(\hat{\pi})$	$PCTIMP$
26	1.61813	0.0119	0.011430	0.01566	0.015030	3.9498
27	1.06524	0.0233	0.020155	0.03066	0.026312	13.4962
28	1.85705	0.0235	0.020023	0.03031	0.025500	14.7958
29	3.22858	0.0372	0.028319	0.05040	0.038705	23.8739
30	1.13044	0.0208	0.018578	0.02541	0.022975	10.6826
31	2.57178	0.0118	0.011286	0.01527	0.014551	4.3583
32	4.22565	0.0298	0.024433	0.04660	0.035949	18.0113
33	1.91692	0.0066	0.006523	0.00833	0.008242	1.1667
34	2.07917	0.0127	0.012104	0.01727	0.016349	4.6896
35	2.41776	0.0440	0.029636	0.05648	0.037893	32.6465
36	2.18351	0.0097	0.009491	0.01217	0.011976	2.1512
37	1.41874	0.0136	0.013033	0.01797	0.017345	4.1681
38	2.39483	0.0191	0.017143	0.02567	0.022665	10.2484
39	2.74191	0.0104	0.010104	0.01308	0.012755	2.8445
40	1.56647	0.0223	0.020640	0.02572	0.024979	7.4418
41	1.58622	0.0156	0.014652	0.02064	0.019428	6.0756
42	8.20855	0.0775	0.035798	0.10540	0.046397	53.8087
43	1.65433	0.0119	0.011554	0.01513	0.014842	2.9060
44	1.84504	0.0064	0.006331	0.00897	0.008869	1.0754
45	2.30226	0.0263	0.021702	0.03568	0.028487	17.4841
46	3.52922	0.0490	0.030204	0.06392	0.038067	38.3593
47	1.68620	0.0110	0.010620	0.01395	0.013484	3.4547
48	1.64145	0.0123	0.011732	0.01573	0.014969	4.6200
49	3.36938	0.0356	0.027687	0.04808	0.037898	22.2287
50	1.37013	0.0111	0.010674	0.01421	0.013639	3.8400
51	0.36386	0.0226	0.020070	0.02905	0.026154	11.1955

Table 6: States arranged in increasing order of V_i

Rank of V_i	State	State ID	V_i
1	California	5	.0000303
2	Texas	44	.0000410
3	New York	33	.0000436
4	Massachusetts	22	.0000846
5	Minnesota	24	.0000941
6	Ohio	36	.0000941
7	Alabama	1	.0000980
8	Pennsylvania	39	.0001082
9	Illinois	14	.0001145
10	Virginia	47	.0001210
11	Wisconsin	50	.0001232
12	Florida	10	.0001369
13	New Jersey	31	.0001392
14	Michigan	23	.0001416
15	Missouri	26	.0001416
16	Tennessee	43	.0001416
17	Maryland	21	.0001513
18	Washington	48	.0001513
19	North Carolina	34	.0001613
20	Oklahoma	37	.0001850
21	Connecticut	7	.0002074
22	Arizona	3	.0002103
23	Arkansas	4	.0002250
24	Kentucky	18	.0002280
25	Georgia	11	.0002403

Table 6 continued

Rank of V_i	State	State ID	V_i
26	South Carolina	41	.0002434
27	Maine	20	.0002528
28	Mississippi	25	.0002723
29	Louisiana	19	.0002924
30	Colorado	6	.0003098
31	Kansas	17	.0003386
32	Oregon	38	.0003648
33	Indiana	15	.0004121
34	New Hampshire	30	.0004326
35	Rhode Island	40	.0004973
36	Wyoming	51	.0005108
37	Iowa	16	.0005198
38	Montana	27	.0005429
39	Nebraska	28	.0005523
40	Hawaii	12	.0006200
41	Utah	45	.0006917
42	New Mexico	32	.0008880
43	District of Columbia	9	.0009242
44	Idaho	13	.0010240
45	Delaware	8	.0011492
46	West Virginia	49	.0012674
47	Nevada	29	.0013838
48	Alaska	2	.0014823
49	North Dakota	35	.0019360
50	Vermont	46	.0024010
51	South Dakota	42	.0060063

Figure 1: Estimates [direct (D), synthetic (S), and EBLUP (C)] Plotted against States Arranged in Increasing Order of V_i (see Table 6 for identifying the states)

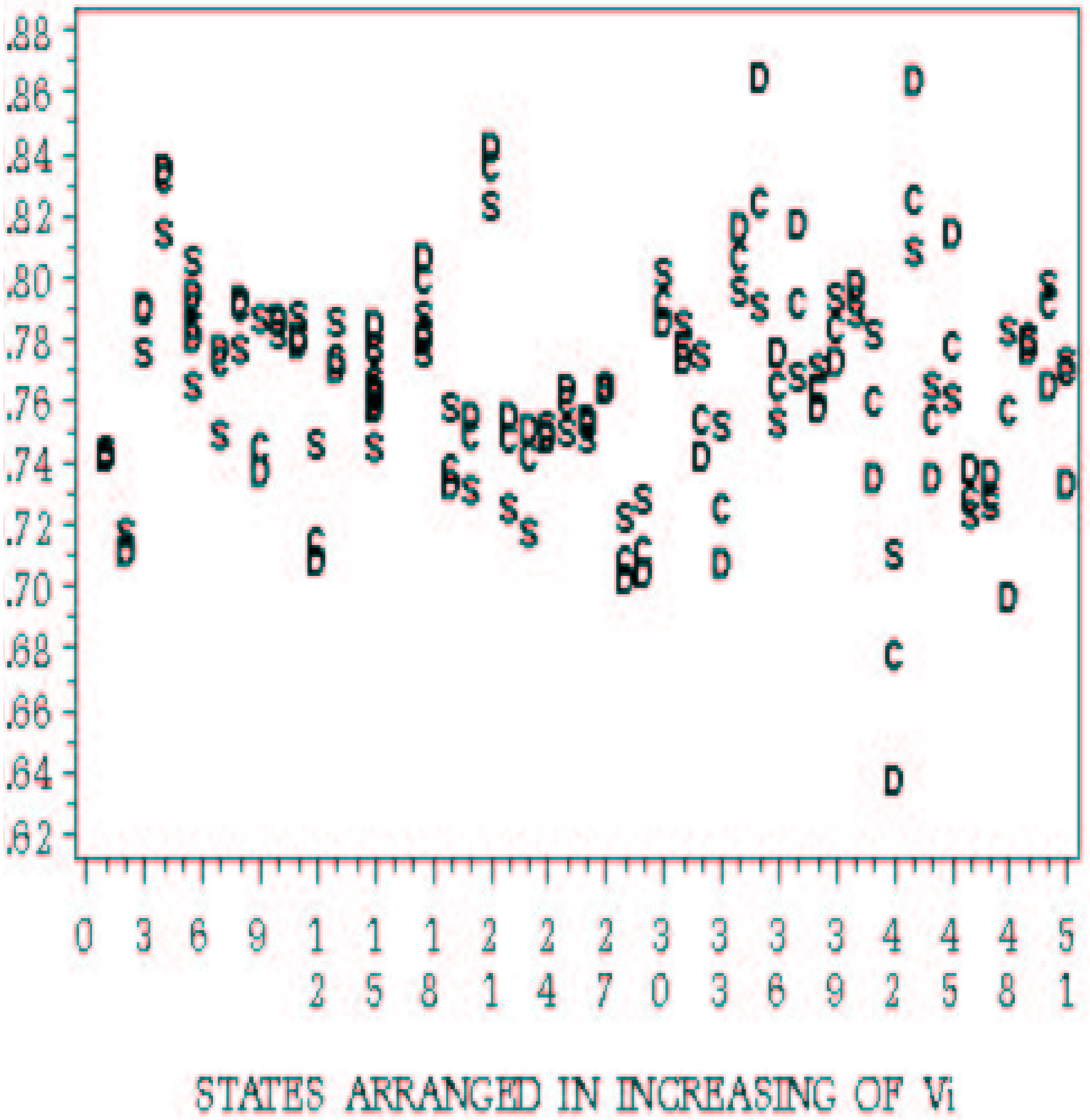


Figure 2: Percent Improvement Plotted against States Arranged in Increasing Order of V_i (see Table 6 for identifying the states)

