National Science Foundation

Recurring Surveys: Issues and Opportunities

A Report to the National Science Foundation
Based on a Workshop Held on March 28-29, 2003

Report Prepared by
Roger Tourangeau

$$\hat{y}_t = \mu + y_{t-1} + \theta(y_{t-1} - y_{t-2})$$
$$\hat{y}_t = \mu + y_{t-1} - \theta e_{t-1}$$

sbe social, behavioral, and economic sciences

Cover Images:

The photo in the lower right-hand corner illustrates the testing and evaluation of the Census 2000 form in 1995. It is available off the web site of the U.S. Bureau of the Census at <http://www.census.gov/pubinfo/www/photos/img/highres/Histo07h.jpg>.

The image in the upper left-hand corner was captured from the web site of New School University < http://www.newschool.edu/>. The drawing was scanned from J.S. Gibbons (1859) *The Banks of New York, Their Dealers, The Clearing House and the Panic of 1857*. New York: Appleton & Co.

Recurring Surveys: Issues and Opportunities

A Report to the National Science Foundation
Based on a Workshop Held on March 28-29, 2003

Report Prepared by Roger Tourangeau

Executive Summary

Recurring surveys, including the "infrastructure surveys" funded by the National Science Foundation (NSF), have a unique role to play in tracking social, demographic, political, and economic trends over time and have made important contributions to every social science. This report summarizes a two-day workshop that was held on March 28-29, 2003 in Arlington, VA. The workshop examined NSF's current group of infrastructure surveys (the American National Election Studies, the General Social Survey, and the Panel Survey on Income Dynamics) and the broader issues raised by surveys done over time.

Recurring surveys share the problems of other surveys and face some unique challenges as well. Prominent among the unique challenges is balancing the need to maintain comparability over time (facilitating the measurement of change) with the need to introduce innovation (in order to maintain quality). Recurring surveys can enhance their value and introduce new features without sacrificing comparability in a variety of ways — by adding linked geospatial or administrative data to the core survey data, permitting international comparisons, collecting contextual data, adding linked samples, embedding experiments in surveys, or supplementing survey data with qualitative data from open-ended items.

Most recurring surveys are either repeated cross-sectional surveys (consisting of separate samples interviewed at different points in time) or longitudinal surveys (consisting of a single sample interviewed at several points in time), although various hybrids are possible. Both designs have their methodological strengths and weaknesses. Panel samples may become less representative over time as the population changes and as panel members drop out; panels may also be prone to certain forms of measurement error, such as conditioning and seam bias. Cross-sectional surveys may also be prone to certain measurement problems (such as telescoping, or misdating events). Although both designs are extremely useful for tracking trends, panels have the added advantage of allowing individual change to be measured and sophisticated causal models to be fit. Both types of design permit relatively low-cost enhancements, such as linked samples and the addition of administrative records data or geodata.

Over the last ten years or so, recurring surveys, like other surveys, have confronted the twin problems of rising costs and falling response rates. To date, there is little evidence that the decline in response rates has produced large biases in survey estimates, but the consensus at the workshop was that the causes and effects of nonresponse in surveys are extremely high priority topics for methodological research. These are important industry-wide problems that could have significant spillover effects on social science. In addition, investigation of new methods of data collection and measurement error remain cutting edge topics for research.

The final session of the workshop identified areas for NSF research support that are likely to yield large returns. These fell into three categories — methodological work, enhancements to the existing infrastructure surveys, and new initiatives. The high priority methods research included work on nonresponse, measurement error, and modes of data collection. Participants recommended a large number of enhancements to the infrastructure surveys, including full funding of the surveys by NSF, increased sample sizes, improved coverage, and greater coordination across the surveys. Several participants also advocated restructuring the governance of the infrastructure surveys at NSF. Finally, the workshop identified a number of new initiatives worth funding, including several new panel surveys.

Table of Contents

Introduction

Every era presents new challenges and opportunities to behavioral and social science researchers. Since sample surveys are a key tool for the social sciences, when new challenges arise to their design or execution, these challenges affect the social sciences more broadly. In fact, the recurring surveys funded by the National Science Foundation and others have made important contributions to every social science and have enabled researchers to address important theoretical and practical questions that could not have been addressed in any other way. In the last few years, rising survey costs and falling survey response rates have presented a formidable practical challenge to survey researchers in the United States and elsewhere (Atrostic, Bates, Burt, and Silberstein, 2001; Groves and Couper, 1998; de Leeuw and de Heer, 2001). The National Science Foundation has substantial investments in ongoing "infrastructure" survey programs, including the Panel Survey of Income Dynamics, the American National Election Studies, and the General Social Survey. In addition, NSF funds support many one-time surveys conducted as part of individual grant projects. The "infrastructure" surveys funded by the Directorate for Social, Behavioral, and Economic Sciences raise their own unique issues, related to their history and their designs as recurring, or over time, surveys. Recurring surveys, including those funded by NSF, have a unique role to play in tracking social, demographic, political, and economic trends over time.

This report summarizes a two-day workshop that examined NSF's current group of infrastructure surveys and the broader issues raised by surveys done over time. The workshop, held March 28 and 29, 2003, is one in a series of workshops that will explore some of the challenges and opportunities for investment by NSF in both existing and new survey programs.

NSF currently funds two types of recurring surveys. The first type — repeated cross-sectional surveys — consist of surveys that interview individual respondents only once, but field similar surveys (with the same basic sample design, a core set of questions administered each time, and the same method of data collection) on some periodic time schedule. For example, over most of its history, the General Social Survey (GSS) has conducted yearly surveys that administer a core set of items to separate annual samples selected via area probability methods. The second type of over-time survey — longitudinal or panel surveys — involves interviewing members of the same sample repeatedly. The Panel Survey on Income Dynamics (PSID) has followed this second model. Both types of survey make it possible to monitor time trends, and various hybrid designs (such as rotation group designs) are possible that combine features of repeated cross-section and panel designs. NSF's third infrastructure survey, the American National Election Studies (ANES), involves repeated cross-sectional surveys (fielded every four years) and two-wave panel surveys (typically before and after a given election). All three of the existing infrastructure surveys have been in existence for more than thirty years.

Recurring surveys face some common issues. First, there is the need to maintain comparability of survey estimates over time, a requirement that is more difficult to meet in the face of rising costs and falling response rates. Surveys must spend more to achieve the same or lower response rates than in previous years. In addition, there are the special measurement problems in both repeated cross-section and panel designs. For example, "core" items administered every time tend to accumulate and may need to be pared back to keep the survey to a manageable length. Or specific items may need to be updated to remain valid measures. Finally, innovation in measurement design must be balanced by the desire to maintain the time series.

The workshop brought together 23 distinguished outside researchers from a number of disciplines — including demography, economics, political science, psychology, sociology, and survey research — with key NSF staff to address these and related issues for NSF's infrastructure surveys. (Appendix A provides a list of the participants.) More specifically, the workshop focused on five issues relevant to NSF's long-term planning:

1) The role of surveys in constructing social knowledge;
2) Special issues in over time designs;
3) Rising costs and declining response rates;
4) New areas and new modes of survey research; and
5) High payoff investments for NSF.

Each of the researchers prepared a short statement summarizing his or her views on one of these issues. These statements are included in Appendix B.

The Role of Recurring Surveys in Social Research

The first session of the workshop dealt with the place of recurring surveys in social research. More specifically, this session addressed three major questions:

- the relation between quantitative and qualitative investigations of social phenomena;
- some of the drawbacks of recurring surveys relative to one-time surveys;
- the feasibility and desirability of hybrid approaches that combine qualitative with quantitative data collection, surveys with experiments, or self-report data with data from external sources, such as administrative records.

Qualitative versus quantitative approaches. As several participants noted, the scientific questions motivating the research ought to drive the research design. Still, the availability of data collected under a specific design may also shape the research questions as well; the capability often creates the research opportunities. Even when a researcher designs a study from scratch, there is almost

always some interplay between the design of the research and the development of the hypotheses the research is intended to answer.

The great strength of the quantitative approach is that, when coupled with probability sampling, it allows generalization to some larger population with a known statistical margin of error. Most qualitative investigations do not permit either straightforward generalization of the results or the assessment of the likely error. In terms of Kish's (1987) three criteria for evaluating research designs (representation, randomization, and realism), quantitative investigations, particularly surveys, are strong on *representation*. Both qualitative and quantitative investigators make the same sorts of claims about the social world and thus face the same types of epistemological challenges, among them the challenge to make valid empirical generalizations. The goal for NSF should be to foster methods that combine the depth and detail of qualitative methods with the breadth and generalizability of sample surveys. Traditionally, surveys have incorporated in-depth interviews, protocol analysis, and other qualitative techniques only in the preliminary stages of developing the questionnaires and other procedures to be employed later in essentially quantitative surveys; the coming years are likely to see an increasing use of qualitative information from the beginning to the end of surveys. Such mixed designs will have to overcome several obstacles, including the heightened disclosure risk in surveys that include lengthy open-ended responses (or other qualitative data) and the cost and difficulty of coding such material. Automated data reduction of rich verbal material remains out of reach.

The theme of creating hybrid designs and data bases that include multiple types of information (such as qualitative and quantitative data), data from different modes of collection, and data from multiple sources (such as self-report data from survey respondents, geodata, and data from administrative record systems) recurred throughout the workshop.

Some shared dangers of cross-sectional and panel designs. A potential danger with recurring surveys — whether they are repeated cross-sections or panel designs — is their inherent tendency toward methodological conservatism. If the goal is to measure change, then it is obviously important not to change the measures. All three of the current infrastructure surveys have had to face the difficult challenge of balancing the need for continuity against the need for innovation. In any survey, much of the instrumentation is likely to reflect the state-of-the-art of the substantive discipline and the conventional wisdom that the measurement error is not a serious problem for these state-of-the-art measures. After all, the measures will have "worked" well in the past. Determining whether this assumption is warranted often involves disrupting the standard measurement procedures for some portion of the sample, procedures that most of the researchers involved probably believe in. Although designs are available for estimating the impact of changes in methodology on the estimates from recurring surveys, the subject matter specialists may be unwilling to go to the trouble and expense of such "calibration experiments"

that compare old and new instrumentation for some period of time and allow the two portions of the time series to be spliced together.

Although all three of NSF's current infrastructure surveys have a good track record for carrying out innovative methodological research, recurring surveys face the continuing challenge of keeping the procedures both fresh and comparable. Some of the tactics for striking the right balance between innovation and continuity involve experiments conducted off-line (for example, in pilots for the main study, a tactic often used by the ANES), setting aside portions of the main sample for experimentation ("methods panels"), embedding experiments (e.g., question wording or order experiments) in the main survey, and incorporating various add-ons (e.g., linked samples) to the main data collection. When major changes in the basic methods of a recurring survey are implemented, calibration studies should be done in which parallel samples receive the old and the new methods and the impact of the switch can be determined. The calibration should continue past the "seam" when the new method becomes the "official" method, since the new method may have somewhat different effects when it is adopted in the main production survey.

Enriching recurring surveys: Hybrid designs. Another trend over the last few decades has been the addition of various enhancements to enrich the data from recurring surveys at relatively low marginal costs. The enhancements include:

- administrative records or geospatial data linked to the survey data;
- international comparisons across parallel surveys;
- contextual data relevant to the survey topic (such as data on employers in a labor force survey, on candidates and campaigns in a survey on elections, on organizations or social networks in a survey on social attitudes);
- experiments embedded in surveys (such as question wording experiments or vignettes designed to illuminate the respondents' understanding of some issue or phenomenon);
- open-ended items or in-depth interviews to supplement the survey data.

These added types of data often bring new methodological challenges. For instance, advances in computer technology have made the linking of large data bases feasible and the costs of creating and analyzing linked survey-administrative data bases is falling, but the confidentiality issues involved in linking records and survey data are formidable. Solutions to these problems are just beginning to appear in the scientific literature. But the trend is clear — surveys will continue to supplement the data collected from respondents with data from outside sources. Similarly, data that used to be collected for methodological purposes (such as data from split-ballot experiments or in-depth interviews) are likely to be collected in the future for substantive reasons as well.

Special Methodological Issues for Recurring Surveys

Most of the new recurring surveys mounted over the last 25 years or so (e.g., the National Longitudinal Surveys of Youth, the Survey of Income and Program Participation, the Health and Retirement Surveys, the Adolescent Health Survey, and the National Survey of Families and Health) have been panel surveys, raising the issue of the unique virtues of repeated cross-sectional surveys. Several participants examined the relative strengths and weaknesses of the two types of surveys. Table 1 provides a summary of the key points made in these discussions.

Nonobservation errors. The top panel of the table focuses on differences in errors of "nonobservation" (i.e., sampling, coverage, and nonresponse error; see Groves 1989). First, repeated cross-sections tend to provide a more accurate picture of changing populations than panel surveys do. Unlike panel surveys, which have to incorporate supplemental samples to represent new population segments, repeated cross-sectional surveys capture new entrants to the population (via births or immigration) in a straightforward way. As a result, cross-sectional samples may be preferred for estimates regarding a single point in time. In fact, high-quality repeated cross-sectional surveys like the March Current Population Survey and the General Social Survey are often used as benchmarks to which other surveys are compared to assess their representativeness or the overall accuracy of their results. Second, the burdens that panel surveys impose on the respondents may produce diminishing representativeness over time, as panel attrition erodes the quality of the original sample. In addition, the need to track members of panels to new locations may require redefining the population of interest (for example, to exclude panel members who move into institutional settings), which may produce further losses of representativeness. Still, there is a bias-variance tradeoff here. The problems of attrition and increasingly poor coverage of the population over time introduce bias into the estimates derived from panel surveys; at the same time, the estimates of aggregate change from panel surveys are generally more precise than the corresponding estimates from repeated cross-sections. Panels can also produce more precise estimates for single points in time as well.

Another potential advantage of repeated cross-sections is that the samples often can be aggregated over time, enabling researchers to study rare populations. For example, researchers have aggregated successive GSS samples to study the gay and lesbian populations (Black et al., 2000). Any single GSS sample contained too few members of these populations to permit analysis, but by combining samples across years, the researchers were able to find enough cases. The sample sizes for panels are fixed from the outset; as a result, panels often oversample various subpopulations to allow for subgroup analyses involving rare subgroups. Such oversampling reduces the efficiency of overall estimates and increases the complexity of the designs. Finally, both types of design encourage the addition of linked samples and data collections to supplement the core sample and questionnaires. The respondents can be used to

generate linked samples of organizations, employers, religious congregations, and so on. With longitudinal surveys the relatively low marginal costs of adding data from external sources (such as employers, medical care providers, insurers, school transcripts, or other administrative records) can motivate such supplemental data collections. The added data may have the disadvantage of raising the start-up costs for outside analysts by increasing the complexity of the resulting data set.

Observational errors. Repeated cross-sections are designed to measure aggregate change over time; panel data can be used for that purpose as well. Of course, the great virtue of panel surveys is that they allow the measurement of change at the individual level, thus permitting the application of much more sophisticated (and more defensible) causal models than can be applied to cross-sectional data. This has been of particular importance in tests of person-level theories, which use the temporal variation of micro-level measures to assist in causal inference. It is these features of longitudinal designs that probably account for their popularity over the last 25 years. Still, repeated cross-sectional surveys may have a unique contribution to make in studying trends. Depending on the timing and spacing of data collection, repeated cross-sections can allow for monitoring the impact of unexpected events; such effects may be difficult or impossible to capture in a panel survey. For example, some tracking studies (consisting of closely spaced repeated cross-sections) collect data as often as once a month. To reduce the burden on panel members, longitudinal surveys rarely collect data this often and, given the size and complexity of most longitudinal surveys, it is generally impossible to alter their schedules of data collection quickly in response to unexpected events. As a result, though, the short-term impact of major events may be difficult to detect and it may be impossible to trace the pulse and decay of their effects. (For an example of a repeated cross-section used to trace the impact over time of a major but unexpected event, see Stewart's write-up in Appendix B.)

Panel surveys allow researchers to fit models that attempt to separate measurement error from true change. In general, special studies have to be undertaken to assess the level of measurement error in cross-sectional surveys. Both types of design are particularly susceptible to certain types of measurement errors. For example, repeated cross-sectional surveys are prone to telescoping errors (in which events are overreported because respondents forget when they occurred). Longitudinal surveys can build-in various memory aids, such as summaries of information provided in prior interviews. On the other hand, longitudinal surveys impose greater burden on respondents and are prone to conditioning and similar time-in-sample effects. As with attrition, the impact of conditioning varies from item to item so it is difficult to measure its effects, let alone compensate for them. Still, several participants questioned whether conditioning and attrition effects are important sources of error relative to other problems (such as poorly specified constructs and omitted variables).

One final advantage of repeated cross-sectional designs is their ability to allow replication of key analyses on independent samples. They also allow for multi-level models, in which long-standing relationships (e.g., between party affiliation and vote choice) may vary across different time periods (e.g., in different elections or different eras).

Costs and timeliness. The costs of a survey reflect many design features, but as a rule, the initial wave of a longitudinal survey tends to be larger and more expensive than a repeated cross-sectional design. As a result, they may require longer lead time to design and get into the field. Because both types of designs encourage add-ons, they may permit a wide range of analyses; at the same time, however, their complexity may discourage new users of the data, who will incur a large entry cost in wading through the documentation and understanding the designs. The complexity of the design as well as the added intricacies imposed by wave nonresponse and item missing data may prevent most users from taking full advantage of panel data. As one workshop participant noted, relatively few analyses are truly longitudinal.

Table 1. Pros and Cons of Repeated Cross-Sectional and Longitudinal Designs

	Repeated Cross-Sectional Design	**Longitudinal Design**
Nonobservation Errors (e.g, coverage, nonresponse, sampling errors)	▪ Better representation of changing populations ▪ Less bias for measuring aggregate trends, but reduced precision ▪ High quality cross-sectional samples can be used to benchmark other surveys ▪ Samples can be aggregated over time for studies of rare populations (unless there are large changes over time) ▪ Can be used to select linked samples (of organizations, congregations, etc.)	▪ Refreshment/supplemental samples needed to maintain representativeness of panel to current population in the face of population change and attrition ▪ Greater precision for change estimates, but increased bias (due to coverage losses over time and attrition) ▪ Because of attrition and population change, panels may lose representativeness over time ▪ Sample sizes fixed at the outset ▪ Low marginal costs for add-ons (linked samples, additional types of data)
Observational Errors (mode, interviewer, questionnaire, respondent errors)	▪ Good for measuring aggregate trends ▪ Depending on frequency and timing of surveys, can capture the spike-and-decay of unexpected shocks to the system ▪ Special studies needed to assess measurement error ▪ Limited ability to collect retrospective data ▪ Less reporting burden ▪ Can replicate key findings over time or study the impact of changes in temporal context	▪ Necessary for measuring individual change ▪ Difficult to alter schedule to capture the effects of unexpected events ▪ Can fit models that separate true change from measurement error ▪ Can minimize recall problems ▪ Data affected by conditioning, time-in-sample effects ▪ New panels needed for independent replication
Costs, Timeliness	▪ Often less expensive to mount ▪ Can be mounted more quickly ▪ Simpler to analyze	▪ Tracking costs can increase data collection costs, though subsequent waves are often cheaper than the initial wave ▪ Long lead time needed ▪ Often large “entry costs” for analysts; complexity and missing data may discourage longitudinal analyses

It is worth noting that it is possible to combine the best features of both designs in various hybrids. For example, the rotation group designs used by the Current Population Survey involves interviewing each rotation group a total of eight times; the sample for any particular month includes both a newly selected rotation group being interviewed for the first time and rotation groups that have been interviewed previously. The ANES design also incorporates both panel (typically within election cycles) and cross-

sectional features (across elections). But the overwhelming consensus was that both types of designs clearly have their value and that both types of studies are essential to social science.

Rising Costs and Declining Response Rates in Recurring Surveys

The last few years have seen sharp increases in survey costs and sharp declines in response rates. The drop in response rates is more noticeable for the initial wave of panel surveys than for their later rounds but losses due to attrition also seem to be edging up. This portion of the workshop examined several questions:

1) Why are response rates falling?
2) What is the impact of this decline on survey estimates?
3) Why are costs rising?
4) What research should be done to address these developments?

Response rates. There are a number of reasons why response rates are declining. People are harder to reach than they used to be. When they are reached, they are less likely to agree to participate in surveys and they are also less likely to be able to complete an interview. Each of these developments reflects long-term societal trends.

Consider, first, barriers to contact. More people live in gated communities, assisted-living settings, locked apartment buildings, and other settings that shield them from unwelcome visitors than lived in such settings ten or 20 years ago. Parallel developments (including the rise of cellular telephones, answering machines, Caller ID, and other technologies for screening telephone calls) have made it harder to reach people in telephone surveys. The rise of one-person households has decreased the likelihood of reaching potential respondents for either face-to-face or telephone surveys. These changes in residential patterns, technology, and the demographic composition of the U.S. population have all made it harder to reach potential respondents and have helped to drive up survey costs, as surveys increase the number of callbacks in an attempt to keep the proportion of sample households they eventually contact high.

The causes of the decline in willingness to participate are harder to pin down, but they also appear to reflect developments outside of survey research. Among the possible contributors are:

- reduced free time;
- heightened concern about privacy;
- greater mistrust of the government;
- increases in the length (and burden) of survey interviews;
- more survey requests than ever before;
- declining civic engagement;
- the rise of telemarketing.

Whatever the exact reasons for the increasing resistance to participation, surveys have been forced to resort a variety of methods in an attempt to counter it. These have included ever-larger respondent incentives, specially tailored appeals and refusal conversion letters, new methods for training interviewers, and offering potential respondents multiple modes for responding. In panel surveys, information obtained in the first interview may be used to tailor tracking and persuasion efforts in later rounds. Wave nonrespondents may be recontacted in subsequent rounds in an effort to persuade them to rejoin the panel. All of these countermeasures tend to drive up survey costs. A final contributor to the declining response rates has been ever growing number of sample members who are unable to complete an interview due to some linguistic barrier (it is often no longer enough simply to translate the questionnaire into Spanish) or to a hearing or other physical impairment.

Aside from the various trends affecting the persons who are selected for surveys, there have been developments affecting those who are supposed to conduct the interviews. The labor pool for interviewers seems to be shrinking. Few jobs involve going door to door anymore. Traditionally, most face-to-face interviewers have been women and women now have more opportunities than they used to in other fields. Interviewers also are more likely to have another job besides survey interviewing than was true 10 or 20 years ago. As a result, recruiting and training interviewers is harder than it used to be and turnover is worse. These trends mean that less talented and less experienced interviewers are being sent into the field and they tend to obtain lower response rates.

Rising costs. The increasing difficulty of maintaining high response rates is one major source of the rise in survey costs. It now takes more callbacks than it used to to reach sample members (whether in-person or by telephone), it may require costly methods to persuade reluctant sample members to participate (e.g., large incentives, tailored appeals), and special procedures may be needed to interview "linguistic isolates" or others who are difficult to interview. Other contributors to rising costs include the change in the labor pool from which interviewers are drawn and the increasing length and complexity of survey instruments. The latter is, in part, itself an outgrowth of rising survey costs. As survey costs have risen, it has become more common for multiple agencies to contribute to their funding; almost inevitably this involves adding items or modules that reflect the different research agendas of the different funding agencies, adding to the length and overall burden of the instrument. These longer and more complicated questionnaires in turn drive up the costs of conducting the interviews, programming and testing the instrument, and training the interviewers.

The rise of Institutional Review Boards (IRBs) and the constraints they impose on survey designs are added sources of increased costs. IRBs at some institutions have required written, signed consent for voluntary surveys, have limited callbacks to sample cases, have constrained refusal conversions, and have limited the use of incentives. The actions of the National Science Foundation's Directorate for Social,

Behavioral, and Economic Sciences (SBE) regarding this have been ameliorative, but the fact remains that many IRBs are increasing the cost of social science inquiries.

In an effort to reduce costs, surveys have adapted multiple modes of data collection. A cross-sectional survey may start with the cheaper modes (e.g., mail or telephone) and reserve more expensive methods (face-to-face interviewing) for the relatively small set of cases remaining after the cheaper methods have been exhausted. This is the strategy used in the American Community Survey. Longitudinal surveys may move to cheaper modes (e.g., telephone) for later rounds after using expensive methods (face-to-face interviewing) in the initial wave of data collection. More surveys now use adaptive strategies in which replicates fielded early are the basis for adjustments to the design for later replicates or in which only a subsample of the nonrespondents is retained for extensive follow-up efforts.

There may be additional hidden costs to the effort to maintain nonresponse rates in the face of mounting resistance. Many survey researchers suspect that reluctant respondents may provide less accurate information than those who are more easily persuaded to take part.

Impact of nonresponse. The most serious consequence of survey nonresponse is that survey estimates can be biased. Several recent empirical investigations have shown that increasing nonresponse *rates* do not necessarily produce increasing nonresponse *bias* in one-shot surveys. Although the general conditions that produce nonresponse bias in survey means or proportions are known (the bias is a function of both the nonresponse rate and the relation between the response "propensity"— the probability that a given case will become a respondent — and the survey variables), it is not clear what circumstances are likely to yield large nonresponse biases and what circumstances are likely to yield small or negligible ones. Several participants noted that the impact of attrition is likely to vary from one survey item to the next and the same thing will be true for total nonresponse as well. Thus, few of the participants took much consolation from the existing findings that relatively high levels of survey nonresponse do not necessarily signal large nonresponse biases. It is usually difficult or impossible to determine the level of nonresponse bias, although when there are external data available on the nonrespondents this can be helpful in assessing the impact of nonresponse; in a longitudinal survey, the data from the initial round can provide information useful for assessing the impact of panel attrition.

Most of the survey literature on nonresponse has focused on its impact on means, proportions, and totals. The impact of attrition may be reduced for more complex, multivariate statistics (such as regression coefficients), but clearly more work is needed to document this difference.

Areas for research. In the past, most research on nonresponse has taken the form of relatively atheoretical "engineering" studies directed at improving response rates for a specific survey. Such studies examine the impact of different methods for reducing nonresponse, for example, by varying the timing or amount of incentives offered. As long as nonresponse rates continue to climb, such studies will continue

to be done and they are likely to remain useful. Another kind of study is likely to assume increasing importance in the coming years; these studies will focus on the issue of when nonresponse produces large biases and when it can be safely ignored. Like investigations of measurement error, these studies may involve disruptions of ongoing efforts to maintain response rates (perhaps even lowering response rates by design) in order to assess their impact on nonresponse bias. In addition, it will be important to demonstrate that falling response rates actually matter (at least some of the time) and to understand the impact of nonresponse on complex statistics derived from survey data.

Research Frontiers in Survey Research

Both societal change and developments within survey methodology itself are likely to define the research agenda for survey researchers for the next decade or so. Nonresponse and the associated rise in survey costs are likely to remain important topics for new research. This research will continue to investigate trends in nonresponse, their causes, and their effects on survey estimates. Survey methodologists are also likely to continue to search for theories that delineate the conditions under which nonresponse introduces appreciable biases into survey estimates. Face-to-face surveys may have reached a tipping point in which the clustering of the sample no longer yields appreciable cost savings; the combination of longer interviews and shorter interviewer hours means that interviewers rarely complete more than a single interview in a trip to a sample segment. Thus, it may be possible to design samples that use clustering for substantive reasons — to provide contextual data on socially meaningful units, such as schools or neighborhoods — rather than for sampling convenience.

Another area that is likely to remain a focus for methods research involves new modes of data collection and mixed mode surveys. Obviously, Web surveys are likely to remain a hot topic for such investigations, but there are a number of broad trends affecting survey data collection over the last 20 years or so and these trends are likely to continue and to inspire new methodological work:

- increased reliance on self-administration rather than administration of the questions by interviewers;
- the move to supplement question text with images and other visual material;
- increase reliance on mobile computing; and
- increasing diversity of response formats, including analogue formats (such as slider bars).

Although some of these topics have a long history, there are many open questions. For example, although self-administration clearly improves reporting in many cases, it is possible that, under some conditions, it encourages sloppy responding ("satisficing") by reducing the respondents' sense of accountability for their answers. Similarly, it is not always clear whether data collected under different modes are truly

comparable. What it will take to produce a Web survey with acceptable levels of coverage and acceptable response rates is likely to remain a lively question.

The cognitive science revolution in survey methodology brought increased attention to measurement error in surveys and research on measurement error in surveys is likely to continue in the next decade, dovetailing with research on differences across modes of data collection in measurement error. Many basic issues in questionnaire design (such as the impact of wording choices) remain lively areas for research. The special reporting problems of panel surveys (e.g., conditioning and seam effects) are still not well understood and it remains an open question whether the positive effects of practice may sometimes offset the negative effects of repeated interviews.

The NSF infrastructure surveys have contributed important methodological work on many of these issues — the GSS, for instance, has been the source of key findings on question wording and context effects; the ANES has conducted important studies on mode of data collection and measurement error; and the PSID has done innovative work on event history calendars as a means of improving the quality of retrospective reports. But all three studies are likely to remain primarily platforms for substantive rather than methodological investigations. These substantive studies encompass both the core items and data collection of each survey and the various add-ons and auxiliary studies each supports. None of them is likely to become "early adopters" of new data collection technologies; according to the workshop participants, this is not an appropriate role for the infrastructure surveys, which should only employ methods that have been proven in other settings.

High Payoff Investments for NSF

The final session of the workshop concerned new investments that NSF might make — to advance methodological research, improve the existing infrastructure surveys, or mount new survey initiatives. Participants made a large number of suggestions for new investments by the National Science Foundation. We list these below, grouping them under three headings — methodological research, enhancements to the existing infrastructure surveys, and other new initiatives. Some of these ideas (notably, establishing a new panel for the study of social and political issues and improving the coverage of non-English speakers in all three surveys) were seconded by several participants. Other key themes involved improving the coordination among the three infrastructure surveys, increasing the level and stability of their funding, and launching various long-term programs of methodological research to address the problems faced by most current surveys — nonresponse, attrition, mode effects, and measurement error. Unfortunately, it would be hard to provide even ballpark cost estimates for most of these ideas, in part because they are not fully developed research plans and in part because many of them could be tailored in various ways to fit the funding available. For instance, the first of the methodological

initiatives calls for a systematic investigation of nonresponse bias, and, depending on the level of support available, such an investigation could take several forms, ranging from a secondary analysis of existing data to a program of field experiments involving extensive new data collection.

Under the heading of methodological research, six initiatives were suggested:

- A systematic investigation of the nonresponse biases in surveys;
- Additional theory-driven research exploring measurement differences across different modes of data collection;
- Studies looking at existing data on nonresponse to determine whether some people are generally nonrespondents across surveys or different people become nonrespondents for different surveys;
- Continued funding for studies on nonresponse and attrition in panel surveys;
- A long-term research program on measurement error in panel surveys; and
- A center for the study of research methods that would provide a conduit for the dissemination of methods work done by commercial and government survey researchers, who currently have little or no incentive to publish.

Clearly, nonresponse and mode effects are likely to remain high priority areas both for survey research as a field and for NSF.

The workshop participants also suggested a number of enhancements to the existing infrastructure surveys. Most of these enhancements would be funded or spearheaded by NSF. The initiatives under this category included:

- Full funding from the NSF for the infrastructure surveys (rather than forcing these studies to find multiple sponsors) and increased samples sizes for the GSS and NES;
- Securing funds for the infrastructure surveys from other federal agencies;
- Improved coverage of non-English speakers in all three surveys and, in particular, a Spanish version of the GSS questionnaire;
- Treating the infrastructure surveys as producers of official statistics and securing long-term support for them like other federal statistical programs;
- NSF funding for GSS and NES pilot studies;
- Coordination of the attitude items appearing on the GSS and NES to provide optimal level of overlap;
- Reorganization of the management of the infrastructure surveys at NSF along functional rather than disciplinary lines;
- A coordinated management structure for the infrastructure surveys, including shared internships and post-doctoral fellows, regular meetings, and auxiliary funding mechanisms for training and innovation;
- Funding for modules to be added to the GSS (such modules now have little chance for funding, since reviewers often think NSF has already invested enough in the GSS);
- A training program for outside researchers to take greater advantage of the infrastructure surveys;
- Oversamples of African Americans and Hispanics in the GSS;

- The addition of in-depth interviews with a subsample of GSS respondents to open the GSS up to qualitative researchers;
- Continued funding for auxiliary studies that gather data on the larger social units (e.g., organizations and social networks) in which the respondents are embedded;
- A long-term panel to examine political socialization and other long-term causal processes related to political attitudes and voting;
- Studies of the electorate during non-election years;
- Enhancing the value of the infrastructure surveys by allowing international comparisons to be made, with the NSF organizing international conferences to increase the comparability of similar surveys conducted in different nations;
- The addition of geodata to the ongoing studies;
- Providing the infrastructure surveys with some degree of insulation from short-term research trends.

Clearly, there were a wide range of suggestions for enhancing the existing infrastructure surveys, ranging from relatively low-cost ideas (translating the GSS questionnaire and fielding a Spanish version) to relatively expensive ones (mounting a new long-term panel on political socialization).

Finally, the workshop participants made a number of suggestions for other new initiatives not directly tied to the ongoing infrastructure surveys.

- A new panel survey to study the dynamics of religious and political behavior, attitude formation and change, and the development and evolution of social networks;
- Work on the impact of local geography and neighborhood on people's lives;
- New legislation that would make the Census Bureau's Master Address File available to other organizations (reducing their sampling costs) and that would remove legal barriers to matching survey data to administrative records;
- Assistance from NSF in obtaining letters of endorsement for surveys from members of Congress, Cabinet secretaries, and other senior officials;
- Bringing together a consortium of federal agencies to recruit and train a national field force of in-person interviewers, who could be hired by any of the major survey firms;
- New studies to capture the impact of events in real time;
- A management structure and training program that encourages systematic work carried out over long periods;
- A Web panel recruited from an area probability sample;
- Support for purely theoretical work on basic social behaviors;
- Mathematical training programs for the social sciences that would help establish a common vocabulary and promote interdisciplinary collaboration;
- A new long-term panel of households on social and economic issues;
- Assistance in establishing an integrated employer-employee database assembled from administrative records via support for work on confidentiality and disclosure issues, for research data centers, and for linkage of these data with data from the infrastructure surveys;
- Encouraging the use of methods panels for the investigation of methodological issues;
- Promoting the simplification of surveys through the use of the use of matrix sampling, modules administered to subsamples of respondents, and other methods.

Once again, some of these ideas would be quite costly to implement (e.g., creating a new panel survey) whereas others involve expenditures of NSF's time and energy rather than its funds (seeking legislative changes).

References

Atrostic, B.K., Bates, N., Burt, G., & Silberstein, A. (2001). Nonresponse in U.S. government household surveys: Consistent measures, recent trends, and new insights. *Journal of Official Statistics*, *17*, 209-226.

Black, D., Gates, G., Sanders, S., & Taylor, L. (2000). Demographics of the gay and lesbian population in the United States: Evidence from available systematic data sources. *Demography*, *37*,139-154.

deLeeuw, E., & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey Nonresponse* (pp. 41-54). New York: John Wiley and Sons.

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

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household surveys*. New York: John Wiley.

Kish, L. (1987). *Statistical design for research.* New York: John Wiley.

Appendix A: List of Workshop Participants

Participants	Discipline	Institution/Organization
John Abowd	Economics	Cornell University
Joseph Altonji	Economics	Yale University
Nancy Bates	Survey Research	U.S. Census Bureau
Nancy Burns	Political Science	University of Michigan
Mick Couper	Survey Research/Sociology	University of Michigan
Michael Dennis	Survey Research	Knowledge Networks
Barbara Entwisle	Sociology	University of North Carolina at Chapel Hill
Robert Groves	Survey Research/Sociology	University of Michigan
John Haltiwanger	Economics	University of Maryland
Guillermina Jasso	Sociology	New York University
Daniel Kasprzyk	Statistics/Survey Research	Mathematica
Jon Krosnick	Psychology/Political Science	Ohio State University
Maria Krysan	Sociology	University of Illinois at Chicago
Samuel Lucas	Demography	University of Berkeley
Peter Marsden	Sociology	Harvard University
William McCready	Survey Research	Knowledge Networks
Colm O'Muircheartaigh	Statistics/Survey Research	University of Chicago
Dawn Nelson	Survey Research	U.S. Census Bureau
Randall Olsen	Economics	Ohio State University
Robert Schoeni	Economics	University of Michigan
Frank Stafford	Economics	University of Michigan
Marianne Stewart	Political Science	University of Texas at Dallas
Roger Tourangeau	Survey Research/Psychology	University of Maryland/University of Michigan
Reeve Vanneman	Sociology	University of Maryland
Norman Bradburn	Assistant Director, SBE	National Science Foundation
Cheryl Eavey	MMS	National Science Foundation
James Granato	Political Science	National Science Foundation
Kwabena Gyimah-Brempong	Economics	National Science Foundation
Richard Lempert	Division Director, SES	National Science Foundation
Joane Nagel	Sociology	National Science Foundation
Daniel Newlon	Economics	National Science Foundation
Laura Razzolini	Economics	National Science Foundation
Frank Scioli	Political Science	National Science Foundation
Patricia White	Sociology	National Science Foundation

Appendix B: Participant Write-Ups

The Role of Integrated Administrative Data in Recurring Social Surveys

John M. Abowd and Julia Lane[1]
March 2003

This paper was prepared for the NSF workshop on repeated cross-sectional and longitudinal surveys. It is based heavily on presentations that we have made at other conferences and colloquia. We would like to thank our many coauthors and colleagues at Cornell, the Census Bureau, the Urban Institute, and INSEE/CREST and other national statistical institutes. Funding was provided by the US Census Bureau, INSEE/CREST, the National Science Foundation (SES- 9978093 to Cornell University), the Department of Labor (ETA), the Bureau of Transportation Statistics, the National Institute on Aging, the Sloan Foundation, and the Russell Sage Foundation.

Modern national and research statistical systems acquire their information from three related sources: censuses, periodic surveys, and administrative records. We want to focus on the benefits of using these three sources in a more integrated fashion. In particular, we outline the benefits to social science, twenty-first century statistical systems, and policy analysis that accrue from careful creation and analysis of integrated employer-employee data. When Haltiwanger, Lane, Spletzer, Theewes, and Troske (1999) organized an international colloquium in 1998 to discuss applications of large scale data integration for employer-employee data there were twenty-three contributions representing data from 14 different countries. When Abowd and Kramarz (1999) reviewed the creation and use of such data, there were already more than 100 scientific studies using data from 18 different countries. Recent years have seen an explosion of such data for three related reasons. First, the unique ability of these data to address important open questions in a variety of social sciences has put academic researchers at the forefront of their creation. Second, continual improvements in information technology have permitted both improved confidentiality protection and better analysis methods. Third, the marginal cost of creating integrated data is orders of magnitude lower than the marginal cost of censuses and surveys, even after properly accounting for the extra costs associated with enhanced privacy protections.

Why is advancing scientific at the top of the list? Like many of the statistical innovations of the past (*e.g.*, national income accounts) work on integrated employer-employee data has been driven by academic social scientists addressing substantive open questions. Many examples come from the study of labor markets (see the attached example of integrating data into the Survey of Income and Program Participation) others (see Davis and Haltiwanger 1999) consider applications to macroeconomics and industrial organization. Demand and supply is sharp tool that works well in many markets. In the labor market, however, the tool has consistently delivered predictions far sharper than the actual data support. Longitudinally-integrated employer-employee data have permitted the first clean empirical decomposition of the sources of this remarkable labor market heterogeneity—and it is this scientific analysis that has stimulated much of the demand to create such data for additional countries and time periods.

The second item on the list is the creation of twenty-first century statistical systems. Such systems, whether used for official national statistics or for research purposes, will inevitably

[1]Abowd: Cornell University, US Census Bureau, CREST, NBER, and IZA; Lane: The Urban Institute, US Census Bureau, NIA, and New Zealand Department of Labor.

involve substantial integration of information from multiple sources. The use of sophisticated information technologies and statistical matching methods already pervades national statistical systems. What is a twenty-first century system? The most precise definition is a system in which, *by design*, the information used to produce the desired estimates is collected from the lowest marginal cost sources then integrated using high-powered information technologies and formal probability models. The important distinction with historical systems is the improvement that comes with designing a system to run with integrated information rather than integrating related data *ex post*. The example here is the Quarterly Workforce Indicators system under development at the US Census Bureau (see LEHD Program 2002). As a part of the discussion of twenty-first century statistical systems, we also want to consider the role that social science researchers can play in the protection of the confidentiality of these data. The research community has not paid sufficient attention to the possibilities inherent in cooperative development of confidentiality protection systems (but see Doyle, Lane, Theeuwes, and Zayatz, 2001 for an important advance in this area). Confidentiality protection has very important implications for the way integrated data are produced and used. When such systems are developed with minimal input from subject-matter specialists, avoidable limitations in the final data product often occur. Because the privacy and confidentiality protection issues associated with integrated employer-employee data have been stressed by all of the agencies that have created them, it is important that the same scientists who helped create the data participate in the design of the confidentiality protection system. This participation will help insure that the fundamental insights that these data permit are preserved in the statistics produced by the confidentiality-protected system (see Abowd and Woodcock, 2001).

Finally, we argue that integrated social data unlock important information is the analysis of the effects of policy changes on the target community or market. There are many examples from the labor market but the same ideas can apply to areas like health care and public finance. The innovation permitted by the use of integrated employer-employee data is the possibility of distinguishing more cleanly the effects of incentives on each side of the market. Researchers in France using INSEE's rich archive of integrated data have analyzed the effects of French labor laws on the way firms adjust employment. Researchers in the US are now using them for the study of welfare systems and transportation patterns.

A twenty-first century statistical system is one that is designed to produce the desired estimates by systematically integrating information from multiple sources using the lowest marginal cost source, formal probability modeling for the integration, and comprehensive confidentiality and privacy protection. Since 1998, we have been associated with the US Census Bureau's Longitudinal Employer-Household Dynamics Program, which is attempting to build just such a system by integrating information from state-level unemployment insurance system records and the Bureau's own economic and demographic surveys and censuses (see Abowd, Lane and Prevost, 2000). The program has several projects and we are going to describe three of them.