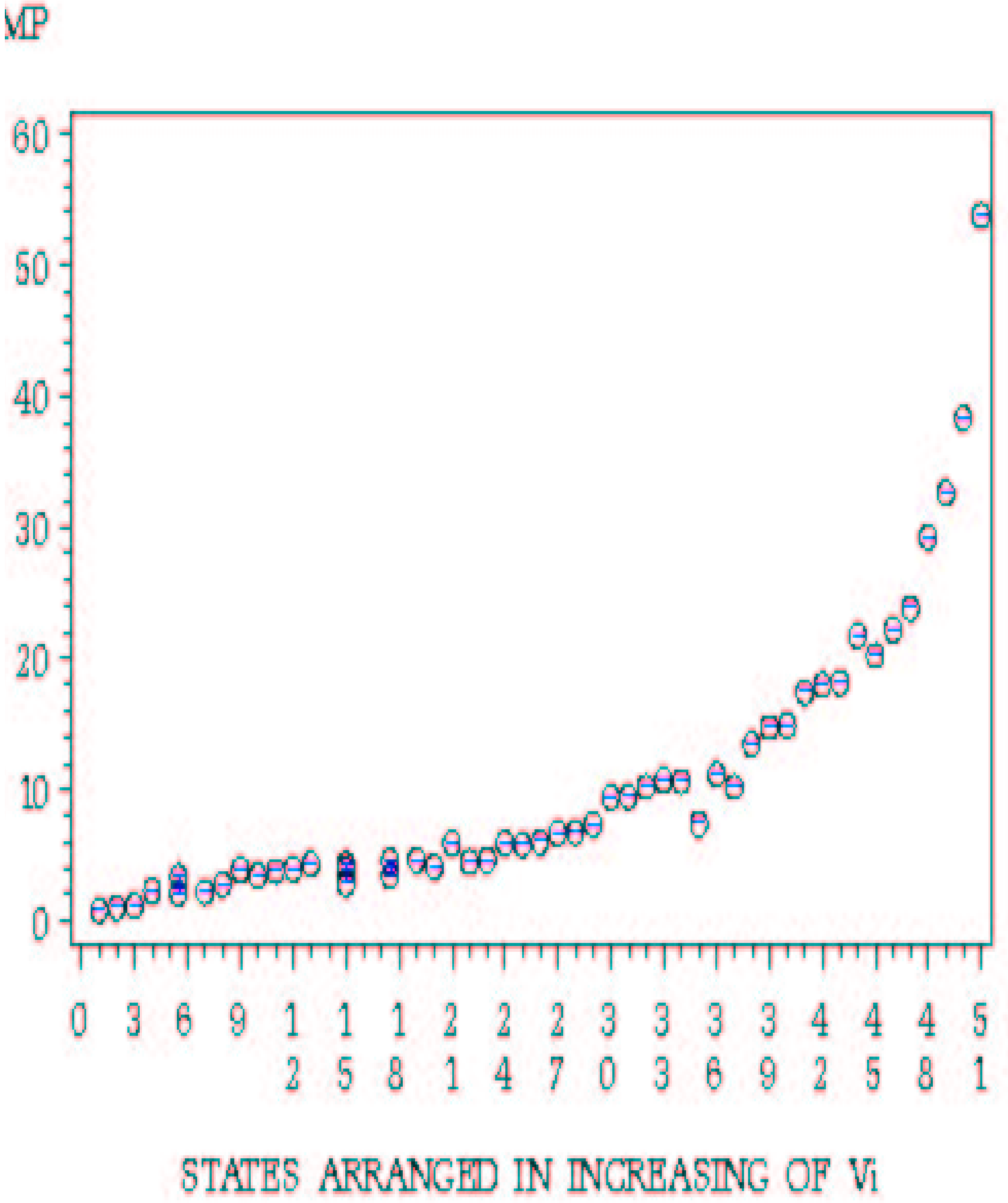
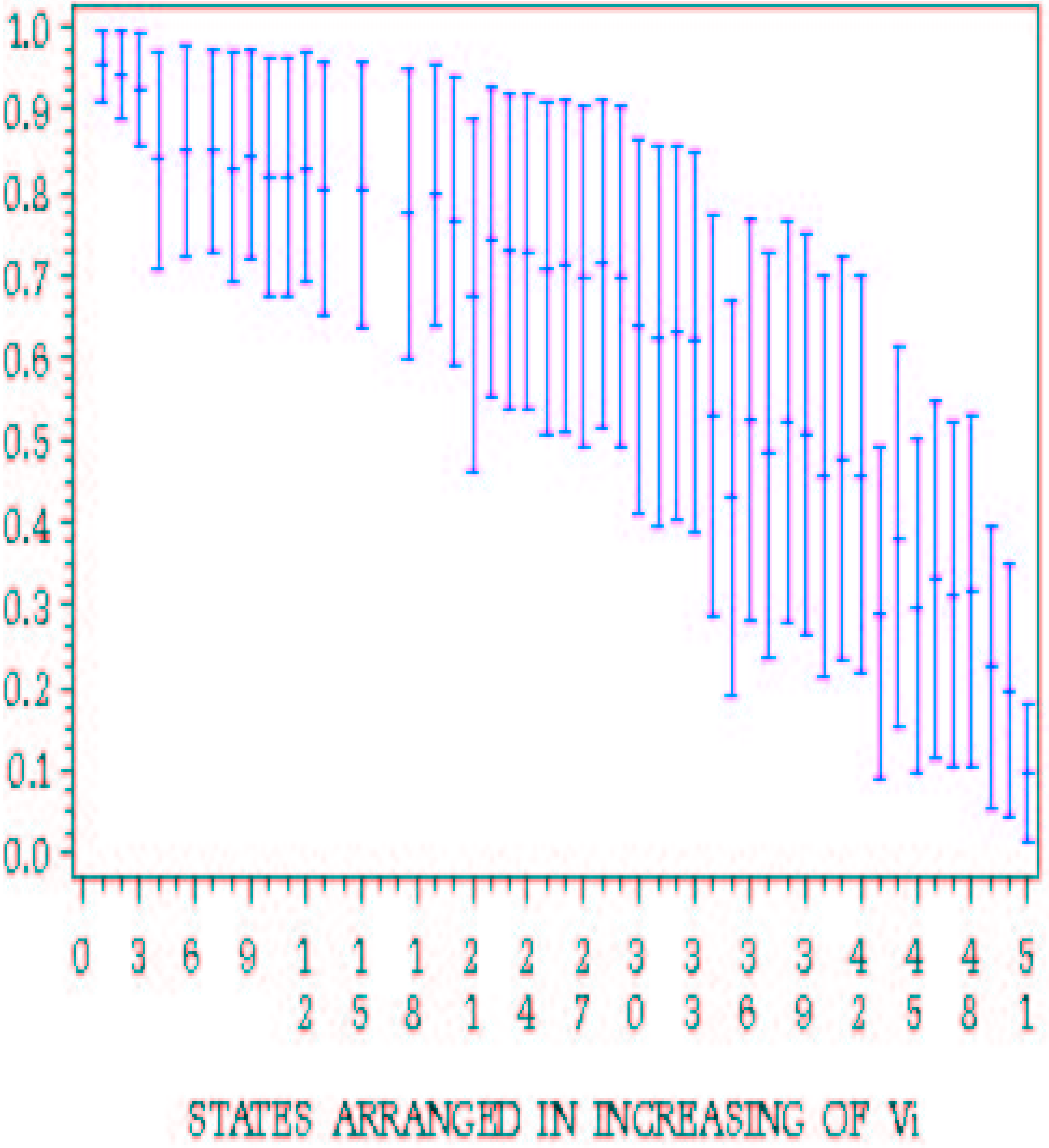