As a part of its statistical mission, the US Census Bureau supplies about three-fourths of all of the data used to produce the American National Income and Product Accounts, which are the responsibility of the Bureau of Economic Analysis, a sister agency in the Department of Commerce. In support of this data collection effort, the Bureau, through its Economic Censuses and Surveys, collects information about the inputs and outputs of most sectors of the American economy. In virtually every sector, the most important input is labor services. The Bureau finds it monumentally expensive to gather detailed information about the labor input. In addition, the Bureau of Labor Statistics, part of the Department of Labor, has primary statistical responsibility for direct measurements of the US labor force and associated statistics. From a scientific viewpoint, it is quite advantageous to have both input and output measures at the same level of economic activity—ideally, the business establishment. Direct collection of detailed information

by the Census Bureau on the labor service component of inputs isn't feasible for the reasons stated above. The LEHD Program uses its integrated employer-employee data to provide just this information at the business establishment level and without additional data collection costs.
The LEHD Program has also developed a system of Quarterly Workforce Indicators (QWI) from its integrated data. Using quarterly unemployment insurance information, and working closely with state-level experts on these data, the Bureau's new QWI system produces quarterly information on job creation, job destructions, accessions, separations, new hires, recalls, and earnings. All of the new estimates, which cover all employment in the unemployment insurance reporting system, can be produced by detailed geography and industry and by age and sex. The US Census Bureau could undertake such a project because the information integrated from the two sides of the labor market is available for use under a specially-created confidentiality protocol that physically separates the identifiers protected by American privacy laws from the identifiers used for the data integration.

A third project underway at LEHD is the development of a public use file that integrates administrative data from the Social Security Administration with survey data from the 1990-1996 Survey of Income and Program Participation panels. Special new Treasury Regulations were enacted to enable the creation of this file because the information in the SSA system that comes from individual and business tax forms is protected under US Code Title 26, which is the operating code for the Internal Revenue Service. Regular Census confidential data are protected under US Code Title 13, which is the operating code for the Bureau of the Census. We have attached a summary of this project to this paper.[2]

Integrated employer-employee data present very challenging confidentiality protection issues (see Abowd and Woodcock, 2001, in particular and other essays in Doyle, Lane, Theeuwes, and Zayatz, 2001, for other challenging confidentiality protection issues). The QWI system described above has its own confidentiality protection system that was designed to permit the release of very detailed estimates without the use of cell suppression, a widely adopted method that would have resulted in the suppression of most of the demographic detail in the EDE system (LEHD Program 2002).

Another new type of confidentiality protection of this type of data—the one that Abowd and Woodcock prototyped for the INSEE data—is being used to prepare the public use integrated SIPP-SSA file. It is reasonable to ask why economists should concern themselves with confidentiality protection issues when they already have access to the confidential microdata. The answer is that there is a perpetual tension between the vigor of scientific inquiry, which requires that multiple research teams have access to the same data, and the realities of confidential microdata, which require access in tightly controlled environments. Subject matter specialists are very good judges of the usefulness of a confidentiality protection system because they understand how the underlying data are used in their fields of specialization. This is why we strongly suggest that other social scientists also become more familiar with these methods.

It is a very difficult problem to provide general confidentiality protection of microdata and to preserve their analytic properties. By developing new confidentiality protection methods, we are trying to increase the research access to the data. By developing systems that build directly on the underlying linked microdata, without destroying any of the original confidential data, we hope to provide a layered path from confidentiality-protected public use data to confidential microdata that is as analytically useful as resources permit. The new techniques are based on masking the microdata using the predictive distribution of the confidential data (see Rubin, 1993; Feinberg

[2] The summary was prepared by Martha Stinson of the LEHD Program.

1994; and Kennickell, 1997). These techniques preserve many of the analytic features of the integrated data, including moments and complicated multi-factor nonlinear relations, because they are implemented using massive amounts of the information in the underlying, integrated, microdata.

We are just beginning to realize the potential of integrated employer-employee data for understanding labor markets. But it is not just labor markets that will benefit from more extensive integrated social data. Regional growth and transportation specialists can exploit the same systems by using the integration of place of work (a characteristic of the employer) and place of residence (a characteristic of the individual) to study mobility patterns and time-of-day populations in urban areas. Industrial economists can study the patterns of births, deaths, and complex recombinations among business using both the mobility of employees and of capital assets to measure the reallocations of economic resources. Social network specialists can use the pattern of connections among individuals and employers to measure the extent to which colleagues and employment histories affect productivity and wages (see Lengermann, 2002).

Health care specialists can use such data to untangle the complex relation among demographic characteristics of the household and employer-provided health care options (see Stinson, 2002). These examples demonstrate that the scientific value of social data, the quality of official statistics programs, and the protection of the confidentiality of the underlying microdata can all be enhanced by the active participation of the research community in the development of these integrated systems.

References

Abowd, John M. and Francis Kramarz, "The Analysis of Labor Markets Using Matched Employer-Employee Data," in O. Ashenfelter and D. Card (eds.) *Handbook of Labor Economics*, Volume 3(B), Chapter 40 (Amsterdam: North Holland, 1999), pp. 2629-2710.

Abowd, John M., Julia I. Lane and Ronald C. Prevost, "Design and Conceptual Issues in Realizing Analytical Enhancements through Data Linkages of Employer and Employee Data" in *Proceedings of the Federal Committee on Statistical Methodology*, November 2000.

Abowd, John M. and Simon D. Woodcock, "Disclosure Limitation in Longitudinal Linked Data," in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), (Amsterdam: North Holland, 2001), pp. 215-277.

Davis, Steven J. and John Haltiwanger, (1999). "Gross Job Creation and Destruction," in O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Volume 3(B), Chapter 41 (Amsterdam: North-Holland, 1999), pp. 2711-2805.

Doyle, Patricia, Julia Lane, Jules Theeuwes, and Laura Zayatz, "Introduction," in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), (Amsterdam: North Holland, 2001), pp. 15.

Fienberg, Stephen E., "A Radical Proposal for the Provision of Micro-data Samples and the Preservation of Confidentiality," Carnegie Mellon University Department of Statistics Technical Report, No. 611, 1994.

Haltiwanger, John C., Julia I. Lane, James R. Spletzer, Jules J.M. Theeuwes and Kenneth R. Troske, (eds.) The Creation and Analysis of Employer-Employee Matched Data, (Amsterdam: North Holland, 1999).

Kennickell, Arthur B. "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances," SCF Working Paper, 1997.

Lengermann, Paul, "Is it Who You Are, Where You Work or With Whom You Work: Reassessing the Relationship between Skill Segregation and Wage Inequality," University of Maryland Ph.D. thesis, October 2002.

Longitudinal Employer-Household Dynamics Program, Employment Dynamics Estimates Project Versions 2.2 and 2.3, US Census Bureau, LEHD Program, Technical Working Paper TP 2002-05 rev1 (May 2002).

Rubin, Donald B. "Discussion of Statistical Disclosure Limitation," *Journal of Official Statistics* 9 (1993): 461-468.

Stinson, Martha H., " Estimating the Relationship between Employer-Provided Health Insurance, Worker Mobility, and Wages," June 2002, Cornell University Ph.D. thesis.

The second data product considered was to choose a larger set of SIPP variables and use a statistical masking technique developed in Abowd and Woodcock (2001).[3] This product has the advantage of preserving confidentiality while allowing the inclusion of 30-40 SIPP variables. Although it will require significantly more data work, it will produce a product that will be enormously useful to the research community. Thus after lengthy discussions among the four agencies involved, this approach has been chosen.

Summary of Progress thus far:
1. January 2002: Committee agrees to pursue masking approach
2. May 2002: LEHD delivers first version of "Gold-Standard" file
3. September-December 2002: LEHD visits SSA to present methodological approach and ask for input on the "Gold-Standard" file. LEHD and SSA hold several meetings to discuss final variable list
4. January 2003: final variable list agreed upon

Schedule for 2003:
1. May 2003: LEHD will deliver second version of "Gold-Standard" file
2. Summer 2003: SSA and CBO will test "Gold-Standard" file using existing statistical research models to establish base-line relationships among the included variables. LEHD begin masking.
3. Oct 2003: LEHD will deliver first masked dataset. Depending on unforeseen difficulties that may arise, this may be a subset of the full dataset.
4. Winter 2003-Spring 2004: SSA, CBO, and Census test masked dataset to determine adequacy of confidentiality protections and ability of masked data to preserve statistical relationships among variables. Give feedback to LEHD.
5. May 2004: Depending on outcome of testing, committee will evaluate next steps.

Possibly create third version of "Gold-Standard" file and do another iteration of masking procedure.

[3]Abowd, John A. and Simon Woodcock. Disclosure Limitation in Longitudinal Linked Data, in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, P.Doyle, J.Lane, L. Zayatz, and J.Theeuwes (eds.), (Amsterdam: North Holland 2001), 215-277.

Some Thought on Declining Response Rates in Panel Surveys

Joseph G. Altonji
Department of Economics
Yale University

March 27, 2003
Very Preliminary

I am not a survey researcher, and my assigned topic---rising costs and declining response rates in over time surveys---is one that I know relatively about. However, the respondents to social surveys are frequently in the same situation, and researchers simply hope that they will respond as best they can rather than fail to respond. I will try to do likewise. I will let the other, better informed members of the panel address the specific questions about the behavior of response rate patterns over time, increases in panel attrition, what is behind non-response, etc that were listed in the conference program. However, I do have a few remarks on the importance of attrition that may be of some relevance to NSF as it plans its research program.

First, most of the admittedly limited research that I am aware of suggests that the problem of attrition is more important for studies that focus on univariate distributions of a variable (e.g., family income) at a point in time or over time (such as the variance of income) rather than for conditional inference in a multivariate setting. The studies that I am familiar with find that panel attrition on multivariate analyses leads to only modest biases. (See for example, the set of papers on attrition published in the Spring 1998 issue of the Journal of Human Resources.) Attrition is very costly to panel studies, in part because with lost cases we lose the opportunity to add new questions and variables that leverage the rich set of variables that came before. However, I do not think that attrition in major US longitudinal datasets, such as NLSY, the PSID, or even the earlier NLS datasets are a large source of bias relative to other problems with observational studies based on survey data, such as measurement error, omitted variables, and simultaneity.

Second, it is extremely helpful to know as much as possible about what leads to a lost case. For example, it matters greatly whether a respondent leaves the sample because of a death, because they have moved and cannot be tracked, or because they refuse to respond. Information about the factors leading to nonresponse are provided in some data sets and can be very helpful for assessing the effects of attrition. Rich baseline data is also very helpful is assessing and addressing bias from attrition.

Third, with the explosion of electronic databases and the Internet, it may be easier than in the past to obtain basic information on individuals who leave the sample, such as residential address, fertility (from birth records), and marriage. Such information might be useful in developing variables that can be used to deal with attrition bias. For example, residential address in combination with property tax records provides information on home ownership and house value, which are key socioeconomic indicators. Furthermore, it may be possible use administrative data to obtain information that is at least a partial substitute for key survey items. Obviously, there are extremely difficult confidentiality and human subjects issues associated with these approaches. But they are worth exploring.

Fourth, studies of attrition bias may be useful in deciding how to allocate scarce resources across cases to reduce attrition. Such studies may indicate the types of cases, based upon baseline

characteristics, that are most important to keep in the sample. The problem with this is that the most important cases will depend on the research question.

Fifth, although it is difficult to survey respondents who have already missed a wave of a panel study, there is a large analytical payoff to devoting large resources to obtain some basic information about them. With such information one can correct for attrition in the main sample. Furthermore, with increases in nonresponse over time, it is more necessary than in the past for panel surveys to go after persons who were not interviewed in the previous wave.

Finally, I wish to close with some observations about the implications of rising attrition rates and rising costs of maintaining sample for the value of preserving existing long running panel data sets, particularly the Panel Study of Income Dynamics. The PSID already has a large sample covering three generations. It covers an extraordinarily broad range of subject areas, with thousands of variables. Each new wave and special module leverages the huge existing stock of variables. The best studies of attrition have generally indicated that after weighting the PSID still provides a very good description of the U.S. population, with the exception of immigrants. It is interesting to speculate on whether, in the current survey environment, it would be possible to start over, recruit a new sample, and after 35 years or so have a sample that is nearly as large or as representative as the PSID seems to be. I doubt it.

Rising Costs and Declining Response Rates in Recurring Surveys

Nancy Bates
U.S. Census Bureau
Prepared for the National Science Foundation Workshop on Recurring Surveys
March 28-29, 2003

Are response rates declining?

Yes. Quantitative indicators collected by the Demographic Surveys Division suggest that response rates have declined over the last decade. Compared to 1990, by 2001 nonresponse for the initial interview increased between 4-7 percentage points for the Current Population Survey (CPS), National Crime Victimization Survey (NCVS), National Health Interview Survey (NHIS) and the Survey of Income and Program Participation (SIPP). For example, CPS and NCVS initial nonresponse was close to 5 percent in 1990, but by 2001 had climbed to almost 10%. The two consumer expenditures surveys, the Consumer Expenditure Diary (CED) and Consumer Expenditure Quarterly (CEQ), experienced even larger declines. By 2001, initial nonresponse had climbed to 28 percent for the Diary and around 21 percent for the Quarterly – increases of over 10 percent for the CED and 8 percent for the CEQ since the 1990.

As a component of total unit nonresponse, the refusal rate has also slowly but steadily increased for Census Bureau recurring surveys since 1990. However, refusals comprise about the same proportion of overall nonresponse at the end of the decade as they did at the beginning. Alternatively, the 'no one home' rate became more of a problem over the years and only recently appears to have leveled out as a component of nonresponse – whether this is a real trend or not is too soon to say.

Is panel attrition getting worse?

Yes. For the most part, the sample loss rate for Census Bureau longitudinal surveys has increased since the early '90's. For the 1996 SIPP Panel, the loss rates were most noticeable in the later waves – for the 2001 SIPP Panel, attrition appears worse in early waves but may be leveling off. This could be a result of methodological follow-up procedure changes and the use of incentives.

What is behind these nonresponse problems?

Research and field experience point to a number of factors contributing to declining response. A discussion with the Assistant Division Chief for Current Surveys (Field Division) suggested the problem has two major perspectives: respondents and interviewers. From the respondent perspective, cooperation is declining because of increased privacy concerns, fear of identity theft, decreased land-line phone use and increased cell phone use, increased language barriers, suspicion of survey legitimacy and mistrust of government. He also cited respondents having less discretionary time due to dual income families and 'sandwich generation' households (those sandwiched between aging parents who need care or help and their own children).

From the interviewer perspective, it is simply more difficult to find qualified field representatives (particularly ones with good language skills, a car, a telephone, and that are available to work at odd hours). Most industries that previously hired a 'door-to-door' workforce are virtually nonexistent. With a talent pool that needs to be developed, it is more difficult to maintain high response rates. The typical age of our field production interviewers is early to mid 50's. This suggests that the interviewer profession

is most appealing to a narrow age group, making the pool even smaller. In addition, the turnover rate is 15-20% annually which adds to the difficulty of maintaining a stable workforce. All of these factors contribute to declining response rates from the other side of the coin.

Are survey costs rising?

Yes. In addition to increased training costs associated with high interviewer turnover, federally mandated pay rates have increased over the decade resulting in higher costs to employ interviewers. Additionally, the number of visits/calls necessary to make contact with households has risen dramatically. This drives up travel costs and the amount of time interviewers must devote to each case. As a result, the cost per case is increasing.

What methods can be used to bring costs down?

A change in program requirements will bring costs down, that is, a decrease in the level of acceptable response rates to Census Bureau recurring surveys. This program change could theoretically occur without severe compromises in data quality if we can distinguish between ignorable nonresponse and non-ignorable nonresponse. By concentrating resources on keeping non-ignorable nonresponse to a minimum, we can cut costs by spending less resources on ignorable nonresponse. Response rates would decline, but should not result in biased estimates.

Other methods for containing costs include: changing modes away from personal visit to telephone contacts, regular use of incentives, drawing more efficient samples (clustered) and allocating sample more in advance to allow additional canvassing and advanced centralized locating research (to obtain good phone numbers) before interviewers try to contact a case in the field. To decrease field costs and manage interviewers more efficiently, the personal visit surveys should implement automated contact histories to record number of visits, interim outcomes, day/time of visits, and reasons for non-cooperation. Finally, surveys should tap the Census 2000 Planning Database to geocode their samples into high and low response areas, making use of the predictive factors that comprise these indices (e.g., linguistically isolated, high percentage renters, high minority population, etc.) when planning assignments and establishing realistic response rate goals.

Designs for the Study of American Politics

Nancy Burns
University of Michigan
Principal Investigator, American National Election Studies
March 24, 2003

In my few pages, I want to talk about the ways in which the study of electoral politics, participation, representation, and public opinion – that is, the study of democratic politics – benefits from two designs the American National Election Studies (ANES) has implemented on a regular basis. The ANES has, for the past fifty years, carried out an unbroken series of repeated cross-sections, organized around Presidential and midterm elections. In addition, the ANES has fielded a number of panels, often as part of the series of repeated cross-sections and sometimes as a supplement to those cross-sections: the 1956-1958-1960 election panel; the 1972-1974-1976 election panel; the 1992-1994-1996 election panel; an election-year panel in 1980; the 1990-1991-1992 Gulf War panel; and the 2000-2002 9/11 panel. A tremendous array of research relies on the capabilities provided by this portfolio of designs.

Repeated cross-sections. Repeated cross-sections, drawing on comparable samples, using a constant mode, and maintaining comparable question wording and implementation, offer opportunities for

- studying both historical and episodic change through systematic data, opportunities that enable the understanding of social dynamics;
- replication crucial to scientific progress, by enabling side-by-side analyses through time;
- specifying how context shapes individual-level results, by offering the opportunity to try out models in systematically different contexts and allowing scholars to put the current election in the context of 50 years of elections.

In general, scholars have used the repeated cross-sections to study shifts in national priorities, the intrusion of third party candidates on the American two-party system, the sweeping changes and remarkable stabilities that have characterized Americans' views on race over the last half century, the emergence and consolidation of incumbency as an electoral asset, fluctuations in support for national policy, participation in and withdrawal from political life, changes in the audience for network news, alterations in the level and meaning of trust in government, the life history and partisan rearrangement of core issues, and the dynamics of electoral choice itself.

A few specific cases from the two leading Political Science journals – the American Political Science Review and the American Journal of Political Science – in the last few years make clear the range of uses of these data. Scholars have drawn on all of the major advantages of a repeated cross-section in their work. They have studied

- the revival of partisanship in presidential elections, refuting the conventional wisdom about the decline of parties in American politics (using all of the cross-sections in the series; the focus is historical change).

- the ways in which actual voters deviate from idealized fully informed voters, and the aggregate implications that follow from information deficits (using all of the Presidential cross-sections from 1972 through 1992; the focus is on replication and on the ability to isolate the effects of electoral context).

- the demobilizing effects of negative advertising (replicating results across election studies to build confidence in the results and in the estimation strategy).

- the ways in which voters coordinate their choices for president and House of Representatives to achieve policy moderation (pooling NES data from 1976 through 1996; allows parameter variation over years; builds the case for a repeated, regular behavior of the American electorate).

Panel data. Panel data offer complementary opportunities

- to understand individual-level change in the face of societal change;
- to develop dynamic accounts at the individual level that complement the dynamic accounts based on repeated cross-sections;
- to adjust for measurement error in survey responses.

Again, to what effect? Scholars have focused on learning, on models of updating, and on opinion stability and change. They have used the panels to study, for example,

- the precise scope and extent of individual updating about party politics (using the 1990-91-92 and 92-93-94 NES panels).

- the extent to which political learning requires emotional involvement in politics (using the 1980 NES panel).

- the extent to which voters offer reasons or rationalizations for their vote (using the 1990-1992 NES panel).

- the extent to which interest and education are endogenous to turnout (using the 1990-1992 panel).

- the extent to which Americans have real attitudes as against "non-attitudes" about the issues of the day (using the 1956-58-60 NES panel).

Repeated Cross-Sections and Panels. Best practice often relies on both kinds of designs. As scholars begin to offer accounts of the consequences of September 11th for democratic practice in America – for public opinion, political engagement, political action, for the link between citizens and government more generally – they will necessarily draw upon both kinds of data. First, they will want to compare comparable cross-sections taken before and after September 11th, to learn whether basic relationships changed. They will be interested in whether the things they see in the comparison between 2000 and 2002 resemble changes over other unsettling periods earlier in American history. And so they might compare the changes from 2000 to 2002 with changes over the Cold War or the Vietnam War or the Gulf War, to figure out just how unusual the constituent parts of American democratic practice were in 2002. They will want to put both 2000 and 2002 in context. Then, they will turn to data on individuals in a panel (NES 2000-2002) to learn just who changed and why. That will help scholars understand the lessons individual Americans drew from 9/11. Finally, scholars will turn to comparable NES panel data taken around other big events to know whether this sort of individual-level change is common or is unique to 9/11. The obvious comparison is with the Gulf War Panel, but creative scholars will be able to use many of the earlier panels to draw out the implications of 9/11. Without repeated cross-sections, scholars wouldn't be able to put 9/11 in context, to say what the aggregate consequences of 9/11 are for American democracy. Without the short-term panels, they wouldn't be able to offer an account of learning and change at the individual level. Without the comparative panel data from other points in

American history, they wouldn't know which parts of the individual-level account are common parts of change over time and which parts are consequences of September 11th.

The combination of panels and repeated cross-sections has methodological advantages as well as these substantive advantages. It is simply impossible to gauge the extent of panel conditioning without identical side-by-side studies drawing panel cases and fresh cross-sectional cases. Scholars have tried, of course, to untangle the effects of panel conditioning without identical side-by-side studies but they have come under sharp criticism (see, for example, the criticisms leveled by Holt 1989).

As Bartels (1999, p. 2) has argued,

> one of the benefits of the NES investment in panel surveys over the last decade is that it provides an unusually strong data base for analysis of methodological problems that arise in all panel surveys. Most importantly, the emphasis of NES on facilitating panel analysis while preserving the integrity of a time series extending over 50 years has produced a series of data sets with parallel panel and fresh cross-section components…such data sets provide important opportunities for assessing and, possibly, mitigating the inferential pitfalls of panel data.

He makes clear that fresh cross-sectional respondents "interviewed at the same time by the same interviewers using the same survey instrument allow for an unusually straightforward and powerful assessment of the effects of both panel attrition and panel conditioning" (1999, p. 5).[4] Bartels finds that, in the panels conducted by the ANES, panel attrition and panel conditioning reduce "the inferential value of panel data by about 20%" across a wide range of analyses. These biases are concentrated exactly where one would expect them to concentrate: in analyses of campaign interest and turnout.[5] Elsewhere, for variables with more tenuous relationships to survey participation, there were few biases to speak of in the short panels NES has traditionally fielded.

Were one considering longer panels as bound up with the elements of democratic practice – action, opinion, engagement, and representation – as the ANES, one would, of course, want to carry identical, parallel fresh cross-sections to make the biases in the panels detectable and thus to make the panels useful.

[4] See, too, Corder and Horvitz 1989, pp. 316-317, and in Waterton and Lievesley 1989.

[5] There are parallel results about variables in this neighborhood in Waterton & Lievesley (1989, pp. 327-8) and in Wang and Cantor (n.d.); see, too, the review of early studies in Kalton, Kasprzyk, and McMillen (1989).

New Modes of Data Collection

Mick P. Couper
Survey Research Center
University of Michigan

Many new methods of survey data collection have appeared on the scene in the last several years and even more new methods appear imminent. The question is, to what extent these are relevant for the NSF infrastructure surveys and how best use can be made of these new methods to enhance and extend the capabilities of the key infrastructure surveys.

Several broad trends in methods of survey data collection can be identified. Briefly, these are:

- The move from interviewer-administration to self-administration. In recent years we have seen the rapid adoption of technologies to facilitate the automation of self-administered surveys, including CASI and its variants (e.g., audio-CASI), interactive voice response (IVR), and the World Wide Web. There are two drivers for this trend: 1) self-administration is generally cheaper than self-administration, and 2) for some types of questions, self-administration leads to "better" reporting.
- The move from verbal (written or spoken) inputs and outputs to visual and haptic/sensorimotor inputs and outputs. To put it briefly, the delivery of survey questions has extended far beyond verbal stimuli (either written on paper or read by an interviewer) to include a wide variety of visual and auditory stimuli, including pictures, videos and sound file. These have the potential to revolutionize how we think of survey questions. Similarly, the responses that are provided are not longer limited to written or verbal inputs. With automated systems, respondents can interact directly with multimedia interfaces, their verbatim responses can be digitally recorded, digital photographs can be captured, new tools such as graphical slider bars and running totals for immediate feedback can be used, and so on. Together with a range of other measurement tools, these can provide a wealth of data to survey researchers, far beyond what is typically obtained in a survey interview.
- The move from fixed to mobile computing, both for data collectors and for respondents. The devices used for computer assisted data collection are increasing the mobility of survey interviewers, again permitting a wider range of measurement opportunities. Similarly, the use of mobile phones, personal digital assistants (PDAs) and so on, now make it possible to put computerized survey devices in the hands of respondents, to record events and reactions closer to the time that they happen.
- The move from discrete surveys to continuous measurement. Given the above trends, it is increasingly possible to conduct survey measurement on an ongoing basis, matching the frequency of the behaviors that are being studied. Daily diaries, hourly measurements, event-driven surveys, the use of transaction data, and so on, are all facilitated by mobile computing technologies.
- The move from single-mode surveys to mixed-mode or multi-mode approaches. For reasons of both cost and efficiency, surveys are increasingly making use of multiple modes of data collection, whether driven by operational needs or respondent preferences. This again can potentially increase the range of measurement options, but has implications for measurement error as well as nonresponse, coverage and sampling errors.

All of these trends—and some have been underway for some time, have implications for the quality of survey data, and for the cost and timeliness of the data collected. These developments have the potential to enhance and extend survey data collection, to permit the exploration of new domains using

new methods. However, much methodological work remains to be done to understand the implications of these new methods for survey measurement.

Will these new developments mean dramatic changes for the ongoing infrastructure surveys? Will the new modes of data collection—especially the Internet, which is receiving the most attention at present—replace the traditional approaches used for these surveys? I believe not—not soon anyway. But will the new developments have an impact on the large-scale ongoing survey efforts? Absolutely. The key question is, how do we exploit the new developments while retaining the core value and quality of these surveys?

There are several ways in which a marriage of the infrastructure surveys and new methods can enrich the quality of information generated. Some brief examples follow:

- The core surveys can (and already do) serve as a benchmark by which many other surveys are evaluated. The responses to Internet surveys are often compared to the GSS or NES, for example, to demonstrate their quality. More so than just benchmarks, these surveys can be used as a gold standard, both for careful adjustment of the estimates from other modes, and for evaluation of the quality of data collected by newer methods.
- Mixed mode designs can be more extensively employed, particularly in follow-up activities. An initial face-to-face interview can be followed up with other modes of data collection that may be cheaper, permitting the introduction of partial panel or rotating panel designs to better explore the dynamics of change. Given the initial investment in contacting sample persons and gaining their cooperation, value can be added by following up these respondents over time.
- Because of their size and complexity, the surveys under discussion are relatively unresponsive to changing circumstances. A mode such as telephone or the Internet is more suited to tracking rapidly changing behaviors or attitudes on an ongoing basis. Given this, the larger, less frequent surveys can serve as calibration points for studying time trends.
- Cheaper, newer methods of data collection can be used to "prove in" content for the less-flexible infrastructure surveys. There is always a tension between the need to keep content for longitudinal analysis, and the desire to incorporate new measures into these surveys. This process could be facilitated by the ability to test alternative approaches using less expensive methods.

In summary then, the newer modes of data collection—IVR, Internet, etc.—may work for some surveys, but not for others. Respondent selection, recruitment, and persuasion are arguably still the most critical elements of these NSF-funded surveys. There is no magic bullet—high quality, high response rate, probability-based surveys, those that provide both rich and detailed information from a large number of respondents and permit inference to the general population with some confidence—still take time and effort. The newer modes of data collection fulfill many specialized niches, and are valuable additions to the survey researcher's toolkit, but they do not replace the current modes of data collection for the current types of survey endeavors under discussion here. The role of the new methods is, I believe, one of enhancing and extending the capabilities of the infrastructure surveys, rather than replacing them. The importance of the infrastructure surveys also suggests careful steps to evaluate and adopt the new approaches, rather than a radical overhaul of existing methods.

The Internet and Longitudinal Statistical Surveys

Mike Dennis and Bill McCready
Prepared for the NSF Workshop on Repeated Cross-Sectional and Longitudinal Surveys
Arlington, VA
March 28 – 29, 2003

Abstract

This paper will discuss several basic issues for designing recurring surveys and especially those involved in creating a representative sample for Internet surveys. It will also discuss the following topics: changes in over-time surveys that are likely to occur in the coming years; new modes of data collection that are in the offing for over-time surveys; new statistical tools that are likely to come into play; new strategies for linking external data with survey data; and optimal sampling plans.

Basic Issues

The basic issues that need to be resolved in designing representative samples for longitudinal surveys accessible via the Internet include sampling frames, attrition, bias & representativeness, response rate enhancements, non-response follow-up, and improved recruitment into longitudinal survey protocols.

Sampling Frames

There are at least six possible types of sampling frames for conducting surveys using the internet:

	Non-Random	**Random**
Single-Use	These are the banner or opt-in surveys that you see in pop-ups on the web and are similar to survey cards inserted in magazines	When there is a listing of the target population, such as all employees in a firm you can generate a random sample and direct them to a website for a survey.
Panels for Cross-sections	Solicit respondents over the internet to join a panel and then draw samples for surveys. (Common with commercial internet panel surveys. These are often weighted using RDD phone surveys.)	Recruit households using RDD phone sampling to join a panel and then provide households with devices to connect to the internet for surveys. (This provides for both cross-section surveys and longitudinal surveys if the panel is large enough - see below.)
Longitudinal Panels	Solicit respondents over the internet to join a panel and then conduct repeated surveys with the same individual. (Seldom used either by commercial or government researchers.)	Recruit households to join a panel and provide households with devices to connect to the internet for repeated surveys with the same individual.

It is difficult to construct true probability samples of households that are accessible via the internet because 43.5% of US households do not have a computer and 49.5% do not have any access to the

internet.[6] There are a variety of adaptations in use such as weighting non-random samples with data from random samples, but the only way to produce a true random sample now is to draw a probability sample of U.S. households and somehow connect the unconnected households to the Internet. Using RDD phone sampling for this initial phase produces a random sample, but the quality of the recruitment rate is very dependent on field costs.

Attrition

Panel attrition happens for many reasons and only some are preventable and not all are knowable. Among the most common causes in our experience are when all the members of a household are not in agreement about participating, when survey exhaustion or boredom sets in, and when several surveys in a row are longer than average.

The consequences of attrition range along a continuum of trivial to very serious. Low levels of attrition to a large panel do not cause much concern, while a high level of attrition to a key panel may be of great concern. It is important to understand if there are any differences in the data produced by those who leave a panel as opposed to those who stay. It is relatively easy to document demographic differences between these two groups, but it is essential to periodically conduct surveys among those who have left the panel.

There are special issues regarding attrition in a large multi-use panel since households rotate off the panel periodically in order to refresh the sample. There should be periodic "exit" surveys of these households as well.

When a project requires specific panelists to complete X number of surveys over Y number of months or years, it is worth devising and implementing attrition prevention measures. Reducing the frequency of surveys is probably useful for reducing attrition. Also, recruiting the panel by in-person field staff instead of by telephone interviewers provides an opportunity to enhance long-term panelist loyalty and for reducing initial attrition caused by failure to participate in the appropriate setup procedures (e.g., equipment setup, completion of the demographic background questionnaire, etc.) In addition, improving the household contact list helps as does acquiring extra-household locator references.

Representativeness & Weighting

It is important to make simple yet informative comparisons for representativeness to such data sources as the US Census and for comparability to gold standard surveys such as the GSS, the PSID, and the NES. A panel should closely mirror the U.S. population on key demographic, geographic, economic and social characteristics.

The challenge to representativeness from RDD-based recruitment for Internet panel development is the obvious noncoverage of non-telephone households and the relatively low cooperation rate to be expected. Knowledge Networks over time has seen a drop in the cooperation rate for telephone-based recruitment from a high of 56% in year 2000 to 36% last year, and more recently an improvement to approximately 42% resulting from recent changes in our recruitment protocol and upgrades to our interviewing staff. To do much better than that, we anticipate needing to spend much more money on respondent incentives and refusal conversion.

[6] September 2001 U.S. Census Bureau's Current Population Survey.

More tantalizing in prospect, but also more expensive, is the alternative of recruiting via in-person using field staff for improving sample representativeness, to serve as a platform for various longitudinal surveys.

For panels developed using RDD recruitment adjust the sample weights to U.S. Census demographic benchmarks to reduce error due to non-coverage of non-telephone households and to reduce bias due to recruitment non-response and other non-sampling errors.

Cross-sectional samples selected from such a panel for individual studies require appropriate sample design weights for each study based on specific design parameters. Apply within survey non-response and post-stratification weighting adjustments to the final survey data to reduce the effects of non-sampling errors (variance and bias).

Mode and Panel Effects and Data Biases

Response modes all generate specific effects and self-administered surveys over the internet address several of these. Personal interviews frequently generate social desirability effects, especially with sensitive or threatening questions. Telephone interviews frequently generate primacy and satisficing effects because of the nature of the interaction. Both formats tend to place a scheduling burden on the respondent because they do not exert exact control over the appointments. Telephone interviews cannot incorporate visual cues to impart and clarify complex stimuli, and personal interviews can only do so in limited and expensive ways.

Internet surveys do not engender desirability effects because there is no interviewer-respondent interaction. Changing and randomizing question order limits primacy and satisficing and the ability of the respondent to view the scales assists in this regard as well. Respondents can complete surveys delivered over the Internet at any time they choose reducing burden.

Response Rate Enhancements

One of the chronic problems that plague longitudinal panels is the combination of reduced response rates at the recruitment stage and attrition as the panel operates. To date, none of the several techniques employed to address these deficits have been very successful. Respondent incentives help promote an "exchange" mentality that provides respondents a reward for participation. Reducing the length or frequency of surveys can also be a reward to respondents that promotes higher cooperation. Providing respondents the ability to communicate with the survey designers also fosters a sense of mutuality and exchange. Finally, arguments for the social usefulness of the information foster a sense of civic contribution that can enhance cooperation.

To limit nonresponse bias caused by panel attrition, we have found it useful in our longitudinal surveys to collect complete locating information at the outset of the baseline survey, and then use this information for mailout surveys to 'lost' panel members for the follow-up surveys. A complement to this strategy is that we invite the sampled 'lost' panel members to respond to surveys using their own Internet connection through a secure web site that we set up.

Future Developments in Longitudinal Surveys

New Modes of Data Collection

The principal way of improving the recruitment response into probability-based internet-enabled panels is to conduct the initial recruitment by in-person interviewing with an area probability sample. This is expensive and applicable chiefly to projects requiring general population samples only, but greatly reduces the cost of longitudinal follow-up interviewing and improves the cumulative response rate.

In order to conduct environmental exposure projects, we could place digital exposure receptors in households that take advantage of the infrared ports in the interactive TV devices. We can also place bar code readers in households that take advantage of the same capability for doing dietary and nutrition projects.

There will be an expansion of integrated behavior-attitude consumer databases where we can link purchases directly to attitudes at the household level.

As more people become digitally connected there will also be an expansion of the surveys that direct respondents to assigned web sites, and some of these will provide longitudinal information.

New Statistical Tools

Having respondents view survey items on-screen allows researchers to capture more complex cognitive decisions. These data lend themselves to a variety of powerful statistical analyses including discrete choice models for economic analysis and complex SEM modeling that is an improvement over the GLM, especially in terms of error analysis.

It will also be possible to apply sophisticated data-mining programs to the longitudinal data-banks. Preserving respondent confidentiality is an utmost concern in this regard.

Linking External data with Survey Data

There are already elementary examples of way to connect household-level survey data with various types of external data including frequent-shopping data, health provider data, and insurance data. Once households grant permission to integrate these data, researchers can address powerful analytic questions with sophisticated tools. These will be particularly important in the areas of nutritional analysis and national healthcare and insurance coverage.

Optimal Sampling Plans

We need to develop new protocols to increase the recruitment rate to internet-accessible panels and improve the overall response rate. One strategy is to use the area probability sampling frame with in-person interviews to place interactive devices in households. Although this initial placement-recruitment cost is high, follow-up costs would be low compared to current models. There is also some evidence that this would improve data quality as well.

Apply Findings from Improved Usability Research

We need to refine and apply the findings from current research about on-screen movement and screen navigation to improve usability and panelist cooperation. We need to do more work with cognitive labs and focus groups to test these protocols and refine them for specific survey projects. Finally we need to initiate within-family ethnographies about how families actually use various interactive devices in their households. We know too little about the human-screen interface and we need to learn much more as these technologies develop.

Non-Response Follow-up

Longitudinal projects generate a lot of data about non-response, if properly designed at the outset. We can compile existing studies that have incorporated a non-response follow-up of some sort. This would enable us to know a good deal more about the sociological, economic, political and psychological natures of the non-respondents and their rationales for not responding. This will go a long way toward allowing us to truly analyze nonresponse bias.

Preparing for Broadband & Wireless

The US is behind many countries in terms of broadband and wireless penetration, but it is likely to gain ground quickly. We need a program of research to examine alternate platforms and evaluate them as to their “research usefulness.”

Making a Case for Repeated Cross-Sectional Surveys: Perspectives on the General Social Survey (GSS)

Barbara Entwisle
Department of Sociology and the Carolina Population Center
March 2003

Researchers are sometimes quick to dismiss cross-sectional surveys, but for some purposes, they may be superior to panel designs. In these comments, I want to consider cross-sectional surveys, with particular reference to the General Social Survey. I would like to highlight the advantages of a nationally representative cross-sectional personal interview survey to evaluate the quality of samples in other surveys, to develop new samples, for the analysis of trends, and to describe rare populations. Longitudinal panel surveys have many strengths. In celebrating these strengths, however, it is important not to overlook or understate the comparative advantages of other longitudinal designs and data sets.

Let me begin with a very brief overview of the General Social Survey, which I will use to illustrate my remarks. The GSS is an "omnibus," personal interview survey of English-speaking adults living in U.S. households conducted by the National Opinion Research Center (NORC) with James A. Davis, Tom W. Smith, and Peter V. Marsden as principal investigators. The mission of the GSS is to make timely, high-quality, scientifically relevant data available to the social science research community. For the purposes of the points I want to make today, key features of the GSS include a nationally representative sample, personal interviews, and the replication of core attitude items over substantial amounts of time. The first survey took place in 1972, and since then, surveys have been fielded every year or two (1972-1978, 1980, 1982-1991, 1993-1994, 1996, 1998, 2000, 2002). Over the three decades of the GSS, more than 38,000 respondents have answered over 3,260 different questions.

As a repeated cross-sectional survey, the GSS has some important advantages to offer. To begin, the GSS has an important role to play vis-a-vis the evaluation of the samples used in other surveys. Because of its national scope, the use of a probability design, and personal interviews as a mode of data collection, in many ways, the GSS sample is a “gold standard” against which other surveys can be judged. Consider a comparison with the samples used in longitudinal panel surveys. Attrition is a major problem in these surveys. For example, in the Panel Study of Income Dynamics, which began in 1968, only half of the original sample was still part of the panel twenty years later (Fitzgerald et al. 1998: 254). Attrition is usually evaluated against the census or the Current Population Survey, but the GSS may also have a contribution to make, especially if the evaluation is to be conducted with respect to data not typically collected in the CPS or census (e.g., religiosity). The consequences of attrition for analysis depends on the goals of the analysis and the particular variables involved (e.g., Fitzgerald et al. 1998b).

Or, consider a comparison of the GSS with surveys using other modes of data collection. Cost differences between personal interview and telephone surveys are large. Personal interviews may be the desired mode, but not all studies can afford them. The GSS is useful as a standard against which to judge the consequences of choosing other modes of data collection. For example, Woodberry (1998) compared GSS data on church attendance with answers to the same question asked in a telephone survey. He found that the telephone survey estimate of church attendance was almost 15 percentage points higher than the GSS estimate. Woodberry attributed the difference to the greater representation of non-churchgoers in a personal interview survey. A similar contrast emerged in an analysis of recent trend in no religious preference (Hout and Fischer 2002), a trend that was much more evident in the GSS than Gallup polls, which are based on a telephone interview. Clearly, the GSS has a role to play in the evaluation of the samples used in other kinds of surveys. It is not perfect, of course. It would be an even better yardstick if its sample were expanded to include non-English-speaking groups, especially Spanish speakers. It would

also be better if the slight recent decline in response rates is checked or reversed. Response rates, which have varied generally between 75 and 80 percent over the three decade history of the project, have been at their lowest level of 70 percent during the last two rounds (2000 and 2002).

The quality and representativeness of the GSS sample is also key to its use in generating samples for which frames may be unavailable or prohibitively expensive to construct. GSS respondents can be used to create nationally representative samples of dyads and of groups or organizations to which respondents belong, such as religious congregations, work establishments, and voluntary associations. As Chavez, Marler and Hadaway (1993) demonstrated, enumerating congregations directly can be a challenge, even if the enumeration is restricted to Protestant churches within a single county. Using the GSS to generate a sample of congregations is much more cost-effective. The 1998 National Congregations Study used this strategy. Respondents in the 1998 GSS who attended religious services more often than "never" were asked the name and location of their religious congregations. The 1998 National Congregations Study is based on the sample that was generated (Chavez 1998). The 1991 National Organization Study pioneered this strategy in its sampling of work establishments. A 2002 National Organization Study was fielded this past fall. It is too early to know what trends might be revealed in a comparison of the 1991 and 2002 NOS data.

The replication of items over waves of the GSS, and comparability in survey procedures over those waves, makes possible a variety of trend analyses. There are many recent examples of this. Hout and Fischer (2002) used GSS data to document the recent upturn in no religious preference and to assess hypotheses about the reasons for it. The length of the GSS series and the periodicity of the surveys were important to establishing both the absence of trend up to 1990, and the presence of a trend thereafter. As another example, Loftus (2001) used GSS data to describe the pattern of change in attitudes towards homosexuality over 25 years. In both examples, the use of common wording and response formats across the waves of the survey was key. The length of the series provides some purchase as to whether change is due to processes of cohort replacement or individuals have changed their attitudes. With a shorter series, it would be much more difficult to make an inference about this. Replication of items over waves of the survey is important to the analysis of trends. With respect to design, there is a strain between making improvements in question design, and losing comparability, and maintaining question wording and format over time, thereby supporting trend analysis.