Figure 3: Confidence Interval for γ Plotted against States Arranged in Increasing Order of V_i (see Table 6 for identifying the states)



Jackknifing in the Fay-Herriott Model with An Example

by Jiming Jiang, Partha Lahiri, Shu-Mei Wan, and Chien-Hua Wu

Discussion (corrected version, 10/19/2001)

by William R. Bell, U.S. Bureau of the Census

The paper by Jiang, Lahiri, Wan, and Wu (hereafter JLWW) considers use of the jackknife to estimate the mean squared error (MSE) of small area estimates from Fay-Herriott (1979) models. The paper notes that Jiang *et al.* (2001) discuss use of the jackknife more generally for estimating MSE with nonlinear and nonnormal small area models. As the present paper restricts consideration to the linear model case, my remarks will focus only on this case. It should be kept in mind, however, that ignoring the generality of the jackknife may be ignoring one of its prime advantages.

The model of Fay and Herriott (1979) for small area estimation can be written

$$
\begin{aligned}
y_i &= \theta_i + e_i \qquad i = 1, \ldots, m && (1) \\
&= (x_i'\beta + v_i) + e_i && (2)
\end{aligned}
$$

JLWW give detailed assumptions underlying this model. Here I simply repeat that the sampling variances $D_i = \text{Var}(e_i)$ of the direct survey estimates y_i are assumed known (actually meaning they are estimated using survey microdata), so that the unknown parameters of the model given by (1) and (2) are the regression parameters β and the model error variance $A = \text{Var}(v_i)$. To apply this model from a frequentist perspective one first estimates the model parameters β and A using the direct estimates y_i, and then applies standard empirical Bayes prediction formulas to produce point estimates of the θ_i. A Bayesian approach can also be used (Berger 1985, Bell 1999).

Assuming the model given by (1) and (2) is true (more on this later), the error in the estimates of the θ_i can be broken into three terms:

$$
\begin{aligned}
\text{error} \quad = \quad & \text{error when all parameters are known} \\
& + \text{contribution to error from estimating } \beta \qquad (3) \\
& + \text{contribution to error from estimating } A
\end{aligned}
$$

The mean square of this error for area i is, under suitable assumptions,

$$
\begin{aligned}
MSE_i &= g_{1i}(A) + g_{2i}(A) + g_{3i}(A) \\
&= A(1-\gamma_i) + (1-\gamma_i)^2 x_i'\text{Var}\left(\hat{\beta}\right) x_i + g_{3i}(A) && (4)
\end{aligned}
$$

where

$$\gamma_i = A/(A + D_i)$$

Since A is unknown to estimate MSE we plug an estimate of A into (4). Results for the term $g_{3i}(A)$ are discussed shortly. The "suitable assumptions" referred to above include normality, which is relevant in regard to asymptotic orthogonality of the second and third terms in (3). My focus, however, will be on comparing the form of (4) with the jackknife estimate of MSE suggested by JLWW, and on examining results from the jackknife for a particular empirical example. For simplicity of notation I will henceforth drop the subscript i that indexes the small areas. It should be understood that all subsequent expressions implicitly depend on i, e.g., through x_i and D_i.

Two problems arise in applying (4):

- $g_1(\widehat{A})$ is biased. In fact, even when $\widehat{A}$ is approximately unbiased, $E[g_1(\widehat{A})] \approx g_1(A) - g_3(A)$.
- There is no exact formula for $g_3(A)$.

Several approaches have previously been suggested to deal with the second problem: ignore $g_3(A)$ (naive approach); estimate $g_3(A)$ using an asymptotic expression (Prasad and Rao 1990, Datta and Lahiri 2000); or use a Bayesian approach (Berger 1985, p. 192, Bell 1999). JLWW propose the jackknife to address both of the two problems.

Before examining how the jackknife addresses the two problems noted, it is worth reminding ourselves of a third problem, which is that use of the MSE result (4) depends on the model being correct. This problem may well be more important and more difficult to address than either of the other two problems. It compromises all the approaches noted to an unknown degree for any particular example.

JLWW's jackknife estimate of MSE is

$$MSE_J = \left\{ g_1(\widehat{A}) - \frac{m-1}{m} \sum_{u=1}^{m} [g_1(\widehat{A}_{-u}) - g_1(\widehat{A})] \right\} + \frac{m-1}{m} \sum_{u=1}^{m} (\widehat{\theta}_{-u} - \widehat{\theta}) \qquad (5)$$

The term in braces estimates $g_1(A)$, with the second term within the braces providing the jackknife bias correction to the plug-in estimate $g_1(\widehat{A})$. The last term in (5) estimates $g_2(A) + g_3(A)$ together, not just $g_3(A)$. These features provide for some generality of the jackknife (e.g., to nonnormal models), though it means that in the context of the linear model ((1),(2)) considered here (5) does not make use of either (*i*) the asymptotic relation between bias($g_1(\widehat{A})$) and $g_3(A)$, or (*ii*) the exact result for $g_2(A)$. The question arises as to when does the jackknife work better, worse, or about the same as alternatives?

In regard to the question of "How well does the jackknife work?," Jiang *et al.* (2001) report simulation results for some linear and nonlinear (GLIM) models. The jackknife works well in the simulations reported, however, so do all the other approaches considered. In fact, the worst case reported in the simulations of bias in estimated MSE for any method is -10.1% for the naive approach (for a mixed logistic model). This is a relatively small understatement of MSE since, if resulting MSE estimates were used to construct prediction intervals, the corresponding "standard error" would be understated by only about 5%. Given uncertainties about normality assumptions needed to construct prediction intervals, this amount of understatement of prediction standard error seems relatively unimportant.

In the present paper, the jackknife results JLWW present for the NHIS application look quite reasonable. I decided to examine results from the proposed jackknife approach for an application I am familiar with: estimation of poverty rates of school-aged (5-17 year old) children for the states of the U.S. and DC. These estimates are an important product of the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program. The Fay-Herriott model used to produce these estimates is developed in Fay and Train (1997). Bell (1999) discusses Bayesian treatment of this model. Further information on the SAIPE program can be found on the SAIPE web site at http://www.census.gov/hhes/www/saipe.html.