Multiple waves of data, in combination with replication of items, can be valuable even when the goal is not to describe or analyze change. When aggregated, repeated cross-sections can be used to identify and describe relatively rare populations, provided that the samples are representative. An excellent example of this can be found in the work of Black and his colleagues (2000), who used GSS data from 1988-91, 1993, 1994, 1996 along with the 1992 National Health and Social Life Survey to document the prevalence of homosexuality given different definitions. Gay, lesbian, and bisexual respondents are relatively few in number in any given wave of the GSS. When the waves are aggregated, however, the numbers become large enough to analyze in a meaningful way. There is much potential in this use of GSS data. If the sample for each wave of the GSS is representative, then the sample obtained by combining waves is representative.

I have heard the GSS characterized as an old-fashioned attitude survey. This is incorrect. GSS modules use innovative techniques to explore topics of great interest to the field. Analyses based on the topic modules often are published in top journals such as the American Sociological Review. The GSS is also part of the International Social Survey Program (ISSP) and includes items that allow direct cross-country comparisons. In my remarks, however, I have focused on the design of the GSS rather than its innovative content. I have focused on the benefits of a representative cross-sectional personal interview survey that includes replication of core items. Using examples from the recent literature, I have shown how these elements of the GSS design make it possible to use the GSS to evaluate the samples used in other surveys,

to develop new samples, to study trends, and to describe relatively rare populations. There are many applications in which a longitudinal panel survey is superior to a repeated cross-section design. I have tried to show that there are some applications in which a repeated cross-section design is superior to the longitudinal panel survey. But each design is strengthened in the presence of the other. From a field perspective, both kinds of surveys are necessary and desirable.

References

Dan Black, Gary Gates, Seth Sanders and Lowell Taylor, 2000, "Demographics of the Gay and Lesbian Population in the United States: Evidence from Available Systematic Data Sources (in Gay and Lesbian Populations)." *Demography* 37: 139-154.

Mark Chaves, 1999. "Religious Congregations and Welfare Reform: Who Will Take Advantage of "Charitable Choice?" American Sociological Review 64: 836-846.

C. Kirk Hadaway; Penny Long Marler; Mark Chaves, 1993. "What the Polls Don't Show: A Closer Look at U.S. Church Attendance." American Sociological Review 58: 741-752. Mark Chavez, Marler and Hadaway.

John Fitzgerald, Peter Gottschalk, and Robert Moffitt. 1998a. "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics." *Journal of Human Resources* 33: 251-299.

----------. 1998b. "The Impact of Attrition in the Panel Study of Income Dynamics on Intergenerational Analysis." *Journal of Human Resources* 33: 300-344.

Jeni Loftus, 2001. "America's Liberalization in Attitudes toward Homosexuality, 1973-1998." American Sociological Review 66: 762-782.

Michael Hout and Claude S. Fischer. 2002. "Why More Americans Have No Religious Preference: Politics and Generations." *American Sociological Review* 67: 165-190.

Robert D. Woodberry, 1998. "When Surveys Lie and People Tell the Truth: How Surveys Oversample Church Attenders." American Sociological Review 63:119-121.

Trends in Survey Costs and Key Research Needs in Survey Nonresponse

Robert M. Groves
University of Michigan
March 23, 2003

1. Introduction

In this age of transaction data produced on almost all events, surveillance recording of near universal human movement, and administrative databases easily analyzed, have surveys lost their value for social science investigations? Are the large costs of household and establishment surveys justified relative to the small costs of a smart social scientist observing and interacting with a small group of persons? Are large-scale surveys merely atheoretical, descriptive information tools that, by definition, cannot be cutting-edge tools of science? Will nonresponse be the death of telephone surveys? How can standardized survey measurement ever expect to offer the richness of insight into social processes obtained from deep, unstructured, dynamic investigations?

With no assurance that diverse social groups are homogeneous, how can social science theories be subjected to useful tests without probability sampling of large populations? How can social knowledge cumulate if measurements underlying the knowledge cannot be replicated – over time, over social groups, over study conditions? Given that our understanding of the dynamics of poverty, of job search and stability, of partnering and family formation, of retirement planning, of major shifts in racial attitudes, are based on surveys, what can replace them?

These and other questions are common challenges to quantitative social science, as it moves into the new century. What will the major tools of social science data collection be in the coming years? Will formal theory construction, qualitative observational studies, and secondary analysis of administrative data be the only tools of social science?

In response to the organizers of the workshop, I address two issues:

a. the rising costs of scientific surveys; and
b. the importance of scientific activities unraveling the puzzle of nonignorability conditions of survey nonresponse.

However, I should lay my premises related to some of the questions above prior to launching on those two topics:

a. I speculate that in the medium-run future, process-generated data (transactions, administrative data, etc.) will not be sufficient to address key questions generated by social science;
b. I believe that the act of invention of measurements is a step in the scientific method that helps to purify theoretical concepts and linked hypotheses that form theories; hence, survey measurement benefits theory evolution;
c. I consider the sample survey to be a blunt instrument, which by itself has a role in testing of well-articulated theories on large populations
    1. through standardization it fosters replicability;
    2. through rigorous selection of subjects it permits inference to important populations within measurable error;
    3. through enhanced attention to nonresponse and measurement error, it forces explicit attention to its weaknesses in ways that much other research does not;
d. I look forward to developments that blend qualitative and contextual methods into the framework of the sample survey;

e. I value the clear evidence that the sample survey is the hub around which empirical social science grew over the past few decades into a central body of techniques and perspectives with effects far beyond the academy.

2. Key Features of Costs Increase in Surveys

In recent years, three forces, linked to one another, have affected the status of scientific surveys:

a. the U.S. household population has experienced large scale changes, leading to more effort required to obtain response rates;
    i. the relative size of single person households has risen (such households require more attempts by interviewers to contact);
    ii. the investment of households in access impediments (telephone answering machines, caller ID, call butler, locked multi-unit structures, walled subdivisions) has increased;
    iii. the rate of reluctance to accept survey requests has increased;

b. the technical infrastructure of surveys has become relatively more expensive;
    i. large scale use of networked computers, with sophisticated questionnaire software, require large investments that must be amortized;
    ii. pressures on the relatively low-skilled labor market traditionally used for survey interviewers from the new technical requirements have led to large training costs and/or large attrition rates of interviewers;

c. the resulting higher costs have led to funding consortia with each member requiring subsets of measures;
    i. longer, more complex interviews have resulted;
    ii. computer-assisted data collection has permitted scientists to implement unusually complex instruments (some so complex that no two respondents get exactly the same questionnaire);
    iii. mixes of questions, physical measurements of respondents, and interviewer collection of observations at the household are more common;

These three forces lead to longer interviews being requested by interviewers with less experience from persons who are more sensitive to the burden of the request. The National Science Foundation has been replaced as a funder of major social science data collections by the National Institutes of Health. However, many social science questions are not able to be morphed into health-related questions.

One common metric for costs is the number of interviewer hours that are required to obtain an interview. There are some empirical comparisons over two editions of surveys with minimal changes in design. The table below shows large increases in costs over the late 1990's. The first two are based on field comparisons; the last, on a deviation between a budget and actual. It appears that cost inflation is sometimes due to efforts to maintain cooperation rates attained in the prior edition of the survey, in the face of greater reluctance.

Interviewer Hours per Interview on Selected Repeated Cross-Section Surveys

Survey	Year	Hours/I'w
General Social Survey	1998	~8
General Social Survey	2000	~12
National Comorbidity Survey	1993	8.5
National Comorbidity Survey	2001	12.4
National Survey of Family Growth (budget)	1999	8.9
National Survey of Family Growth (actual)	2002	11.3

3. Research in the Causes and Consequences of Survey Nonresponse

Quantitative social science has largely taken the fact that complete measurement of a probability sample of persons is a desirable feature of sample surveys, and inferred that higher response rates are always better than lower response rates. While the elimination of nonresponse error of statistics can easily be demonstrated with 100% response rates, there is little statistical guidance possible about the relationship between nonresponse rates and nonresponse error. In the last three years, three important papers have challenged the notion that the linkage between nonresponse error and nonresponse rates is straightforward (Keeter *et al.*, 2001; Curtin *et al.*, 2001; Merkle and Edelman, 2002).

There is growing evidence that traditional viewpoints on the nexus between nonresponse rates and nonresponse error were naïve. Just reviewing the typical lay questions about this helps to illustrate this:

a. what response rate is needed for a survey? (Answer: nonresponse error is a property of individual statistics not surveys as a whole; thus the relevant nexus between nonresponse rates and error lies at the statistic level)
b. is there a minimum response rate that is needed for a survey to be useful? (Answer: no; characteristics of nonrespondents on the key statistic of interest is an issue as well as rates)
c. don't nonrespondent pools consist of the same kinds of person across all surveys? (Answer: no, there appear to be variations across survey types; e.g., race correlates vary across survey topics)
d. in short, do higher response rates mean lower nonresponse error? (Answer: not always)

There is growing evidence that many participatory decisions made to survey requests are shallow, heuristic-guided acts; that reactions to survey features vary across sample persons, and that the salience of a feature varies over survey introductions. Statistically and logically it can be shown that error arises from nonresponse to the extent that causes of the participatory decision are connected to survey statistics

computed. (For example, survey-based measures of at-home behavior are empirically quite sensitive to noncontact nonresponse error.)

The unanswered scientific questions critical to guiding NSF investments in further survey data collection are:

a. what are the necessary ingredients of survey design, sponsorship, topic, collection mode, and target population to have nonresponse rates be linked to nonresponse error?
b. what design features (or mix of features) might be introduced to dampen the linkage between nonresponse rates and errors?
c. what design features might be introduced to provide scientists with indicators of nonresponse error?

Answering these questions is a central challenge of the field of survey methodology in the near future; funders of survey efforts should support efforts to obtain these answers.

References

Curtin, Richard, Stanley Presser, and Eleanor Singer (2000) "The Effects of Response Rate Changes on the Index of Consumer Sentiment," *Public Opinion Quarterly*, 64:4: 413-428.

Keeter, Scott, Andrew Kohut, Carolyn Miller, Robert Groves, and Stanley Presser, (2000) "Consequences of Reducing Nonresponse in a Large National Telephone Survey", *Public Opinion Quarterly*,

Merkle, Daniel and Murray Edelman (2002), "Nonresponse in Exit Polls: A Comprehensive Analysis", Chapter 16 in Groves, R., D. Dillman, J. Eltinge, R. Little, *Survey Nonresponse*, New York: Wiley, pp. 243-258.

Longitudinal Surveys and Social Science Analysis

Guillermina Jasso
New York University
March 2003
Prepared for the Workshop on Repeated Cross-Sectional and Longitudinal Surveys, sponsored by the National Science Foundation, March 2003.

There are many reasons for doing longitudinal surveys. In this note I discuss three related reasons, two of them rooted in substantive aspects of sociobehavioral science, the third rooted in the statistical procedures of empirical work. In general, all research asks a question and seeks to answer it. The nature of the question dictates the portfolio of protocols that can provide answers, including repeated cross-sectional surveys, longitudinal surveys, and experiments, and including as well hybrid protocols such as combinations of experiments and surveys. Given that time is a fundamental dimension of the world we live in, collecting information over time would seem to be a natural enterprise. Of course, there are special challenges and costs associated with longitudinal surveys, and we shall learn much about them in subsequent sessions. It is important to expand the kit of cost-saving tools, so as not to deprive questions which require longitudinal designs of appropriate data.

1. *Three Reasons for Longitudinal Surveys*

Reason 1. *Sociobehavioral Phenomena and Processes Occur over Time*

Sociobehavioral phenomena and processes occur over time, and many of them cannot be fully understood without explicit dynamic designs. This is particularly the case when there is movement in all three elements from the age-period-cohort trio. Here are some examples.

Assimilation of immigrants. Assimilation occurs over time, and is shaped by characteristics of the immigrant at entry and by conditions in both origin and destination countries at the time of immigration as well as at subsequent periods. Becoming American is a journey in the progressive embrace of core American values. Understanding assimilation thus requires following representative samples of new immigrants over time. Consider the core value of the dignity and equality of all humans – an idea that becomes both more sharply understood and more powerful with every passing year. New immigrants may think in their minds that they already embrace equality, but they may not realize for some time how deeply inegalitarian they may be – in the language they use to address diverse others, in the boundaries they erect for choosing spouses and friends, in the way they raise their children. Ten years down the road they may realize that they are replicating the very stratification structure they sought to leave behind and that it is incompatible with their burgeoning American vision. The point of discovery – for example, discovery of the inegalitarian views embedded in their native language, such as gender and *tutoiement*/*vouvoiement* elements – may occur at different durations since entry, and may differ for immigrants of different cohorts or who immigrated at different ages or by different pathways. To examine systematically this journey, it is necessary to collect data over time.

Human development. Many sociobehavioral processes involve growth and development--cognitive development, emotional development, psychosexual development, moral development. Indeed, human development, in all its facets, lies at the heart of the human experience. How a person's mind grows, and information is acquired, and health habits formed, and new skills perfected, all while confronting often serious obstacles; these require research over time. Similarly, moral development occurs over time. For example, in the study of justice, an important question focuses on how a person's ideas of justice change over time and, concomitantly, how the discrepancy between true and disclosed ideas of justice changes over time.

Attachments. A person's attachment to the labor force, to religion, to a hobby, or a spouse is also a process that unfolds over time. Attachments obviously include love, as in the celebrated question, "How does love grow?" which set medieval minds to thinking about continuous dimensions.

Effects of ideas about reality on behavior. A class of important questions focuses on how ideas about the way the world works ("beliefs", for short) affect behavior – for example, how beliefs about the effects of diet, exercise, and smoking on life expectancy influence diet, exercise, and smoking behavior; how beliefs about the effect of spousal schooling similarity on marital happiness influence choice of spouse; how beliefs about the effect of school, major, and grades influence decisions and behavior among college students and their parents. While some of these relationships may be contemporaneous, others are time-linked in diverse ways, for example, via a lag structure, or via a link between strength of the effect of the belief and onset of the belief and/or duration of the belief.

Effects of normative ideas on behavior. A related class of questions focuses on how normative ideas influence behavior, for example, how ideas of justice influence voting behavior, allocation decisions, consumer choices, charitable contributions. These effects also may exhibit a lag structure or other elaborate intertemporal form whose precise understanding requires longitudinal data.

Reason 2. ***Individuals May Be Unique***

Longitudinal data can help address the longstanding questions whether or how individuals are unique (and, concomitantly, whether sociobehavioral nature is governed by universal laws, and, if so, whether the laws of sociobehavioral nature are quantitative or qualitative) – questions which in one form go back to the earliest discussions of free will, and in another, to the late-19th-century and early-20th-century speculations of Dilthey, Rickert, and Windelband.

Person-Specific Regressions. Consider the usual regression equations of earnings, voting, or other individual-level behaviors: Humans may show their uniqueness by having idiosyncratic intercepts and/or idiosyncratic slope coefficients, as in the following form:

$$Y_{it} = \beta_{0i} + \sum \beta_{ki} X_{kit} + \varepsilon_{it}. \qquad (1)$$

To assess idiosyncratic intercepts, it is sufficient to have two observations per person. To assess idiosyncratic slopes, many more observations per person are required; for example, in a classical fixed-parameters approach, the relevant calculation would be the degrees of freedom (the number of data points minus the number of parameters estimated).

To illustrate, consider the Mincer earnings function, a workhorse in contemporary social science. Suppose that the parameter that converts schooling into earnings is not only responsive to cohort or geography or the state of the economy but is also an individual signature. Similarly, the parameters that convert experience into earnings may be person-specific. Estimating such person-specific parameters requires data over the life course. The great American panels (the early NLS panels and the PSID) are nearing the point where estimation of person-specific equations becomes a reality.

If the coming work shows that numerical parameters are person-specific (but of the same sign), then the new answer to Dilthey, Rickert, and Windelband may be that (1) individuals are indeed unique (in a specified sense), and, simultaneously, (2) there may still be universal laws (of a specified sort, stopping just short of being numerical laws).

But longitudinal data are not always needed to establish uniqueness. The uniqueness of individuals may also be established via factorial surveys. These have the characteristic that although they are cross-sectional they involve repeated measures. Obtaining multiple ratings from each respondent enables estimation of person-specific regressions and testing of the full complement of homogeneity hypotheses. Use of this protocol has in every case found uniqueness, further supporting the idea that though universal laws governing sociobehavioral nature may indeed be mathematical, they will not always be numerical.

Reason 3. ***Control Biasing Effects of Unobserved Heterogeneity***

A third reason for carrying out longitudinal surveys is to control for time-invariant unobservables which may be correlated with explanatory variables. Even if the process or relation under study is not inherently dynamic, even if there is no interest in estimating person specific parameters, longitudinal data are still useful for controlling the biasing effects of unobserved heterogeneity, via estimation of fixed-effects models.

2. ***Two Remarks***

Remark 1. ***The Question Dictates the Design***

As noted above, the question dictates the design. Some questions may dictate a one-shot survey, others a repeated cross-sectional survey, still others a longitudinal survey. Moreover, it may happen that a full picture of a process may require a combination of designs addressing related questions.

To illustrate, there is a tradition of measuring income (or wage, or earnings) inequality in repeated cross-sectional samples from a population defined by geographic residence. This design answers an important question; but it overlooks other questions. A full understanding of stratification processes would require posing further questions, for which initial steps in the relevant designs would include (i) measuring inequality within each of several cohorts, at different points in time; and (ii) measuring inequality over the life cycle for each person in a sample.

Remark 2. ***Hybrid Designs***

A variety of hybrid designs may prove useful for addressing particular questions. We have already mentioned the combination of factorial surveys and cross-sectional surveys. Here we briefly discuss two further kinds of combined experiment-surveys.

Survey participation as experimental treatment. There are longstanding questions concerning the effects of participating in a survey. The New Immigrant Survey (NIS), whose first-cohort baseline round goes in the field in June, provides wide scope for studying some of these issues. Consider the decision to naturalize (for which most immigrants will become eligible five years later). It will be possible to contrast naturalization rates among two groups: nonsampled immigrants; and sampled immigrants who participate. In principle, it would be possible to further contrast a third group, namely, sampled immigrants who refuse to participate, and as well to incorporate further specificity on attriters.

Future NIS cohorts can be studied in more elaborate ways. For example, a large probability sample can be drawn, with one randomly chosen subset invited to participate, a second randomly chosen subset invited to join a panel (i.e., be tracked) but not asked any questions, and a third randomly chosen subset not contacted for several years (but tracked via administrative records) and interviewed for the first time at varying durations since admission to permanent residence. Several years later the various subsets can be compared with respect to such behaviors as learning English, entering the labor force, religious attachment, health behaviors, and so on.

Obviously, more elaborate designs are possible, for example, randomizing interviewer nativity and language of first approach.

Surveying groups in a natural experiment. In immigration research, a perennial question is, "What is the pure effect of immigration?" The pure effect of interest may be on the immigrant him or herself, or on the immigrant's children, or on the origin country. This type of question is difficult to answer because immigrants are self-selected. A new design currently being formulated takes advantage of a natural experiment in U.S. visa allocation, namely, the diversity visa lottery. By following a sample of lottery winners and a sample of lottery losers, over time, it will be possible to assess the true and pure effects of immigration.

Nonresponse Issues in Longitudinal Surveys

Dan Kasprzyk
Mathematica Policy Research

Background

The issues associated with the design and analysis of surveys across time were discussed in a number of venues during the mid 1980's and early 1990's. The Social Science Research Council, with funding from the National Science Foundation, organized a "Working Group on the Comparative Evaluation of Longitudinal Surveys" chaired by Bob Boruch (Boruch and Pearson 1985). The working group fostered an extensive discussion of the advantages and disadvantages of longitudinal surveys (Boruch and Pearson 1988; Pearson and Boruch 1986; Pearson 1989; Duncan and Kalton 1987). The Section on Survey Research Methods of the American Statistical Association organized an international conference on panel surveys in late 1986 that resulted in an edited volume (Kasprzyk, Duncan, Kalton, and Singh 1989). The manuscripts in this book highlighted methodological issues in the design, conduct and analysis of large-scale longitudinal surveys. The U.S. Census Bureau initiated a large-scale longitudinal survey, the Survey of Income and Program Participation, and even though the survey was in its infancy in terms of its operations and output, the Committee on National Statistics of the National Academy of Sciences conducted a formal evaluation of its design and operations (Citro and Kalton 1993). In 1992, Statistics Canada dedicated their annual symposium to the "Design and Analysis of Longitudinal Studies."

The interest continues today in many places and on many topics. There are international surveys of economic well-being, following the Panel Study of Income Dynamics model; the National Institute on Aging sponsors the Health and Retirement Study (HRS); there are large-scale longitudinal surveys of medical expenditures, such as the Agency for Health Care Research and Quality's Medical Expenditure Panel Survey, longitudinal surveys of Medicare beneficiaries, the Bureau of Labor Statistics National Longitudinal Surveys, and longitudinal surveys of children of different age cohorts, sponsored by the National Center for Education Statistics. Interest has not diminished over the last 20 years or so and neither have the problems.

Nonresponse and sample attrition are important topics in the survey research literature that provide substantial research opportunities for survey analysts. Of course, with longitudinal surveys and their multiple interviews, the issue on nonresponse is exacerbated and complicated when compared to cross-sectional nonresponse. Nonresponse, itself, has received a great deal of attention over the last twenty years, no doubt because it is frequently used as an overall indicator of the quality of a survey. Unfortunately, it often is the only quantitative quality indicator available and as a result the indicator takes on a high-level of importance. Its virtues are that it can be calculated fairly easily in most circumstances and that most people think they know what it means and how to interpret it—unfortunately indifferent to the fact that bias in the estimate is also a function of the differences between the respondents and nonrespondents on the variables being estimated.

Complications of Longitudinal Surveys

Longitudinal surveys complicate the problem of nonresponse in several ways. As a general rule, most think of the missing data patterns in longitudinal surveys in terms of monotone nonresponse - - where a sample unit (person/ establishment) stops participating in a survey at one data collection cycle and continues to be a nonparticipant for each collection cycle thereafter; that is, once a sample unit is a nonrespondent, it remains a nonrespondent. A longitudinal survey's data collection procedures, however,

need not take that approach. Rather, survey procedures can stipulate return visits to nonparticipating sample units for the purpose of converting them to respondent status. In this case, the patterns of missing responses can become quite complicated, leading to non-monotone missing data (wave) patterns - - that is, the patterns that occur when sample units are respondents for one or more data collection cycles, become nonrespondents for one or more cycles only to return again as a respondent. This situation results in the survey director having to make decisions about the appropriate treatment of wave nonresponse over the course of the longitudinal study. On the other hand, the amount of information available to compensate for nonresponse in these situations can be overwhelming. These data, however, should improve the nonresponse adjustments (in the reduced bias sense). All of this is but a short prelude to say that the missing data problem in longitudinal surveys is complicated and is not solely limited to the problems of attrition, defined as the monotone nonresponse described above.

Nonresponse in sample surveys and its reduction has been an important topic in survey research. Nonresponse is a significant and growing problem in all surveys and it shows no sign of getting any better. Even though there are many successful situations in which surveys obtain good response rates (possibly the topic, incentives, or the survey sponsor) the problem will be with us for a long time. And, of course, in the longitudinal survey context, the issue is not just obtaining a good response rate one time; the issue is maintaining both a good response rate at each data collection cycle and maintaining the representative nature of the sample over the course of the survey.

One can address the attrition issue and the ways to reduce bias in longitudinal surveys in several ways, though not mutually excusive: 1) through the data collection process; and 2) through adjustments to compensate for nonresponse. I think most would agree the best way to address the issue of nonresponse is at the data collection stage. That is, the importance of conducting methodological research to test and develop "best" methods and procedures to encourage survey participation cannot be overemphasized. In general, the strategy is to consciously address design issues so that attrition is not exacerbated. In that respect, it is important for longitudinal survey directors to develop a cost benefit analysis on important data collection issues and determine appropriate strategies with the available financial resources.

Attrition Reduction through Data Collection

Longitudinal survey directors ought to invest resources to understand the components of nonresponse. Knowledge and understanding of the problem will result in a better use of survey resources. If refusals are the primary concern, research aimed at identifying successful refusal conversion methods is important. Is more effort required in the field necessary, possibly extending the data collection period? Do the interviewers need additional training? Are the survey materials helpful to the interviewer? Do the sample cases understand the importance of their individual response? Do incentives encourage response? In the rush to field a survey, it is easy to not pay attention to some of the mundane issues of data collection. More attention to research methods to inform critical field decisions is informative and helpful. However, if nonresponse is due to an inability to find and establish contact with the sample case, then more resources ought to be put into tracking or tracing previous interview respondents. This is exceptionally important at the start of the survey, because if a good follow-up strategy is not implemented, the cases are usually lost forever.

Attrition Reduction through Data Collection: Following Rules

A related data collection decision is concerned with the decision rules on who ought to be followed in a longitudinal survey. There are several ways to look at this. First, in a national survey like the Survey of Income and Program Participation (Citro and Kalton 1993) that follows all adult individuals enumerated at a sample address, one must decide whether the rules for following sample cases include following them anywhere in the United States. When personal visit interviews was the SIPP's primary method of data

collection, sample cases were only followed if the forwarding address were within 100 miles of a SIPP primary sampling unit. So cases outside the 100 miles were lost to nonresponse. This is not an issue if telephone numbers are available and telephone interviews allowed.

Similarly, when sample cases are institutionalized, whether for health or criminal justice reasons, one must decide whether the status of the individual sample case is verified at each data collection cycle and once again followed if he/she leaves the institution. Sample cases will be lost to attrition if the interview status is not verified and updated.

If interest in maintaining high recontact and response rates is paramount, an investment in data collection procedures that maximize following sample cases to determine eligibility and obtain interviews seems very important.

Attrition Reduction through Data Collection: Nonrespondent Follow Up

The decision on whether to follow-up nonrespondents in subsequent waves of data collection is important for maintaining response and coverage in a longitudinal survey. Dropping cases at the first refusal leads to the monotone patterns of response as described earlier. A decision to try to convert past refusals to response status can lead to significant additional interviews. In a small three wave study of high students, Mathematical Policy Research was able to obtain interviews with about 10% of the nonrespondents to the initial interview in one or both of the two subsequent interviews. In the early years of the SIPP, the 1985 – 1991 SIPP panels, between 4% and 5 % of the sample cases had non-monotone patterns of missing data – with only one interview missing (U.S. Bureau of the Census 1998). Obviously, large number of cases can be excluded from the survey solely based on survey design decisions.

Decisions related to improving nonresponse follow-up often lead to more aggressive tactics to obtain an interview. Abreu and Winters (1999), for example, offered incentives to nonrespondents from the previous wave in an effort to bring them back into sample and found the conversion rate was significantly improved by the use of incentives. The Survey of Program Dynamics developed a program to offer incentives to nonrespondent households from earlier interview waves as well as the current wave reluctant respondents and found that "incentives are effective in improving data quality." (Kay, Boggess, Selvavel, and McMahon 2001).

Methodological research that develops and tests approaches to improve response rates, including the use of incentives, ought to remain high on the list of ways to address nonresponse; unfortunately, the methods are not likely to be revenue neutral.

Attrition Reduction through Data Collection: Time in the Field

Operational decisions, such as the amount of time in the field available for refusal conversion may affect response rates. An extension of the data collection period will result in a higher response rate than had the collection period not been extended. This is a potentially expensive strategy with sometimes only marginal response rate improvements. During the early years of the SIPP program, discussions occurred on several occasions to extend the closeout deadline to improve response rates. Complications in the field and processing operations usually won the argument over potentially better response rates. Response rates can usually be improved with longer field times and more effort, but the cost to obtain such interviews increases.

Attrition Reduction through Data Collection: Targeting Refusal Conversion

In continuing and repeated surveys, data can be developed about sample cases that do not cooperate or reluctantly cooperate in the survey. In this way, a survey director might be able to target refusal conversion efforts where they are most likely to be successful. This strategy tries to use the data collection budget in the most effective way to maintain or improve response rates over time. This approach might make sense if adequate models can be developed; a substantial amount of research using multiple data sets is appropriate.

Reducing the Effects of Attrition through Survey Design

If one of the goals of the survey program is the production of cross-sectional estimates over the course of the longitudinal survey, then maintaining a representative cross-sectional sample is critical. There are several ways to approach this. Through internal sample regeneration, that is to include in the sample, cases associated with the original sample, where a rule of "sample membership association" is defined; for example, if individuals at a sampled address move to a new address, then all persons at the new address become part of the estimation scheme for cross-sectional estimates. Estimates from the SIPP are handled this way. The prime purpose of including the "new" person(s) in the sample is usually to describe the social and economic context of the individuals in the original sample. An important issue that arises after adopting such procedures is whether to include these "non-sample " cases in the panel after they no longer live with the sample individual. A few years, Graham Kalton and Rod Little addressed this issue for the NSF PSID Board of Overseers.

Other rules can be established that account for new entrants to a panel. For example, the survey can choose to include babies born to sample panel members after the start of the panel, and can identify sample individuals who have not attained the age to be interviewed, follow them, and interview them when they reached the appropriate age. In this way the sample accounts for new entrants to the survey.

Another approach to maintaining a representative sample is the use of a rotating panel design, in which two or more panels of sample are in the field at the same time, although each panel was implemented at different times. The Current Medicare Beneficiary now retires one third of the existing sample (the oldest panel) each year, and a new panel is selected to replace it (Lo, Chu, and Apodaca 2002). For over a decade the SIPP's design required beginning a new panel sample each year, and the National Academy of Sciences Panel recommended the design be retained (Citro and Kalton 1993). This is a useful strategy if the goals of the survey include both cross-sectional and longitudinal estimation. In a similar way, the National Center for Education Statistic (NCES) "freshens" its student cohort samples to allow for improved cross-sectional point-in-time estimates. The Panel Study of Income Dynamics (PSID) added a Latino supplemental sample in 1990 to improve the coverage and reliability for Latinos and Latino subgroups (Hill 1992).

Reducing the Effects of Attrition through Statistical Adjustments

Most approaches described above have important multi-year cost implications. Aggressive nonresponse follow up and tracking efforts, offering incentives even if targeted to certain sample cases, and decision rules that track currently ineligible sample cases require financial resources, and, in truth, compromises are made. Even with substantial financial resources, however, panel attrition will occur. The goal with the approaches described above is to minimize attrition to the largest extent possible. In the final analysis, though, panel attrition remains and methods to compensate for the missing information must be developed. The development of nonresponse adjustments for longitudinal surveys has engaged a wide range of research interest, but no relatively simple answer has emerged.

The SIPP program, for example, has had high levels of attrition and has tried conscientiously to reduce those levels. During the early 1990's the Census Bureau supported research on the effects of sample attrition on estimates from SIPP and on methods for developing longitudinal weights that adjust for attrition. Results were reported at the 1994 meetings of the American Statistical Association. Hill (1994), Rizzo, Kalton, Brick, and Petroni (1994), An, Breidt, and Fuller (1994), and Folsom and Witt (1994) studied the problem with the idea of reducing attrition bias in the SIPP cross-sectional and longitudinal estimates. The Census Bureau approach to explore alternative methods is practical and helpful. One important conclusion comes from the Rizzo et al. (1994) paper - - namely, that alternative weighting methods produced similar estimates and that there was no evidence that any specific method for compensating for panel nonresponse was more effective than another. This somewhat disappointing result suggests several approaches.

First, continued sponsorship of research related to methods to compensate for panel nonresponse should continue to be a priority. Obviously, methods and models to correct for attrition bias are critical and further study is warranted. Second, however, is the need to pay considerably more attention to the kinds of information available for use in nonresponse adjustment procedures. In addition to sampling frame variables, variables from the interviews in which a response was obtained are probably not used as much as they should be. Furthermore, auxiliary variables do not need to be a result of the survey data collection. Recent work by Vaughan and Scheuren (2002) illustrates the benefits of nonresponse compensation using variables from an administrative record system. In this application, earnings records from the Social Security Administration were used to adjust for attrition in the Survey of Program Dynamics (SPD) with some success in measuring net change in earnings and variables related to earnings. Despite the difficulties of obtaining access to administrative data, research that expands the set of auxiliary variables to related administrative records is desirable.

Panel Attrition: What Next?

There is no reason to believe panel attrition as a survey research problem will be reduced to the point of not being a serious and important issue. Financial and staff resources often drive field operations decisions on the issues related to improving response rates. As a result, compromises are often made that may affect response rates, in general, and the overall attrition rate, in particular. After addressing data collection approaches to improve response rates, panel attrition can only be handled through post survey modeling, adjustments and documentation.

For large-scale longitudinal surveys, these strategies include:

- allocating staff and financial resources to conduct investigations of nonresponse, its components and correlates;
- identifying critical auxiliary variables to improve nonresponse adjustment models. This includes the use of data collected during survey waves in which the attritor is a respondent. It also means serious consideration be given to identifying and obtaining access to auxiliary variables in administrative record systems that are strongly related to the survey estimates;
- providing documentation and access to information that assists analysts in understanding how their analyses may be affected by nonresponse. Methodology reports, Users' Guides, and Quality Profiles generally provide such information, and, consequently, should be an important component of the longitudinal survey program;
- sponsoring and encouraging independent researchers to study the bias due to nonresponse and attrition. The Panel Study of Income Dynamics provides a good example of this approach. In 1994, the PSID sponsored a conference on attrition in longitudinal surveys. Several conference contributions analyzed attrition in the PSID and were later published in the Spring 1998 issue of the *Journal of Human Resources*. The SIPP is a good example of a slightly

different approach, whereby the survey program supported research on alternative nonresponse adjustment methods and reported results at a topic contributed session at the annual meeting of the American Statistical Association.

Obviously, doing everything possible to keep sample loss, and consequently attrition, as low as possible is very important. Substantial energy and resources ought to be dedicated for that purpose. In the final analysis, though, attrition will always exist. This is why every large-scale survey program, particularly a large-scale longitudinal survey program, ought to allocate a portion of its survey budget to support research on statistical and survey methods. In addition to a methodological research program, a program dedicated to the support of data users is also very desirable. Such a program would inform data users of data quality issues, in general, and problems related to nonresponse and attrition, in particular.

References

Abreu, D. and Winters, F. 1999. "Using Monetary Incentives to Reduce Attrition in the Survey of Income and Program Participation." *Proceedings of the Section on Survey Research Methods.* American Statistical Association. Alexandria, VA. pp. 533-538.

An, A., Breidt, F.J., Fuller, W. 1994. "Regression Weighting Methods for SIPP Data." *Proceedings of the Section on Survey Research Methods.* American Statistical Association. Alexandria, VA. pp. 434-439.

Boruch, R. and Pearson, R. 1986. "Assessing the Quality of Longitudinal Surveys." *Evaluation Review.* Sage Publications, Vol. 12, No. 1, pp.3-58.

Boruch, R. and Pearson, R. (eds.) 1986. *Survey Research Designs: Towards a Better Understanding of Their Costs and Benefits*. Lecture Notes in Statistics No. 38. Springer-Verlag. New York.

Boruch, R. and Pearson, R. 1985. *The Comparative Evaluation of Longitudinal Surveys*. Social Science Research Council. New York.

Citro, C. and Kalton, G. 1993. *The Future of the Survey of Income and Program Participation.* Washington, DC: National Academy Press.

Duncan, G. and Kalton, G. 1987. "Issues of Design and Analysis Across Time." *International Statistical Review.* Vol. 55, No.1, pp. 97-117.

Folsom, R. and Witt, M. 1994. "Testing a New Attrition Nonresponse Adjustment Method for SIPP." *Proceedings of the Section on Survey Research Methods*. American Statistical Association. Alexandria, VA. pp. 428-433.

Hill, D. 1994. "Adjusting for Attrition in Event-History Analysis." *Proceedings of the Section on Survey Research Methods.* American Statistical Association. Alexandria, VA. pp. 417-422.

Hill, M. 1992. *The Panel Study of Income Dynamics: A User's Guide*. Sage Publications, Inc.

Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M.P. 1989. *Panel Surveys*. John Wiley and Sons. New York.

Kay, W., Boggess, S., Selvavel, K. and McMahon, M. 2001. "The Use of Targeted Incentives to Selected Respondents on Response Rate and Data Quality." *Proceedings of the Section on Survey Research Methods.* American Statistical Association. Alexandria, VA.

Lo, A., Chu, A., and Apodaca, R. 2002. "Redesign of the Medicare Current Beneficiary Survey Sample." *Proceedings of the Section on Survey Research Methods.* American Statistical Association. Alexandria, VA.

Pearson, R. 1989. "The Advantages and Disadvantages of Longitudinal Surveys." *Research in the Sociology of Education and Socialization.* Vol. 8, pp. 177-199. JAI Press.

Rizzo, L., Kalton, G., Brick, M., and Petroni, R. 1994. "Adjusting for Panel Nonresponse in the Survey of Income and Program Participation." *Proceedings of the Section on Survey Research Methods.* American Statistical Association. Alexandria, VA. pp. 422-427.

U.S. Bureau of the Census. 1998. *SIPP Quality Profile*. Washington, DC.

Vaughan, D., and Scheuren, F. 2002. "Longitudinal Attrition in SIPP and SPD." *Proceedings of the Section on Survey Research Methods.* American Statistical Association. Alexandria, VA.

Basic Methodological Work for and in Repeated Cross-Sectional and Longitudinal Surveys: A Few Thoughts

Jon A. Krosnick
Departments of Psychology and Political Science
Ohio State University

This memo begins with some observations about the current status of survey methodology in social scientific exploration, highlighting some of the most important challenges survey researchers face. Then, I consider ways in which the large federally-funded infrastructure survey projects have helped to address these issues in the past, quite effectively in fact. Finally, I suggest some potential directions for future basic methodological research that could be accomplished in the context of the infrastructure surveys and could help to inform their design.

The Current Status of Surveys in the Social Sciences and Their Principal Challenges

These are very exciting times for survey methodology. The excitement comes from a fast-growing interest in the survey method across the social sciences within academia and outside. Never before in the history of the world has so much survey research been done and has it had so much impact on academic scholarship, on government policy-making, and on business decision-making.

Perhaps one reason for this surge in interest in survey research is generational replacement. The cohort that was running government and business 30 years ago went to college in about 1940 (plus or minus). And 1940 was before survey-based social science had been born and reaped real rewards, so these leaders learned little or nothing about survey-based social science in college. In contrast, the generation that is running government and business now went to college in the 1960s and 1970s, and these generations nearly all took introductory psychology courses and other social science courses that made great use of survey data. As a result, these decision-makers have long-appreciated the value of social science data collection and analysis and especially surveys.

Another likely reason for the surge in the survey method's appeal is the increasing sophistication and practicality of the method itself. First with the advent of telephone surveys and then with the introduction of Internet-based surveys, turn-around time for survey data collection has been decreased considerably as compared to the face-to-face surveys that began the field. So the fast pace of today's decision-making in government and business can be served well by surveys.

All this creates a real sense of value placed on the methodology, which is exciting for its practitioners. And the demand for people trained in the sophisticated application of the methodology has grown considerably as well, especially as evidenced by the degree-programs newly-established graduate training programs in survey methods by the Joint Program in Survey Methodology at the University of Maryland, the University of Michigan, Ohio State University, the University of Nebraska, and other such schools. Federal statistical agencies and private sector companies want competent survey professionals in greater numbers even than those these programs can supply. Fabulous.

At the same time, the use of survey data across the social sciences has been growing dramatically. One documentation of this trend was provided partially by Stanley Presser (1984), who conducted a content analysis of papers published in the most prestigious journals in various social scientific disciplines to assess the extent to which survey data were employed. He focused his analyses on 1948-1059, 1964-1965, and 1979-1980 and defined survey research rather narrowly, omitting the use of questionnaires in laboratory experiments, for example. His results appear in the first three columns of numbers in Table 1.

The rows are ordered here to progress from the field that used surveys the least in 1949-1950 (political science, at 2.6%) at the top to the field that used it the most then (sociology, at 43.0%) at the bottom. There was very substantial growth in survey data use from 1949-1950 to 1964-1964 in political science, economics, and sociology: six-fold, six-fold, and two-fold increases, respectively. The lone

hold-out here was social psychology, which showed a slight decrease in survey data use, from 22% to 15%.

Between 1964-1965 and 1979-1980, social psychology regained ground to return to 21%, and economics and sociology remained steady. Only political science manifested growth in use of surveys during that time period, rising to 35%.

But the really striking numbers in Table 1 appear in the final column. These are brand-new figures generated by political scientist Willem Saris (University of Amsterdam) and his colleagues (Saris, Gallhofer, van der Veld, & Corten, 2003). During the fifteen years beginning in 1980, the use of survey data increased in all disciplines. And social psychology played leap-frog with political science and economics, surpassing their usage rates. Fully half of articles in the Journal of Personality and Social Psychology used survey data during the 1994-1995 period. 70% of articles in sociology did so, and more than 40% of articles in political science and economics did so. Thus, during that 15-year period, the use of survey data increased by 18%, 47%, 138%, and 25% in the four disciplines, respectively.

These increases are gratifying for researchers who believe in the value of the method, and with increased use of the method has come increased interest in understanding how it works and how to make it work best to yield the most accurate measurements. This methodological interest addresses a wide variety of issues, including sampling procedures, data collection techniques, questionnaire design, and statistical analysis. We are now more sophisticated then ever about how to design representative samples, how to collect data in unbiased ways, how to design questions to accurately tap the information in people's minds, and how to conduct sophisticated analyses to model psychological and social processes with survey data to provide strong support for causal conclusions in addition to simply reporting frequency distributions and correlations between variables.

But as this sophistication has been rising, so have the challenges of conducting high-quality surveys using traditional methods. Most notably, the financial costs of collecting data have been skyrocketing. Face-to-face surveys with area probability samples have always been expensive, but the costs of training, supervising, and equipping a talented and experienced interviewing staff across the country has risen dramatically in recent years. Telephone interviewing has also become more expensive, most notably because of the substantial challenges in reaching people, given the impediments posed by caller-ID, call blocking, answering machines, and other technologies. The increase in cell-phone use poses another substantial challenge. And the costs of conducting self-administered mail surveys are now recognized to be just as substantial as the costs of telephone interviewing if a researcher wishes to achieve as high a response rate as possible by using multiple mailings and financial incentives.