For applying the jackknife to the SAIPE example I used the method-of-moments (MOM) estimator of A used by JLWW in their NHIS example. Table 1 shows these estimates of A for nine years of data to which the model was applied. (1994 is omitted because sampling variances have not been estimated for this year due to complications caused in that year by transition to a redesign of the Current Population Survey which supplies the direct estimates y_i.) Both the not truncated and truncated (at 0) MOM estimates are shown. Maximum likelihood (ML), restricted maximum likelihood (REML), and Bayesian estimates (posterior means) are shown for comparison. These all assume normality, and the Bayesian estimates use flat priors for β and A.

Table 1. Alternative Estimates of A(SAIPE example)

year	ML	REML	Bayes	not truncated MOM	truncated MOM
1989	0	0	1.7	−.1	0
1990	0	0	2.2	1.1	1.1
1991	0	0	1.6	−3.1	0
1992	0	0	1.6	−3.2	0
1993	.4	1.7	3.4	5.8	5.8
1995	0	.2	2.0	.5	.5
1996	0	0	1.9	2.0	2.0
1997	0	0	1.5	−1.3	0
1998*	.7	2.0	3.7	5.8	5.8

*Preliminary results

We notice that the MOM estimates of A are rather unstable. The truncation at zero is frequently required, and for 1993 and 1998 the MOM estimate is quite large relative to the other estimates. ML and REML, though more stable, are not very appealing since these estimates are zero in most years. Bell (1999) notes how estimating A at zero leads to unreasonable MSE estimates for ML and REML. On the other hand, the Bayesian estimates of A appear much more reasonable (and Bell (1999) notes that resulting Bayesian posterior variances are more reasonable than the frequentist MSE estimates.)

Though I shall omit giving detailed results, it turns out that the jackknife MSE estimates look unreasonable both when A is estimated to be zero and when the MOM estimates of A are large. The poor performance is not generally the fault of the jackknife, however, but simply a result of getting unreasonable estimates of A from MOM. In principle the jackknife could be applied with ML or REML estimation of A, at significantly higher computational cost, though the results in Table 1 suggest this would rarely help in this example. Rather than dwell on possibilities for improving jackknife results by using alternative estimates of A, I will compare jackknife MSE results (using $\widehat{A}_{MOM}$) for two years to illustrate a particular problem that arises when A is estimated at or near zero, and that is more pertinent to the performance of the jackknife.

Notice that when $\widehat{A} = 0$ $g_1(\widehat{A}) = 0$, and all of MSE_J comes from the second and third terms in (5). Tables 2 and 3 below examine the components of MSE_J for two years (1991 and 1989) for which $\widehat{A}_{MOM} = 0$. For both years results are shown for a small number of states for illustration. In the table headings $\widetilde{A}_{MOM}$ denotes the original, not truncated MOM estimates of A from Table 1. Bayesian posterior variances are shown for comparison. It is worth noting that for most states in most years, these posterior variances are very close to what one obtains by substituting the posterior mean of A into $g_1(A)$.

Table 2. Jackknife estimation of MSE for 1991 (SAIPE example)
($\widetilde{A}_{MOM} = -3.1$, $\widehat{A}_{MOM} = \max(0, \widetilde{A}_{MOM}) = 0$)

state	$g_1(\widehat{A})$	$\widetilde{A}_{-u}$	$\widehat{bias}[g_1(\widehat{A})]$	$\widehat{g_2 + g_3}$	**MSE**$_J$	**Bayes**
AL	0	−2.7	0	.7	.7	2.0
AK	0	−2.8	0	.5	.5	2.3
AZ	0	−3.1	0	.5	.5	2.0
AR	0	−3.2	0	1.3	1.3	2.5
CA	0	−2.9	0	.6	.6	1.4
CO	0	−2.8	0	.2	.2	1.6

The $\widetilde{A}_{-u}$ columns in Tables 2 and 3 give the leave-one-out not truncated MOM estimates of A. In 1991 (Table 2) all of the $\widetilde{A}_{-u}$ are negative, with the result that all of the truncated leave-one-out MOM estimates of A ($\widehat{A}_{-u}$) are zero. This is not surprising given that the full sample not truncated MOM estimate for 1991 ($\widetilde{A}_{MOM} = -3.1$) is well below zero—dropping any one observation does not have enough effect to turn any of the $\widetilde{A}_{-u}$ positive. As a result $g_1(\widehat{A}_{-u}) = 0$ for all states u, and since $g_1(\widehat{A}) = 0$ as well, the second term in (5) estimating the bias in $g_1(\widehat{A})$ is zero. Hence, both terms in the braces in (5) are zero for every state, and MSE_J comes entirely from the third term in (5). This term (labelled $\widehat{g_2+g_3}$ in the tables) is a jackknife estimate reflecting variation in the small area point estimates $\widehat{\theta}$ due to variation in the leave-one-out estimates of β and A. The resulting MSE estimates tend to look too small both in an absolute sense and relative to the Bayesian estimates—note Colorado (CO) in particular. The MSE estimates also exhibit the same sort of pattern problems noted in Bell (1999) for the ML and REML estimates based on (4). For example, MSE_J for California (CA), despite its large CPS sample, is as high or higher than that for many other states with much smaller samples.

Table 3 shows a different problem that arises for the jackknife estimate of MSE. For 1989 the not truncated MOM estimate of A, $\widetilde{A}_{MOM} = -.1$, is very close to zero. Dropping one observation alters the not truncated MOM estimates as shown in the $\widetilde{A}_{-u}$ column, sometimes yielding positive values, and sometimes yielding negative values. When $\widetilde{A}_{-u} < 0$, $\widehat{A}_{-u}$ is truncated to 0, and $g_1(\widehat{A}_{-u}) = 0$. These states make no contribution to the second term in (5). When $\widetilde{A}_{-u} > 0$, however, $g_1(\widehat{A}_{-u})$ is positive, and these states do contribute to the second term in (5). In fact, since $g_1(\widehat{A}) = 0$ here, the term in braces in (5) is simply minus the sum of these positive terms (multiplied by $(m-1)/m = 50/51$), which turns out to be around 2 for each state. This is the jackknife estimate of bias in $g_1(\widehat{A})$. Subtracting off this bias estimate of around 2 overwhelms the third term in (5), $\widehat{g_2+g_3}$, resulting in negative estimates of MSE for every state. (The estimates of bias in $g_1(\widehat{A})$ vary only slightly over states since, reintroducing the state subscript i, for state i this term is actually $-\frac{m-1}{m}\sum_{u=1}^{m}[g_{1i}(\widehat{A}_{-u})]$ where $g_{1i}(\widehat{A}_{-u}) = \widehat{A}_{-u}D_i/(\widehat{A}_{-u}+D_i) = \widehat{A}_{-u}/(1+\widehat{A}_{-u}/D_i) \approx \widehat{A}_{-u}$ since the D_i are much larger than the $\widehat{A}_{-u}$.)