Mainly because costs have been rising faster than survey budgets, observed response rates for many of the best quality surveys are thought to have dropped in recent decades. This is known to have been true for some top-level surveys conducted by the federal statistical agencies (though their response rates began exceedingly high and have dropped only slightly), and it is thought to be true as well for many academic and commercial surveys. Conventional wisdom among survey professionals is that the drops in response rates are not due to inevitable increases in the number of people who decline to participate. Rather, the challenge is simply to reach people in the first place – "non-contacts" are the obstacle, not actively opposed "refusals."

It is interesting to note that remarkably little evidence of systematic drops in response rates and the explanations for those drops have made it onto professional conference programs and into print in academic or nonacademic publications. Each investigator and survey organization knows the response rates they are obtaining, but very few organizations have gone public with the details of their response rates. This is so importantly because of the commercial nature of the data collection enterprise and the "prisoner's dilemma" that each organization confronts: "If I reveal my response rates and they turn out to be lower than those of my peer organizations, I'll lose reputation and business." So people have had an incentive to keep these rates private. Journals such as Public Opinion Quarterly have recently begun to require that all authors publish their response rates calculated using standard formulas, which is a terrific step forward, but the details behind those response rates are not yet being published routinely.

Furthermore, very few survey organizations are in a position to document trends over time in

response rates clearly, because the nature of the studies they conduct and the methods they use have changed considerably. One exceptional project in this regard is the Survey of Consumers conducted by the University of Michigan's Survey Research Center, for which a monthly RDD telephone survey has been conducted continuously since 1977 using data collection methods and questionnaires that have been held remarkably constant over time. Projects like that have unique abilities to track trends in response rates and to explain those rates. Yet even they do not typically have the incentive and resources to do such analyses and to publish them.

Exceptions to this rule of response rate invisibility are the federal infrastructure surveys, which routinely make public their response rates. And those for the General Social Survey have manifested no meaningful trend over the years:

1975—76 percent
1976—75 percent
1977—77 percent
1978—74 percent
1980—76 percent
1982—78 percent
1983—79 percent
1984—79 percent
1985—79 percent
1986—76 percent
1987—75 percent
1988—77 percent
1989—78 percent
1990—74 percent
1991—78 percent
1993—82 percent
1994—78 percent
1996—76 percent
1998—76 percent
2000—70 percent

This is powerful evidence suggesting that declines in response rates in recent years have been neither uniform nor inevitable. But they are thought to be occurring nonetheless.

Faced with the challenges posed by increasing costs and decreasing response rates, at least two principal innovations have been mounted by survey professionals: (1) basic methodological work to track the dynamics of response rates over time and explore ways to improve them, and (2) alternative modes of data collection that circumvent the challenges of conventional methods. The most notable innovation in the latter category is Internet-based data collection. Knowledge Networks, Harris Interactive, and many other firms have established Internet-based panels of respondents to provide survey data. Nearly all of these firms have been using quota-sampling, but Knowledge Networks has been employing probability sampling to recruit representative samples of American households via Random Digit Dialing telephone interviews and offered to equip households with free WebTV equipment in exchange for completing a questionnaire each week.

This work on dropping response rates and on use of the Internet has sparked a great deal of excitement among survey research professionals in government, academia, and the commercial world. But no silver bullets have been generated yet by this work. We have learned that advance letters and incentives do indeed improve response rates. And we have learned that Internet-based data collection is indeed viable. But we have a long way to go with basic research into inducing participation and mode optimization before we can have the confidence in survey data that researchers enjoyed in the 1960s, when face-to-face surveys were dominant and effective.

The accumulated social science literature provides us with much more, in fact a veritable mountain of evidence, on optimal questionnaire design. During the last one hundred years, thousands of studies have been published in psychology, sociology, political science, economics, education, linguistics, medicine, public health, communication, and many other disciplines comparing different ways of asking questions and yielding solid experimental evidence about how to optimize the structure and wording of questions (for a review, see Krosnick and Fabrigar, forthcoming). But even this mountain of evidence, manifesting a great deal of coherence and replication across studies, time periods, and respondent populations, raises as many important questions as it answers. So there is much to be done to improve this aspect of survey methodology as well.

The Role of Infrastructure Survey Projects in Methodological Innovation

During the last 50 years, the federal government has done a great deal to help the development of our understanding of the survey method by funding large-scale infrastructure surveys that have served as platforms for social scientists to collect survey data. These platforms include the National Election Study, the General Social Survey, the Panel Study of Income Dynamics, the National Longitudinal Surveys, and more. Of course, these studies have primarily yielded substantive data for the study of social phenomena. But of necessity, the principal investigators running these studies have been deeply concerned about optimizing methodology, both to conserve costs and to maximize data quality.

The two cases I know best are the National Election Study and the General Social Survey. Since their inception, both of these projects have involved a continuous stream of methodological innovations and methodological experimentation to better understand their techniques and to improve their practice. The NES website currently lists 116 pilot study reports and 63 technical reports. Some of these reports have focused on technical issues involving the main biannual studies of elections, but the vast majority of these reports have described the findings of analyses of data from the "Pilot Studies" conducted by NES in non-election years.

These reports describe numerous experiments designed to address basic issues in survey methodology, including the impact of branching vs. one-step measurement of attitudes, comparisons of the quality of data obtained from telephone vs. face-to-face interviewing, permitting respondents to place themselves in ranges of points on rating scales rather than at single points, comparing the measurement of regular behaviors using questions about a "typical" week vs. questions about the past week, identifying optimal ways of identifying knowledge-based expertise in a domain, measuring perceptions of the personality traits of other people and the emotions evoked by other people, measuring identification with social groups, optimizing measurement of the frequency of behaviors that are subject to strong social desirability pressures, measuring values and ideologies, using framing experiments to identify the causes of attitudes, measuring religion and religiosity, testing a model of the cognitive processes entailed in reporting attitudes, the impact of pre-notification and incentives on survey participation, measuring social connectedness and social networks, measuring attitude strength, measuring news media exposure, the impact of offering "don't know" response options on the validity of knowledge measures, developing efficient survey-based measures of respondents' personality traits to replace psychology's typical long item batteries, comparing the reliability and validity of rating scales of various lengths, documenting the impact of question order, and much more.

The ICPSR/GSS resources website lists 90 methodological reports addressing such issues as house effects on measurements, measurement of ethnic identities, comparisons of seemingly synonymous words in survey questions, the causes of "don't know" responses, question order effects, recalling attitudes from prior times, comparison of rating and ranking methods for measuring priorities, comparing the quality of data obtained by face-to-face and telephone modes of data collection, measuring ideologies cross-nationally, the accuracy of proxy reporting, documenting social networks, measuring religion and religiosity, the impact of response choice order on answers, measuring income and poverty, measuring occupations, measuring voluntary memberships in organized social groups, the optimal measurement of confidence in institutions, and more.

Clearly, these two infrastructure studies have done a great deal to conduct basic methodological research that has use not only in their own questionnaire design activities but in the design activities of many survey researchers conducting their own independent projects. This work is not merely fine-tuning small details – the infrastructure survey platforms have been the launching pads for genuine substantial methodological innovation and evaluation, much of it published in prestigious journals and widely recognized as valuable. Indeed, these are especially efficient contexts for such work, because so much other information is collected from respondents to complement and supplement any given experiment on question design.

No in-depth investigation of an important methodological innovation can be conducted in the vacuum of a single stand-alone, small-scale study – comparing two question wordings, for example, is best done with a long questionnaire and measurements made on the same respondents in other interviews that can serve as validation criteria. It would be impractical for each methodologist to mount his or her own independent study of such magnitude – and indeed, doing so is unnecessary, because the infrastructure studies have offered platforms for methodological testing. Such work is best done in two stages, as the National Election Studies have done, whereby pilot studies involving reinterviews of respondents offer low-cost opportunities for initial tests, to be followed by large-scale replications for measures or techniques that prove worthy initially.

It is worth noting that the infrastructure projects are especially important because they have offered opportunities for methodological innovation to the community of scholars more broadly than any conventional granting program could. As a graduate student without a track record of grant-getting or independent publication or the skills to write a full grant proposal, I was able to submit a suggestion to the National Election Studies for a methodological study that was judged on the merits of the ideas offered and that led to an economical test and ultimately a series of publications. Projects like NSF's TESS continue this tradition, but in more of a stand-alone fashion. TESS is a fabulous resource for many reasons, but it cannot offer the rich array of criterion variables that studies like the GSS and NES can for the development of measures intended to contribute to a long-standing intellectual tradition of studying a particular phenomenon (e.g., voting, public opinion, and elections). So the pilot study model for methodological innovation is especially appealing, because it affords investigative opportunities to a broad array of scholars, from the most junior to the most senior.

Another way in which infrastructure projects have offered unmatched opportunities for methodological innovation has to do with reinterview rates in panel studies. Studies such as the Panel Study of Income Dynamics and the National Longitudinal Surveys have been in the business of creating and maintaining long-term panels of respondents who are interviewed many times. Knowledge Networks has done the same thing. And the directors of these projects have learned that maintaining panel reinterview rates (and sample representativeness) across waves of data collection is increasingly difficult.

This has led researchers to conduct systematic experimental investigations comparing various techniques for increasing respondent willingness to be reinterviewed. Some of this work finds its way into publications and is disseminated throughout the social sciences via presentations at research conferences, but some of this work is not thusly disseminated, and it would be nice if more dissemination were to occur. The challenge facing these investigators is to find the time to write such publications while at the same time doing their primary non-methodological work, and the incentive system does not make that as easy as it might. Nonetheless, this work is being done and is very valuable.

Basic methodological work is also being done on how to maximize response rates in RDD telephone surveys by researchers involved in another major federal infrastructure survey project: the CDC's Behavioral Risk Factor Surveillance System, which involves coordinated RDD surveys done in every state of the nation. That project has always paid close attention to response rates and has seen those rates drop in recent years. As a result, the investigators have begun an aggressive research effort to identify ways they might improve response rates and implement those techniques. Their findings are being disseminated at conferences, though I am not aware of publications reporting them yet.

Parallel efforts are also being made by other non-academic/non-government research organizations. For example, the ABC News Polling Unit is in the business of constant RDD surveying of the American

public, and they have implemented experiments recently to attempt to increase response rates. Personnel in that office have no time available to publish their findings, but they are valuable findings that could be of great use to the profession if they were to be disseminated along with the findings of other such organizations doing similar work. Federal funding to promote dissemination of such findings would be a very worthwhile investment. In the meantime, the NSF infrastructure survey studies have done a great deal to advance our understanding of how to optimize the survey method, even thought many of the insights thusly generated have not yet been widely disseminated.

Are the infrastructure surveys incompatible with methodological innovation? Are these studies by their very nature "stuck in the mud" established by their past practices? The answer is "yes and no." As detailed above, these studies have not been shy about exploring new methodologies, and the vast majority of the methodological research these groups of scholars have conducted have led to innovations in their procedures or commitments to existing procedures that are shown to be optimal through comparative research. But it is true that the commitments of the NSF infrastructure surveys to tracking trends in attitudes, beliefs, and behavior patterns over time create powerful incentives against changing measurement procedures.

I believe this incentive is just as it should be. One very powerful illustration of the potential problems that can result from methodological innovation is the landmark book, The Changing American Voter (Nie, Verba, & Petrocik, 1976). Its authors found that sharp shifts in the structure of political attitudes took place in 1964 in the National Election Study datasets that were sustained thereafter and suggested that a tremendous change had taken place in Americans' thinking about politics due to the events of the early 1960s. This finding was broad and multifaceted and had wide-reaching implications detailed in this very long book published by Harvard University Press.

But most likely, the book is wrong about its central point. The reason is that the National Election Study implemented a radical shift in measurement technology in 1964. The formats, wording, and ordering of many questions were changed significantly, most likely for the better in terms of psychometric precision. But researchers were inattentive to the potential impact of such shifts, and they were accorded relatively little attention in The Changing American Voter. Later, Bishop, Tuchfarber, and Oldendick (1978) and Sullivan, Piereson, and Marcus (1978) conducted experiments comparing the old wordings/structures to the new and showing that the wording/structure shift alone was most likely responsible for what amounted to an illusory shift in the nature of American political cognition.

Therefore, methodological change, even for the better, must be done very carefully in the large-scale long-term infrastructure surveys. The "mistake" made in interpreting the shift in NES data in 1964 occurred before NES instituted its tradition of methodological pilot studies. Little is known about the developmental work that may have preceded the shift in question wording/structure by way of such research, but pre-testing on the scale of typical NES pilot studies seems unlikely to have been done. If it had been done, the Changing American Voter would no doubt have looked quite different and may not have been published at all.

So does this mean that innovation should be done rarely if ever? I would argue that such conservatism would be a mistake. Instead, innovation should take place regularly, but via splicing after careful pretesting. That is, in order to permit over-time comparisons unconfounded by methodology shifts, an old question wording and its new replacement should each be asked of randomly selected halves of the respondents for at least two waves of data collection. Then, trends can be tracked for the same period using the two measurement methods, and adjustments can be made to create an overlap.

Continuing Methodological Research in the Infrastructure Projects

Although we have learned a great deal about survey methodology from experimental work done in the context of the infrastructure projects, there is much still to learn. One of the most pressing issues at the moment involves modes of data collection. As the challenges of telephone interviewing have increased, researchers have begun to reconsider the advantages and disadvantages of all modes. The most common comparison done during the last 30 years has been between face-to-face interviewing with

telephone interviewing (for a review of these studies, see Holbrook, Green, & Krosnick, in press; Ellis & Krosnick, 2003).

When telephone interviewing was introduced to the survey methodology toolbox almost 40 years ago, researchers were attracted by the many practical appeals of the method, and early comparisons of data collected by telephone with data collected face-to-face were reassuring in suggesting few differences in obtained results as compared to face-to-face surveys (e.g., Groves & Kahn, 1979). But that early work and much work since then was relatively atheoretical and did not develop theory-based predictions about what sort of mode differences would be expected for which sorts of questions, why, and when. Recent work bringing such a theoretical approach has turned up consistent evidence of mode effects, suggesting that you get what you pay for when it comes to the contrast of telephone and face-to-face interviewing (see Holbrook et al., 2003).

For a variety of reasons, face-to-face interviewers appear to develop a greater sense of rapport and trust and more task engagement from respondents, especially respondents with the least education, and this elicits from these respondents more honest answers and more effort in doing the cognitive work necessary to generate accurate answers when reporting attitudes, beliefs, and behavior patterns. Furthermore, telephone surveys have consistently achieved lower response rates than face-to-face surveys, and sample representativeness in terms of demographics has been consistently lower for telephone surveys than for face-to-face surveys, the former bringing a greater bias toward the well-educated, the wealthy, and white respondents (Ellis & Krosnick, 2003). This work suggests that the long-term infrastructure studies that have routinely employed face-to-face interviewing would pay a considerable price in terms of response rates, sample representativeness, and reporting accuracy if they were to shift to the telephone. The evidence documenting these differences comes from studies conducted in the 1970s, 1980s, 1990s, and very recently, so the patterns seem quite general.

But there is still more basic work to be done to understand precisely why those mode differences occur and what might be done to overcome them. This work can be done partly in the laboratory (comparing face-to-face interviewing with interviewing via intercom, for example), but it must also be done partly in the field. Furthermore, there are other modes worth investigating. One obvious alternative is self-administered paper-and-pencil questionnaires. For a long time, this mode seemed seriously problematic because of apparently inevitably low response rates, but Dillman's (1978, 2000) thorough and innovative approach to this problem identified techniques that can produce response rates comparable to those obtained by telephone surveys. Self-administered paper-and-pencil questionnaires are often used as supplements to face-to-face surveys to allow respondents to report potentially sensitive information confidentially in such a way that the interviewer does not know the respondent's answers.

However, obtaining a high response rate is not the only goal in mode selection, and there is reason to worry that self-administration on paper may bring with it some costs in terms of response accuracy. Most importantly, the notion of accountability has been a central focus of study in social psychology in recent years, and the research thusly generated points to a potential danger of self-administered paper-and-pencil questionnaires. When people answer questions completely anonymously on paper and pencil, they have no sense that they may be held accountable for the accuracy of the answers they provide. This lack of accountability has been shown to yield a wide range of errors and biases in reasoning in psychological experiments comparing answers given to questions under such conditions with answers to the same questions under conditions of high accountability (for a review, see Lerner & Tetlock, 1999). Recent evidence suggests that self-administered paper-and-pencil questionnaires used in survey settings may indeed bring with them a lack of accountability and a consequent decrease in respondent effort, yielding less accurate reports (see, e.g., Krosnick, Holbrook, Berent, Carson, Hanemann, Kopp, Mitchell, Presser, Ruud, Smith, Moody, Green, & Conaway, 2002; Silver & Krosnick, 2001).

For example, Silver and Krosnick (2001) found that in an experiment in which respondents were randomly assigned to complete a questionnaire either on their own with paper and pencil or via telephone interviews, people completed the paper and pencil questionnaire in about half time tie (15 minutes) as it took to complete the telephone interviews (31 minutes on average), and the quality of responses provided by the self-administration respondents was 30% lower than the quality of responses provided by the

telephone respondents. This sort of evidence suggests that people may rush through self-administration exercises on paper. And if mode quality is transitive, then taking together a series of past studies suggests that face-to-face interviewing may yield the highest data quality, followed by telephone interviewing providing somewhat lower quality, followed by paper-and-pencil self-administration, which may yield the lowest quality. However, direct comparisons across these three modes have yet to be conducted while at the same time eliminating design confounds. Such research is worth conducting in the future.

In light of this set of findings, it is interesting to consider the new kid on the block: Internet-based surveys. Most Internet surveys being conducted these days involve haphazard samples implemented with quotas. This approach is inherently severely limited, and some recent comparisons suggest that such samples significantly compromise representativeness (see, e.g., Chang & Krosnick, 2001a, 2001b). Because Knowledge Networks has built its panels via systematic probability sampling methods, it has the potential to yield more valid and useful data.

Of course, if the problem with paper-and-pencil self-administered questionnaires is attributable to the self-administered quality of the experience, then there would be reason for concern about Internet-based surveys, even those conducted by Knowledge Networks. But in fact, Knowledge Networks' national samples are nearly as representative of the country's population as telephone survey samples are, and the amount of random and systematic measurement error appears to be lower in the Internet data than in the telephone data (Chang & Krosnick, 2001b, 2002). There are a variety of likely reasons for this, and three of the most important are (1) the commitment to a long-term reciprocal exchange relationship, (2) the practice that respondents get from repeatedly completing questionnaires in the same mode (see, e.g., Donovan & Radosevich, 1999), and (3) the slow pacing and thorough consideration of each question that comes of necessity from a questionnaire presented with a single question per screen. This suggests that long-term panels conveying survey data to researchers via the Internet may indeed offer the potential for high quality data collection.

The challenge confronting researchers interested in employing the Knowledge Networks approach is that sampling via RDD telephone interviewing limits the representativeness of the sample due to coverage and nonresponse error inherent in telephone contacting. Therefore, an interesting avenue for future work would be to explore applying the Knowledge Networks approach with initial contacting done via area probability sampling and initial face-to-face interviews. Any one of the many large-scale infrastructure survey projects conducting face-to-face interviews could complete an interview and then ask respondents whether they would like to accept free WebTV equipment in exchange for weekly survey completions. This could be a very cost-effective way of assembling a maximally representative sample for web-based surveys, but only when it is attempted via developmental research will we learn if this approach can work in practice.

One of the likely keys to the viability of Knowledge Networks' approach is the constant involvement of respondents in completing questionnaires every week on a wide variety of topics. Because questionnaire completion becomes such a routine activity for these people, and because the topics are constantly changing from week to week, it seems less likely that this sort of data collection will focus respondents' interest and attention and thereby change the very phenomena being measured (e.g., Bridge, Reeder, Kanouse, Kinder, Nagy, & Judd, 1977). However, there is always the risk that repeated interviewing on the same topic may indeed produce effects on respondents and thereby make them unrepresentative of the populations to which the data are to be generalized. Therefore, much research work is needed exploring how reinterviewing changes people in specific life domains so that the representativeness of samples can be maintained while benefiting from the richness of data that come from long-term panels. Indeed, the long-term panels that have been maintained as part of the Panel Study of Income Dynamics and the National Longitudinal Surveys should be routinely compared to fresh cross-sectional samples to assess impacts of repeated interview conditioning, but any observed differences should not automatically be attributed to conditioning. It is quite possible that the practice at survey reporting acquired by panel participants may improve the accuracy of their responses, so it is worthwhile to explore the explanations for observed differences empirically.