Table 3. Jackknife estimation of MSE for 1989 (SAIPE example)
($\widetilde{A}_{MOM} = -.1$, $\widehat{A}_{MOM} = \max(0, \widetilde{A}_{MOM}) = 0$)

state	g_1	$\widetilde{A}_{-u}$	$\widehat{bias}[g_1]$	$\widehat{g_2+g_3}$	**MSE**$_J$	**Bayes**
AL	0	.16	1.97	.52	−1.4	2.1
AK	0	.04	1.96	.70	−1.2	2.4
AZ	0	−.09	1.97	.37	−1.5	2.1
AR	0	.18	1.97	.71	−1.2	2.4
CA	0	−.03	1.85	.60	−1.2	1.1
CO	0	.10	1.96	.18	−1.7	1.6
CT	0	−1.50	1.96	1.30	−.6	3.0
DE	0	.16	1.97	.38	−1.5	1.8
DC	0	.02	1.98	.90	−1.0	3.3
FL	0	−.13	1.91	.55	−1.3	1.4

The negative estimates of MSE result not just from the poor estimation of A by MOM (though this is a necessary part of the problem), but also from poor estimation of the bias in $g_1(\widehat{A})$ by the jackknife. This would appear to be a potential problem for the jackknife any time the not truncated estimate of A is very close to zero.

Since the jackknife MSE estimates JLWW present for their NHIS application look quite reasonable, this raises a question about how the SAIPE application differs from the NHIS application. Table 4 provides some answers. It shows, for both surveys, the maximum and minimum values of the estimated signal-to-noise ratio $(\widehat{A}/D_i)$ across states, as well as the ratio of the maximum to the minimum state sampling variances $(\max(D_i)/\min(D_i))$.

Table 4. Comparing the NHIS and SAIPE Examples

	$\max\left(\frac{\widehat{A}}{D_i}\right)$	$\min\left(\frac{\widehat{A}}{D_i}\right)$	$\frac{\max(D_i)}{\min(D_i)}$
NHIS	20 (CA)	.1 (SD)	200
SAIPE	1 to 1.5 (CA)	.07 to .1 (DC)	15 to 20

For the NHIS application the signal-to-noise ratio ranges from a very high value of 20 to a very low value of .1. In contrast, for the SAIPE application (for which the results vary some over the years of data) the smallest signal-to-noise ratio is about the same as that for NHIS, but the largest is only around 1 or 1.5. The corresponding ratio of the maximum to minimum sampling variances is 200 for NHIS, reflecting a very wide range of sampling variance across states, but is only 15 or 20 for SAIPE. These data suggest that in the NHIS application the states with large samples provide enough information for reasonably reliable estimation of A (here by MOM), which leads to reasonable looking estimates of MSE by the jackknife (and presumably by other approaches). Small area estimation is needed for those states with small NHIS samples (low signal-to-noise ratios). On the other hand, the CPS direct estimates used in the SAIPE application have sufficiently high levels of sampling error that estimates of A are more unreliable, and conventional frequentist estimates of A frequently run into trouble (as can be seen from Table 1). Resulting estimates of MSE can be unreasonable, and if the not truncated estimate of A is near zero, this can lead to the problem illustrated for the jackknife estimate of MSE.

To generalize the conclusions a bit, estimation of A in the Fay-Herriott model appears to be of more fundamental importance than the choice of alternative approaches to estimating MSE, in the sense that when the data do not provide enough information for reliable estimation of A by conventional frequentist methods, any resulting estimates of MSE are suspect. The Bayesian approach appears to yield more reasonable results in such cases at least by preventing estimates of A near zero. The appeal of the jackknife may be more for cases where nonnormality is a serious concern or the model is nonlinear (e.g., GLIM models), though its performance is still likely to depend on whether or not the data provide sufficient information for reliable estimation of variances or other dispersion parameters of the model.

References

[1] Bell, William R. (1999) "Accounting for Uncertainty About Variances in Small Area Estimation," *Bulletin of the International Statistical Institute*, 52nd Session, Helsinki.

[2] Berger, James O., (1985) *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

[3] Fay, R. E. and Herriott, R. A. (1979) "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association*, **74**, 269-277.

[4] Fay, Robert E. and Train, George F. (1997) "Small Domain Methodology for Estimating Income and Poverty Characteristics for States in 1993," American Statistical Association, Proceedings of the Social Statistics Section, 183-188.

[5] Jiang, J., Lahiri, P., and Wan, S. (2001), "A Unified Jackknife Theory," paper submitted for publication.

Can What Partha Lahiri and Company Have Done Help the National Agricultural Statistics Service?

Phillip S. Kott; USDA/NASS

The National Agricultural Statistics Service (NASS) is using component-of-variance small-domain estimation to help estimate the undercoverage in the US Census of Agriculture. The backbone of the Census of Agriculture is an extensive (but not exhaustive) list of farms in the US. Although farms on this list are responsible for over 95% of most agricultural activities, NASS estimates that 13% of all farms operating in 1997 were not on the Census list. One reason for this is that the federal government uses a very liberal definition of a farm: an operation producing at least $1,000 of agricultural output in a year or capable of producing that output.

Leaving aside the merits of the government's definition of a farm, NASS wants the 2002 Census of Agriculture to do a better job than it has in the past providing state aggregates for the farms *not* on the Census list. NASS's key instrument in that endeavor is an area-frame sample designed primarily to estimate the total corn, wheat, soy beans, cotton, and potato acreage and production in the US. A secondary use of this sample is the estimation of aggregates for farms not on NASS's survey and Census lists. For the 2002 Census effort, this sample will be supplemented to better measure the Census-list undercoverage. Nevertheless, NASS believes that only the total numbers of farms missing form the Census list will be reliably estimated at the state level (and even then, certain states like those in New England will have to be combined). For aggregates like the number of missing farms that have horses or the number of missing farms operated by blacks, small-domain techniques will be needed that draw strength from states outside the particular state of interest.

For the purposes of this discussion, let us focus on one particular estimate: the fraction of farms not on the Census list that have a black operator. In truth, NASS will be estimating 20 or 30 fraction like this one, from the fraction of missing farms with horses to the fraction of missing farms with annual sales in a given range. Conceptually, however, they are all the same. One thing to note is that the fractions are many and disparate. Consequently, unlike the problem in Jiang *et al.* (2001), NASS uses a single covariate the fraction of farms on a state's Census list *with* the attribute in question (e.g., a black operator). To begin, we will ignore even that covariate.

Background

Partha Lahiri and his team of collaborators, Jiming Jiang, Shu-Mei Wan, and Chien-Hua Wu, have written a number of papers based on research funded by federal statistical agencies through the National Science Foundation. The question I will address here is whether that research can be of service to NASS in its attempt to estimate the fraction of black-operated farms among those missing from a state's Census list. I begin with some notation borrowing liberally from Jiang *et al.*, the particular paper under review. Throughout this discussion, I will refer to the "Lahiri team." That should not be construed as a denigration of Jiang and the other collaborators' contributions.

Let z_{ij} be 1 if a missing farm j in the area sample of State i has a black operator; 0 otherwise. The direct, randomization-based estimator for π_i, the fraction of black-run operations among those missing from the Census list in State i, is

$$z_i = \sum_{j\in S(i)} w_{ij}z_{ij} / \sum_{j\in S(i)} w_{ij},$$

where w_{ij} is the sampling weight attached to farm j, and S(i) is the set of farms in the area sample of State i but not on the State's Census list.

That variance of this model under a simple Bernoulli model is

$$\begin{aligned} \mathrm{Var}(z_i) &= \{\textstyle\sum_{j\in S(i)} w_{ij}^2 / [\sum_{j\in S(i)} w_{ij}]^2\}\, \pi_i\,(1 \quad \pi_i) \\ &= \pi_i\,(1 \quad \pi_i)/n_i^*, \end{aligned}$$

where $n^* = \{ [\sum_{j\in S(i)} w_{ij}]^2 / \sum_{j\in S(i)} w_{ij}^2\}$ is the *effective sample size* in State i. NASS believes that the simple Bernoulli model is appropriate in this context. It ignores the effects of stratification and cluster sampling on variance and assumes that the only role sample weighting plays is in increasing variance and decreasing effective sample size. Nevertheless, weights are used in determining z_i as protection against model failure.

Jiang *et al.* assumes that the D_i can be determined reliably with randomization-based methods. That is not my experience. If sample sizes are not large enough to estimate π_i directly, then estimates of the variance of the direct estimator are even more suspect. Still, nothing is lost if Jiang's randomization-based D_i is replaced by my model-based one.

Suppose we have M "states" for which we need estimates (recall that some states are collapsed together for this purpose). To draw strength from the other states, NASS assumes that each z_i can itself be modeled:

$$\begin{aligned} z_i \;=\; &\pi_i \quad + \quad e_i \\ &\pi + v_i \quad + \quad e_i, \end{aligned} \tag{1}$$

where $E(v_i) = E(e_i) = 0$, $\mathrm{Var}(v_i) = A$, and $\mathrm{Var}(e_i) = D_i = \pi_i\,(1 \quad \pi_i)/n_i^*$.

Consider the estimator

$$\begin{aligned} z_i^{(\gamma)} &= (1 \quad \gamma_i)z_i \;+ \gamma_i \textstyle\sum^M w_k z_k / \sum^M w_k\,, \\ &= (1 \quad \gamma_i)z_i \;+ \gamma_i\, z \end{aligned} \tag{2}$$

where $\gamma_i \approx D_i/(A + D_i)$, w_k is the sum of the sampling weights within S(k), and z is the randomization-based estimator for the fraction of black-run operations among the farms missing from the Census list *nationally*. Since D_i approaches 0 as the effective sample size in i becomes arbitrarily large, $z_i^{(\gamma)}$ is randomization consistent whenever z_i is.

The estimator, $z_i^{(\gamma)}$, is not quite optimal under the component-of-variance model in equation (1). Many (Lahiri and his team included) would not weight the z_k by w_k. Moreover, $z_i^{(\gamma)}$ ignores the variance of z and the covariance between z_i and z. Nevertheless, we will assume for simplicity that M is large enough that such issues hardly matter. More important is the requirement that we estimate A and the D_i before $z_i^{(\gamma)}$ can be operationalized.

To estimate D_i, we need to estimate π_i first, but that is precisely the goal of the entire exercise. It is common to estimate D_i using z_i in place of π_i. Indeed, that is what NASS has been doing for the most part. Note, however, that when $z_i = 0$, D_i must also be zero. This is suspect. Just because we find no black-run operations that not on the Census list in a state area sample does not mean there are no back-run operations missing from the state's Census list anywhere. To avoid this silliness, NASS has been setting an arbitrary lower bound on its D_i estimate. Nevertheless, the agency's calculation of γ_i remains dependent on a very rickety estimator for π_i.

The variance component A can be estimated using the methods of moments:

$$a^* = [\ \sum^M z_i^{\ 2} \quad (\ \sum^M z_i)^2/M]/(M \quad 1) \quad [\ \sum^M z_i\,(1 \quad z_i)/n_i^*]/M\ .$$

This formula can produce negative estimates. It is therefore popular to estimate A with

$$a = \max\{0, a^*\}.$$

NASS uses something a bit different:

$$a_{NASS} = \max\{a^*, (1/2)[\ \sum^M z_i^{\ 2} \quad (\ \sum^M z_i)^2/M]/(M \quad 1)\}.)$$

Although NASS does not think it has the sample sizes to estimate the π_i directly. It does think that it can estimate directly the fraction of black-operated farms among the farms missing from the Census list *nationally* with z. Consequently, if z_i^* denotes NASS's final estimator for π_i, it desires the z_i^* satisfy

$$\sum^M w_i z_i^*/\ \sum^M w_i \ = \ z\ .$$

This is the *bookkeeping constraint*.

The Arcsine-root Transform

Jiang *et al.* hits upon a clever way to remove the dependence of the D_i on π_i. Instead of applying the components-of-variance model in equation (1) to the z_i, he applies it to a transform of the z_i:

$$y_i = 2\sin^{-1}(\downarrow z_i).$$

(I added a factor of 2 to the transform. It does not effectively change anything, but it makes the arithmetic a bit cleaner.)

One can show that $\mathrm{Var}(y_i) \approx 1/n_i^*$, which is invariant to π_i! Thus, if we replace the z_i in equation (1) by the y_i, the D_i become (nearly) $1/n_i^*$, and the need for early estimates of the π_i is avoided. If NASS were to follow this suggestion, then it would not have to set an arbitrary lower bound on the D_i.

The problem with this transformation is that in invoking it one needs to assume the D_i and A are small. Otherwise, one could not go forward and backward between the original and arcsine-root spaces and preserve near unbiasedness (in particular, the back-transformed solution may not even be unconditionally unbiased for π). If A *is* small, however, then

$$\begin{aligned} D_i &= \pi_i\,(1 \quad \pi_i)/n^* = (\pi + v_i)(1 \quad \pi \quad v_i)/n_i^* \\ &\approx \pi(1 \quad \pi)/n^* \end{aligned}$$

where the near equality gets better when we take expectations.

This suggests that instead substituting z_i for π_i in $D_i = \pi_i (1 \quad \pi_i)/n_i^*$, NASS begin with $d_i^{(1)} = z(1 \quad z)/n_i^*$, because z is much more stable than z_i, never zero in practice, and fairly close to π_i. The agency can calculate a set of γ_i based on the $d_i^{(1)}$ (and an estimator for A), and then use the computed $z_i^{(\gamma)}$ from equation (2) within $d_i^{(2)} = z_i^{(\gamma)}(1 \quad z_i^{(\gamma)})/n_i^*$. This leads to an iterative process that will likely converge fairly quickly. It should be noted, however, that the Lahiri team's arcsine-root transformation removes the need for iteration.

NASS's Single Covariate and the Bookkeeping Constraint

Unlike in the Lahiri team's formulation, NASS uses a single covariate and no intercept. Instead of the model in equation (1), NASS bases its small-domains estimation on

$$z_i \;=\; c_i\,(\mu + v_i) + \;e_i,$$

where c_i is the fraction of farms on the Census list in State i that have black operators. How to incorporate this type of information in arcsine-root space is not a trivial question.

Although more intuitively appealing than a model at least half in arcsine-root space (it is unclear whether the c_i should also be so transformed), the model NASS uses has its own conceptual drawback. It is not symmetric in that the model for $1 \quad z_i$ is not linear in $1 \quad c_i$. For fractions like black-operated farms, it is clear that NASS wants to look at z_i rather than $1 \quad z_i$ because it is much smaller. For other fractions, like the fraction of farms with hog production, that is not so straightforward. Indeed, whether z_i or $1 \quad z_i$ is smaller depends upon the state.

Often a simple ratio adjustment is used to enforce the bookkeeping constraint. That is to say, each near-optimal $z_i^{(\gamma)}$ is multiplied by the common factor necessary for the constraint to hold. NASS, however, has been incorporating the constraint directly into the optimality requirement. Rather than minimizing the mean squared error of each state separately. NASS attempts to minimize the weighted sum of the state mean squared errors under the bookkeeping constraint. This would be nearly impossible to do in arcsine-root space.

Variance Estimation

Perhaps the Lahiri team's single largest contribution is in the area of variance estimation, where they propose two simple jackknives to adjust for the asymptotic biases of the conventional variance estimator for the optimal $z_i^{(\gamma)}$, $v_i = D_i\, a/(a + D_i)$, where a is a method-of-moments estimator for A, and D_i is known.

Even if we accept my model-based formulation, D_i is not known, except approximately in arcsine-root space. More to the point, because of the restriction on a (or, more precisely, a_{NASS}) and the bookkeeping constraint, NASS will not be using a near-optimal $z_i^{(\gamma)}$, although it's estimator can still be put in the form of equation (2) (which I will continue to call $z_i^{(\gamma)}$, without, I hope, undo confusion). With this in mind, it is not clear to me that treating the γ_i NASS uses as fixed in the variance formula:

$$v' = (1 \quad \gamma_i)^2\, z_i^{(\gamma)}(1 \quad z_i^{(\gamma)})/n_i^* + \gamma_i^2\, a$$

is inappropriate when M is large. When M is less than large, an analogous formula can be derived.

This does not render the Lahiri team's variance formula useless, however. It can still be calculated to estimate just what NASS loses by not using the optimal $z_i^{(\gamma)}$.

What About Confidence Intervals?

Wald confidence intervals for proportions can extend below 0 and beyond 1. Jiang *et al.* point out that the arcsine-root transform appears to be a cure for that.

An alternative is to extend Wilson's (1927) method; e.g., compute a 95% confidence interval by solving the following for π_i :

$$\frac{|z_i^{(\gamma)} - \pi_i|}{\{v'\pi_i(1 - \pi_i)/[z_i^{(\gamma)}(1 - z_i^{(\gamma)})]\}^{1/2}} < 1.96\ ,$$

Squaring both sides leads to a easily solvable second degree polynomial in π_i, which can be converted into a asymmetric confidence interval around z_i*.

Concluding Remarks

Although NASS will likely not use the jackknives proposed by Lahiri and his collaborators, I find them extremely useful in principle and remarkably intuitive (where what they extend was not). NASS will also likely not use the arcsine-root transformation. Nevertheless, I think it is fair to say that the exercise of studying what the Lahiri team had done will sharpen what NASS finally does.

Reference

Jiang, J.P., Lahiri, P., Wan S.-M, and Wu, C.-H., Jackknifing in the Fay-Herriot Model with an Example. [earlier in the Proceedings]

Wilson, E.B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, **22**, 209, 212.

Session 6

Future of the Funding Opportunities in Survey Research

The Immediate Future of the Funding Opportunity in Survey and Statistical Methodology

Monroe G. Sirken
National Center for Health Statistics

Introduction

There are two distinct periods in the future of the Funding Opportunity in Survey and Statistical Methodology. My remarks deal with the immediate future - the period of time remaining in the existing arrangement between the National Science Foundation (NSF) and 12 Federal agencies of the Interagency Council on Statistical Policy (ICSP) to jointly support the Funding Opportunity. Also, I'll briefly discuss benefits of the Funding Opportunity already realized and others yet to be realized. Nancy Kirkendall will discuss some of the options for continuing the Funding Opportunity after expiration of the existing NSF/ ICSP arrangement.

The Immediate Future Of the Funding Opportunity

Table 1 is essentially a roadmap of the Funding Opportunity during the entire period covered by the NSF/ICSP funding arrangement. The table heading shows three funding years, namely, 1999, 2001, and 2002, and the table stub lists four milestones in each funding year cycle:

1) NSF Funding Opportunity announcement issued;
2) Two-tiered project review and selection process conducted by NSF and Federal agency panels;
3) Seminars, such as this one, convened to review and discuss findings of funded research projects;
4) Funded projects are closed out.

The 12 cells in Table 1 are labeled by the calendar years in which the milestones occur. The length of a single funding year cycle, from issuance of the NSF announcement to close-out of funded projects, is about 5 years. The combined length of the three funded years cycles, from issuance of the NSF announcement in the first funding year to close-out of projects in the third funding year, spans an 8-year period, 1998- 2005. The Funding Opportunity is now about midway into this 8-year period. Milestones represented by the six cells above the jagged line have occurred or are in process of occurring. Milestones represented by the six cells below the jagged line is the roadmap of immediate future of the Funding Opportunity. If we stay on course, research projects being funded this year and those funded next year, respectively, will be presented at seminars, like this one, in years 2003 and 2005. These seminars will provide opportunities for direct discourse between the principal investigators of funded research projects and the statistical staff of Federal agencies.

The outlook for the immediate future of the Funding Opportunity is quite bright. NSF and Federal agency funding is confirmed for the current funding year, 2001, and seems most likely to be forthcoming next year, 2002, the last funding year of the current NSF/ICSP arrangement.

Benefits of the Funding Opportunity

Establishing, developing and successfully testing the mechanism that runs the Funding Opportunity is, I believe, the outstanding achievement so far. The mechanism has two unusual features:

a) a consortium of 12 Federal agencies in the ICSP and the NSF collaborates in supporting the Funding Opportunity; and

b) the FCSM serves as the liaison between the ICSP and the NSF in administering the Funding Opportunity.

The first listed feature is unusual because the agencies in the consortium are pooling their resources in order to collectively sustain a basic survey and statistical research program that benefits the entire Federal statistical system. This kind of interagency coordination and collaboration is particularly remarkable because it is occurring in a decentralized Federal statistical system in which short-term demands of individual agencies are the norm.

The second listed feature is unusual because, for the first time, the FCSM is actively engaged in fostering and coordinating an interagency collaborative research program.. This new FCSM activity represents an exciting extension of FCSM's normal activities involving preparation and distribution of statistical policy working papers and sponsorship of research and statistical policy seminars.

It is fair to say that the Funding Opportunity would never have happened except for the efforts of two individuals. Cheryl Eavey, Chief of NSF Methodology, Measurement, and Statistics Program got the ball rolling with her offer to the ICSP to underwrite half of the funding for the Funding Opportunity during a three year period if Federal agencies agreed to provide matching funds. Katherine Wallman, Chief Statistician, Chair of the ICSP, and head of the OMB Office of Statistical Policy that sponsors the FCSM, got the job done by lending her strong support and astute leadership to the cause.

Concluding comments

Ultimately, the success the Funding Opportunity will be judged by the impact that the research it supports will have on the programs of Federal statistical agencies. Though far too early to make that judgement, the quality of the papers and discussions presented at this Seminar are encouraging signs.

TABLE 1

ROADMAP OF THE FUNDING OPPORTUNITY IN SURVEY AND STATISTICAL METHDOLOGY

	FUNDING YEAR		
MILESTONES	**1999**	**2001**	**2002**
NSF announcement (closing date)	1998	2000	2000
Project review & selection	1999	2001	2002
Seminar presentation	2001	2003	2005
Project close-out	2002	2004	2005

Future of the Funding Opportunity In Survey Research

Nancy J. Kirkendall, Energy Information Administration

To talk about the future, let me set the background with a little detail about our current operations. In 1998, the heads of the 13 largest statistical agencies, members of the Interagency Council on Statistical Policy (ICSP) agreed to participate in this jointly funded research program for 3 years. In essence, the financial agreement was that each agency would provide $25K (budgets permitting) toward the funding opportunity, and the National Science Foundation (NSF) matched the aggregate contribution providing a total of about $600K. However, in addition agencies can, and do, decide to contribute to the support of individual proposals that are of particular interest to them. In the first year, only the Census Bureau contributed extra funding for specific projects. In the second year, three agencies contributed extra funding. This is significant, because it increases the pool of money available to fund proposals, thereby encouraging researchers to apply.

The $25K contribution essentially buys agency participation in the government panel. This funding opportunity uses two panels of experts. The first is the usual peer review panel of experts selected by NSF. This panel reviews the proposals to determine whether they have scientific merit. Those that are judged to have scientific merit are further classified as to whether they are high, medium, or low importance. The government panel of experts reviews the proposals that have scientific merit, as judged by the NSF panel. The government panel is particularly focused on deciding which proposals have the greatest potential to provide results useful to Federal statistical agencies. These are also classified as to whether they have high, medium, or low potential for providing useful results. The opinions of both panels are used to determine the winning proposals. NSF makes the final determination, in consultation with representatives of the Federal Committee on Statistical Methodology (FCSM), another partner in this research opportunity.

Bad News and Good News.

NSF has decided that in the future we will not have a separate proposal for Research in Survey and Statistical Methodology, at least in part because of the time and effort to manage a separate program. However, NSF has agreed that the statistical agencies can participate in the regular Methodology, Measurement and Statistics program, and we can add a summary statement of our needs to the description of that funding opportunity. On the positive side, this gives proposals of interest to the statistical agencies access to a larger amount of NSF money (if they are judged to have high importance by the scientific panel.) Agencies will still be able to contribute funds, both as a consortium and to support research of interest to them. The government panel can continue to operate in the same way and review the proposals that NSF's expert panel judges to have scientific merit.

The new approach may have an added advantage of providing Statistical Agencies more control over how their money is spent. The question we need to answer for the future is how best to coordinate the activities of the statistical agencies. Here are some options.

Option 1a: Agencies each pay X dollars to "buy a seat" at the table of government experts. X will be fairly small, because some of the agencies are small. In addition, agencies will be encouraged to authorize their expert to identify particular proposals of special interest to them (if any) and to contribute additional funding to those proposals.

Pro: This helps to assure that the government experts review proposals from a Federal statistical system point of view.

The government panel could actually make the decision as to which proposals to fund for the benefit of the statistical system. (In the current model, the government panel advice is advisory.)

Agencies can also decide to contribute to projects of special interest.

Con: Even though X is small, smaller agencies object to paying for research that they do not think benefits them.

Option 1b: Similar to option 1a except large agencies pay X dollars, and small agencies pay Y dollars (X greater than Y) to "buy a seat" at the table of government experts. (Same pros and cons).

Option 2: Statistical agencies convene a government expert panel, with no requirement for "up-front-funding". The government experts will be authorized by their agency to identify particular proposals of special interest to them and to contribute additional funding to those proposals.

Pro: Agencies only pay for research that they think benefits them.

Government panel facilitates agency access to research proposals.

Individual agencies can decide to jointly fund specific proposals.

Con: Agencies do not necessarily act for the common good of the statistical system.

These options are the ones I have thought of, and I would be very interested to hear comments on them, as well as any additional ideas anyone may have. I would also be interested in expanding the list of pros and cons for these options. One shortcoming of all options presented above is that none of them addresses the issue of how to make the program more broadly accessible to smaller statistical agencies, those that are not represented on the ICSP.

Our challenge over the next year is to come up with a specific proposal to take to the Interagency Council on Statistical Policy. I think the leaders of the Statistical Agencies generally like the program we have now; or, at least, they like the idea of jointly sponsoring research. The grumbling I have heard is from smaller agencies with limited budgets. Part of the reason for our current common research program is to provide the smaller agencies with access to research by leveraging our common interests. The larger agencies already have mechanisms in place to sponsor research. As a result, they do not feel that they necessarily gain as much from the collaboration. Although, one might say that the government panel provides a convenient way to identify specific proposals that could benefit their programs.

Reports Available in the Statistical Policy Working Paper Series

1. *Report on Statistics for Allocation of Funds*, 1978 (NTIS PB86-211521/AS)
2. *Report on Statistical Disclosure and Disclosure-Avoidance Techniques*, 1978 (NTIS PB86-211539/AS)
3. *An Error Profile: Employment as Measured by the Current Population Survey*, 1978 (NTIS PB86-214269/AS)
4. *Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics*, 1978 (NTIS PB86-211547/AS)
5. *Report on Exact and Statistical Matching Techniques*, 1980 (NTIS PB86-215829/AS)
6. *Report on Statistical Uses of Administrative Records*, 1980 (NTIS PB86-214285/AS)
7. *An Interagency Review of Time-Series Revision Policies*, 1982 (NTIS PB86-232451/AS)
8. *Statistical Interagency Agreements*, 1982 (NTIS PB86-230570/AS)
9. *Contracting for Surveys*, 1983 (NTIS PB83-233148)
10. *Approaches to Developing Questionnaires*, 1983 (NTIS PB84-105055)
11. *A Review of Industry Coding Systems*, 1984 (NTIS PB84-135276)
12. *The Role of Telephone Data Collection in Federal Statistics*, 1984 (NTIS PB85-105971)
13. *Federal Longitudinal Surveys*, 1986 (NTIS PB86-139730)
14. *Workshop on Statistical Uses of Microcomputers in Federal Agencies*, 1987 (NTIS PB87-166393)
15. *Quality in Establishment Surveys*, 1988 (NTIS PB88-232921)
16. *A Comparative Study of Reporting Units in Selected Employer Data Systems*, 1990 (NTIS PB90-205238)
17. *Survey Coverage*, 1990 (NTIS PB90-205246)
18. *Data Editing in Federal Statistical Agencies*, 1990 (NTIS PB90-205253)
19. *Computer Assisted Survey Information Collection*, 1990 (NTIS PB90-205261)
20. *Seminar on Quality of Federal Data*, 1991 (NTIS PB91-142414)
21. *Indirect Estimators in Federal Programs*, 1993 (NTIS PB93-209294)
22. *Report on Statistical Disclosure Limitation Methodology*, 1994 (NTIS PB94-165305)
23. *Seminar on New Directions in Statistical Methodology*, 1995 (NTIS PB95-182978)
24. *Electronic Dissemination of Statistical Data*, 1995 (NTIS PB96-121629)
25. *Data Editing Workshop and Exposition*, 1996 (NTIS PB97-104624)
26. *Seminar on Statistical Methodology in the Public Service*, 1997 (NTIS PB97-162580)
27. *Training for the Future: Addressing Tomorrow's Survey Tasks*, 1998 (NTIS PB99-102576)
28. *Seminar on Interagency Coordination and Cooperation*, 1999 (NTIS PB99-132029)
29. *Federal Committee on Statistical Methodology Research Conference (Conference Papers)*, 1999 (NTIS PB99-166795)
30. *1999 Federal Committee on Statistical Methodology Research Conference: Complete Proceedings*, 2000 (NTIS PB2000-105886)
31. *Measuring and Reporting Sources of Error in Surveys*, 2001 (NTIS PB2001-104329)
32. *Seminar on Integrating Federal Statistical Information and Processes*, 2001 (NTIS PB2001-104626)
33. *Seminar on the Funding Opportunity in Survey Research*, 2001 (NTIS PB2001-108851)

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161; telephone: 1-800-553-6847. The Statistical Policy Working Paper series is also available electronically from FCSM's web site <*http://www.fcsm.gov*>.