There is also lots of room for more investigation of questionnaire design issues in the large-scale infrastructure surveys. Some issues worth investigating may appear particular to certain domains of study. For example, researchers studying political behavior routinely ask people whether they voted in an election and for which candidates they voted. These reports routinely depart from actual rates of turnout and actual proportions of votes cast for particular candidates (e.g., Clausen, 1968; Parry and Crossley, 1950; Traugott & Katosh, 1979). It is tempting to presume that these discrepancies are due to intentional misreporting by respondents, but remarkably little evidence documents such a conclusion (because of the inherent difficulties involved in validating reports about voting). And it is possible that the observed discrepancies may be attributable to sample unrepresentativeness rather than misreporting (e.g., Greenwald, Vande Kamp, Klinger, & Kerr, 1988). Indeed, some evidence indicates that pre-election interviews actually increase voter turnout, thereby producing discrepancies with aggregate figures (e.g., Greenwald, Carnot, Beach, & Young, 1987). Yet much more work is needed to document why the discrepancies exist and what can done to solve the problems. Although this work's implications might seem to be applicable only to research on voter behavior, they in fact have broad implications for measurement of any socially sensitive behavior.

There is also lots of potential for important research to be done to help in optimizing other, more general aspects of questionnaire design. A partial review of recent work on some specific topics appears in a recent chapter in the Annual Review of Psychology (Krosnick, 1999). This review outlines the theory of survey satisficing and explains how low respondent ability, low respondent motivation, and high task difficulty can lead particular respondents to shortcut the cognitive processes entailed in generating optimal answers to survey questions and to settle for merely satisfactory or minimally acceptable answers instead. This theoretical perspective is now validated by a large body of empirical studies dating back more than 70 years, the findings of which have been known but rarely were integrated or accounted for in theoretical terms. The theory of survey satisficing is therefore proving to be very helpful by bringing some coherence to a set of minimally-coherent findings. Although we now have a lot of evidence documenting survey satisficing, we know much less about whether and how we can design surveys to minimize satisficing. This work certainly seems merited in the future.

Another area for investigation is the measurement of experienced event frequencies. Many survey studies seek to ascertain how often people have performed a particular behavior or experienced a particular type of event and to date the behaviors they have performed or events they have experienced. In general, we know that recall longer in the past tends to be less accurate than recall regarding the recent past. But our literature offers us remarkably little guidance about how far back in the past to go with a survey question. A current NASA-funded research project began with a comprehensive review of the existing literature on this issue (Silver & Krosnick, 1999) and proceeded to conduct experiments systematically varying factual recall period and permitting comparisons of data quality across those periods (Silver & Krosnick, 2000). The methods used there could be employed to help questionnaire designers get a better handle on optimizing recall question design.

This is a tiny list of the many potentially fruitful directions for future basic methodological research that could be done very effectively in the context of the infrastructure surveys. Having gone on way too long in this memo, I will stop here and promise to provide more such suggestions if they would be useful in the future.

Conclusions

In sum, the NSF-funded infrastructure surveys have done a great deal of good for the study and refinement of survey methodology, but much more work is left to be done, and those platforms offer exceptional opportunities for constructive investigation and innovation. There are many unanswered questions about how to optimize surveys that can and should be investigated both in the context of these studies and more broadly outside of them as well, to inform their design and preserve these valuable national resources.

References

Bishop, G. F., Tuchfarber, A. J., & Oldendick, R. W. (1978). Change in the structure of American political attitudes: The nagging question of question wording. American Journal of Political Science, 22, 250-269.

Bridge, R. G., Reeder, L. G. Kanouse, D., Kinder, D. R., Nagy, V. T., & Judd, C. M. (1977). Interviewing changes attitudes – sometimes. Public Opinion Quarterly, 41, 57-64.

Chang, L., & Krosnick, J. A. (2001a). The representativeness of national samples: Comparisons of an RDD telephone survey with matched Internet surveys by Harris Interactive and Knowledge Networks. Paper presented at the American Association for Public Opinion Research Annual Meeting, Montreal, Canada.

Chang, L., & Krosnick, J. A. (2001b). The accuracy of self-reports: Comparisons of an RDD telephone survey with Internet Surveys by Harris Interactive and Knowledge Networks. Paper presented at the American Association for Public Opinion Research Annual Meeting, Montreal, Canada.

Chang, L., & Krosnick, J. A. (2002). Comparing self-administered computer surveys and auditory interviews: An experiment. Paper presented at the American Association for Public Opinion Research Annual Meeting, St. Petersburg, Florida.

Clausen, A. (1968). Response validity: Vote report. Public Opinion Quarterly, 32, 588-606.

Dillman, D. A. (1978). Mail and telephone surveys: The total design method. New York, NY: Wiley.

Dillman, D. A. (2000). Mail and Internet surveys: The tailored design method. New York, NY: Wiley.

Donovan, John J., and David J. Radosevich. 1999. "A Meta-Analytic Review of the Distribution of Practice Effect: Now You See It, Now You Don't." *Journal of Applied Psychology* 84:795-805.

Ellis, C., & Krosnick, J. A. (2003). Sample representativeness in telephone and face-to-face surveys. Manuscript under review.

Greenwald, A. G., Carnot, C. G., Beach, R., Young, B. (1987). Increasing voting behavior by asking people if they expect to vote. Journal of Applied Psychology, 72, 315-18

Greenwald, A. G., Vande Kamp, M., Klinger, M. R., Kerr, K. L. (1988). A contactability bias in surveys of voter turnout. Unpublished manuscript. Seattle, WA: University of Washington

Groves, R. M., & Kahn, R. L. (1979). Surveys by telephone: A national comparison with personal interviews. New York: Academic Press.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (in press). Telephone vs. face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. Public Opinion Quarterly.

Krosnick, J. A. (1999). Survey methodology. Annual Review of Psychology, 50, 537-567.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., Presser, S., Ruud, P. A., Smith, V. K., Moody, W. R., Green, M. C., & Conaway, M. (2002). The impact of "no opinion" response options on data quality: Non-attitude reduction or an invitation to satisfice? Public Opinion Quarterly, 66, 371-403.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. Psychological Review, 125, 255-275.

Nie, N. H., Verba, S., & Petrocik, J. R. (1976). The changing American voter. Cambridge, MA: Harvard University Press.

Parry, H. J., & Crossley, H. M. (1950). Validity of responses to survey questions. Public Opinion Quarterly, 14, 61-80.

Presser, S. (1984). The use of survey data in basic research in the social sciences. In C. F. Turner and E. Martin (Eds.), Surveying subjective phenomena. New York: Russell Sage Foundation.

Saris, W. E., Gallhofer, I., van der Veld, W., & Corten, I. (2003). A scientific method for questionnaire design: SQP. Amsterdam: University of Amsterdam.

Silver, M. D., & Krosnick, J. A. (1999). Selecting a recall interval to maximize the accuracy of pilot recollections: A literature review. Unpublished manuscript, Ohio State University, Columbus, Ohio.

Silver, M. D., & Krosnick, J. A. (2000). NAOMS field trial results report. Unpublished manuscript, Ohio State University, Columbus, Ohio.

Silver, M. D., & Krosnick, J. A. (2001). An experimental comparison of the quality of data obtained in telephone and self-administered mailed surveys with a listed sample. Paper presented at the American Association for Public Opinion Research Annual Meeting, Montreal, Canada.

Sullivan, J. L., Piereson, J. E., & Marcus, G. E. (1978). Ideological constraint in the mass public: A methodological critique and some new findings. American Journal of Political Science, 22, 233-249.

Traugott, M. W., & Katosh, J. P. (1979). Response validity in surveys of voting behavior. Public Opinion Quarterly, 43, 359-77.

Table 1: Use of Survey Data in the Social Sciences Since 1949

Discipline	1949-1950	1964-1965	1979-1980	1994-1995
Political science	2.6%	19.4%	35.4%	41.9%
	(114)	(160)	(203)	(303)
Economics	5.7%	32.9%	28.7%	42.3%
	(141)	(155)	(317)	(461)
Social psychology	22.0%	14.6%	21.0%	49.9%
	(59)	(233)	(377)	(347)
Sociology	24.1%	54.8%	55.8%	69.7%
	(282)	(259)	(285)	(287)

Sources: Presser (1984) for the first three columns of numbers; Saris, Gallhofer, van der Veld, and Corten (2003) for the final column

Incorporating Experiments, Open-Ended Questions, and In-Depth Interviews in Surveys: Racial Attitudes as an Illustration of the Possibilities

Maria Krysan
University of Illinois at Chicago

What do people think and feel about people of different races and racial issues? What is the basis for these thoughts and feelings? How have attitudes changed—or not changed—over time? And why? Survey research—and in particular recurring surveys such as the National Election Studies and the General Social Survey—have provided much of what we know by way of the answers to these and other questions about how race is "lived" in America, including, but not limited to, information about racial attitudes in particular (for one compilation, see Schuman et al. 1997).

These survey data reveal much about what has—and has not—changed about how Americans think about race. A detailed look at this record of survey data goes well beyond the scope of this short paper, but suffice it to say that the story they tell is a complex one. For example, one category of racial attitude questions reveals much change in a liberal direction: today, Americans rarely object to the principles of racial equality (e.g., whether or not whites and blacks should be allowed to go to the same schools or whether public accommodations should be integrated) and seldom endorse biological explanations for racial differences. But at the same time, a substantial percentage of whites continues to endorse negative racial stereotypes, attribute racial inequality to a lack of motivation by blacks, and deny that discrimination against African Americans is an important cause of racial inequality. Fair housing policies, various forms of affirmative action, and other policies intended to ameliorate racial inequality often garner little support, have not shown much change over time, and in some cases support has eroded.[7] In addition to time series analyses, survey data are also used to adjudicate between theoretical explanations for various patterns of racial attitudes and thus provide important information about the sources and foundations of racial prejudice.

In short, survey data can be seen as a "social telescope" (Converse 1987) revealing if and how attitudes have changed, and providing the empirical data necessary for understanding the sources of these attitudes.

---

[7] Repeated cross-section surveys in which the same survey questions are repeated over time have provided these insights and serve as the foundation for analyses of the causes and implications of such patterns. Thus, it is crucial that survey researchers retain some of these "old" measures of racial attitudes. At the same time, we must continue to develop "new" measures to create a survey record that reflects the complexity of contemporary race relations. The major on-going surveys, which allow for tracking change—or stagnation—over time, have begun to do the latter, by including questions on such topics as affirmative action, affect toward racial groups, modern stereotypes, and perceptions of discrimination and reverse discrimination. However, a review of the National Election Studies and the General Social Survey suggests a growing overlap between these two surveys in the questions they ask about race—an overlap that was not true in past decades (see Schuman et al. 1997 and cf. Hyman and Sheatsley 1956; Hyman 1964). While replication across different survey organizations provides valuable information about the reliability of the results, it also means that the survey record on racial attitudes we are constructing may be more narrow and, therefore, less able to reflect the complexity of racial attitudes. In addition, as our country becomes increasingly multi-racial/ethnic, there is a serious need to incorporate questions that are asked of—and about—other racial/ethnic groups, such as Latinos and Asians. Without such inclusion, we run the risk of creating a less complete record for the future about how race has mattered in the hearts and minds of Americans.

But surveys can also be the foundation and backdrop for detailed examinations that are more akin to a "social microscope." One way of doing so is through a coupling of traditional surveys with other data collection methods. In the remainder of this paper, I use racial attitudes as a "case study" to illustrate how experiments, open-ended questions, and in-depth interviews can be combined with traditional surveys to provide further depth and nuance to social scientific knowledge on this topic. While my focus is on racial attitudes—simply because it is the topic about which I am most familiar—the application to other social phenomenon should be clear.

To the degree that traditional survey research designs incorporate experiments and open-ended questions, and conduct supplemental in-depth interviews, the emphasis is often on using these tools to answer particular methodological questions about how best to design a survey or word a question, such as: Do we include or exclude the "don't know" response? Do we offer a middle alternative? Do we phrase a question this way or that? What is the effect of a $1.00 versus $5.00 incentive on response rates? Should we match race of interviewer and respondent? To be sure, answering questions like this—in some cases drawing on experiments, open-ended questions and in-depth interviews—has been critical to the advancement of survey research as a rigorous social scientific method. But in this paper, I focus on the use of hybrid data collection efforts to also make important *substantive* contributions.

Three features of racial attitudes make it particularly fertile ground for illustrating the benefits of a multi-method approach. First, race in America and racial attitudes in particular have become *sensitive* (given the current climate in which social desirability pressures to not appear racist are considerable) and, second, *subtle* (in the sense that expressions of racial attitudes and the influence of race can be indirect and nuanced). Third, the topic is notably *complex*—there is no single type of racial attitude, and no "simple and sovereign" (Allport 1954) theoretical interpretation of what underlies racial attitudes.

The Sensitivity of Race In America. In the contemporary racial climate, there are strong social pressures against expressing racial prejudice and negative racial attitudes. Because surveys are based on self-reports, this poses a challenge to survey researchers since the social pressures of a survey may result in under-reporting of negative racial attitudes and thus create social desirability bias in the data. Evidence of this sensitivity with respect to racial attitudes is clear from studies showing that responses can vary depending on the social context of the interview: a change in the race of the interviewer or the privacy of the setting in which the questions are asked can have sometimes-dramatic effects on the reported attitudes (Anderson et al. 1988a, b; Schuman and Converse 1971; Krysan 1998; Davis 1997). Indeed, some argue that survey data on racial attitudes, for this reason, are of little value. Others are more sanguine and suggest that this sensitivity (e.g., to race of interviewer) should be treated not as a methodological artifact, but as a "fact of social life" which can be used to shed light on the very nature of racial attitudes and inter-racial interaction (Schuman and Converse 1971).

The Subtlety of Race in America. One feature of racial attitudes in America today is the increasing subtlety of expressions of racial bias or negative racial attitudes. For example, biological or genetic explanations for inequality have been replaced by more subtle "cultural" assertions about group differences (e.g., Bobo, Kluegel and Smith 1997; Schuman et al. 1997). Or, respondents may object to blatant group-based negative characterizations, but nevertheless hold to the somewhat more subtle belief that one group, on average, is simply "less good" than another. Still others have argued that there exist implicit racial attitudes—racial biases of which the individuals themselves are unaware (e.g., Wittenbrink, Judd and Park 1997; Devine 1989). Because of this subtlety of expression, tools must be designed to capture the nuances of racial attitudes which, though perhaps more subtle, have consequences—in the form of persistent discrimination in housing, jobs, and other social institutions—that are far from subtle.

The Complexity of Race In America. For all their sensitivity and subtlety, however, existing survey data using traditional methods nevertheless paint a complex and in some cases contradictory picture: there is no single "trend" among racial attitude questions and no single theoretical model that can explain

the bases of these attitudes. For example, a contentious debate persists about whether attitudes about racial policies (such as affirmative action) are driven more by racial prejudice or by political beliefs (see Sears et al. 2000 for a review of this debate). Similarly, when trying to explain whites' avoidance of neighborhoods with African American residents, there is a growing debate about whether this avoidance is because of race, per se (i.e., driven by racial prejudice), or whether it is because of negative (social-class based) characteristics that whites associate with integrated neighborhoods. Finally, the increasing racial/ethnic diversity of our nation adds layers of complexity to studies of racial attitudes: how do different racial/ethnic groups feel toward each other and about policies intended to benefit one or the other group? The increasing complexity, sensitivity, and subtlety of racial attitudes and racial issues calls for increased methodological savvy on the part of those interested in using surveys to shed light on this topic. One way is by embedding other methodological tools—survey-based experiments, open-ended questions, and in-depth interviews—within surveys.

Using Hybrid Data Collection Efforts to Understand Racial Attitudes.

Experiments within surveys. One way that experiments within traditional surveys of racial attitudes have been used is to determine if the manner by which the data are collected (e.g., the race of the interviewer or the privacy of the setting) influences reports of racial attitudes. These studies show that different dimensions of racial attitudes are more and less sensitive to such changes—for example, questions about the principles of racial equality are less susceptible to social context changes, while questions on racial policies and perceptions of discrimination are more susceptible (Krysan 1998; Krysan and Couper 2002; Schuman and Converse 1971; Hatchett and Schuman 1976). These experiments randomly assign respondents to different interview conditions and the results, by focusing on the *kinds* of questions that show effects, can provide substantive conclusions about inter-racial interactions and tensions.

A second use of experiments—and one that is increasingly common—is to conduct question-wording experiments that are designed to help untangle substantive questions and puzzles. In this case, respondents are randomly assigned to answer one version of the question or another, and a comparison of the responses to the different versions is used to test hypotheses about the role of race, often vis a vis other characteristics, in shaping attitudes. For example, in studies of residential preferences, the racial composition—both the numbers and particular racial/ethnic groups involved (Asians, Latinos, African Americans, white)—are experimentally varied (Emerson et al. 2001; Charles 2000; Schuman and Bobo 1988). In studies of racial policies, the target of the policy (women, poor people, African Americans, etc.) or the framing of the policy (to highlight one value or another) is experimentally varied. This is done to determine, for example, whether there is a racial double standard or what role particular values (e.g. individualism, equality) as against racial prejudice play in shaping racial policy attitudes (e.g. Sniderman and Piazza 1997; Kinder and Sanders 1997; Bobo and Kluegel 1993). For many of these experiments, the question is: all else being equal, does race continue to play a role in shaping reactions to neighborhoods, government programs, potential employees, or university admissions policies? By using experimental variations like this, respondents are unaware that different variants of the questions are asked, and so the effect of race can remain more subtle and indirect.

The ability to incorporate experiments into general population surveys is a powerful tool that combines the benefits of the controls that are possible in experiments with the generalizability of results from representative samples. Sniderman and Grob (1996) provide an overview of the recent innovations in the use of experiments in survey research and highlight that technological advances allowing for easy randomization have greatly simplified this process and added to the power of this method. Moreover, with the use of laptops in face-to-face interviewing and web surveys, there is now the possibility of incorporating even more subtle manipulations. For example, a recent experiment by Hutchings and colleagues (2001) embedded subtle racial cues (in the form of images in a videotape that respondents

viewed during the interview) in a computer assisted self-administered instrument to assess—indirectly and through experimental manipulation—the effect of race in political campaign advertising (Hutchings, Traugott and Valentino 2001).

Open Ended Questions within Traditional Surveys. Several decades ago, Schuman (1966) advocated the use of random probes in surveys, whereby a selection of respondents were asked an open-ended follow-up question to a closed-ended item. This was done largely to obtain information about the meaning of the question to the respondents. However, use of open-ended questions also has the possibility of providing information on the mechanisms, motivations, and foundations of attitudes—in short, the kinds of insights that are generally the domain of more qualitatively-oriented data collection tools. Because they are embedded in a traditional survey, though, they have an advantage over traditional qualitative methods in that the data come from large numbers of representative samples of respondents. Again, the increasing use of technology in survey data collection makes the incorporation of open-ended questions less labor intensive than when Schuman (1966) first advocated the practice (e.g., laptop computers can audio-tape the responses and computer software make the coding and analysis of the data easier).

In several articles examining the role of racial residential preferences in contributing to residential segregation, I have made use of open-ended follow-up questions to provide depth and nuance to the findings from a set of standard closed-ended questions (Krysan 2002a; 2002b; Krysan and Farley 2002). After asking respondents, using closed-ended questions, to characterize their preferences about living with people of different races, they were asked to explain the reasons behind their choices, in an open-ended question. The analysis of the verbatim responses to this question provided information about the motivations underlying whites' and African Americans' preferences, and in particular whether they were driven by racial prejudice, social class, concerns about racial climate, etc. The open-ended questions provided empirical evidence about these motivations, where past research based only on the data from the closed-ended questions relied mainly on speculation and assumption about what shaped particular preferences. In addition, the open-ended data revealed nuances in how such explanations are articulated. For example, whites expressed three different kinds of racial stereotypes, which varied in their level of subtlety, when they were asked to explain their objections to integrated neighborhoods: respondents articulated stereotypes about (1) African Americans themselves; (2) the neighborhoods they lived in; and/or (3) the property values of such neighborhoods (Krysan 2002a) .

In-Depth Interviews In Concert With Traditional Surveys. The use of in-depth interviews to supplement and augment survey research is perhaps the least commonly used of these hybrid data collection methods. Although cognitive interviews—a specialized type of in-depth interview—have become increasingly common during the questionnaire development and pretest stages of survey research, in-depth interviews used (in conjunction with surveys) for mainly substantive purposes are relatively rare. The studies that do exist demonstrate that this combination has great potential. In the area of racial attitude research, Apostle et al. (1983) provide an example of one way to use in-depth interviews. They began their study of racial attitudes by conducting a series of in-depth interviews to help them identify white Americans' central "modes of thinking" about the causes of racial inequality. Based on the themes that emerged in these interviews, they designed a survey to allow them to establish the prevalence of various modes of thinking, and to gather the data necessary to test hypotheses the demographic and social correlates of different explanatory modes. The end result was a manuscript that wove quantitative and qualitative results together, thus capitalizing on the breadth and generalizability of the survey data and the depth and detail of the in-depth interviews to create a richer story about racial attitudes than either would have done on its own.

In addition to using in-depth interviews to conceptualize and design subsequent large-scale surveys, it is also possible to conduct such interviews during, or after, the survey data have been collected. A useful

example of this approach comes from outside racial attitude research. Pearce (2002) describes and illustrates an innovative hybrid data collection design where in-depth interviews are conducted after the data have been collected, and even after some initial statistical analysis has been completed. She uses these preliminary statistical models to identify "anomalous cases"—that is, cases where the model did not do a good job of predicting the individual's response to the key dependent variable (in this case, preferred family size). These anomalous cases constituted the sample for her subsequent in-depth interviews. Based on the in-depth interviews, Pearce (2002) was then able to identify additional "independent variables" that should have been included in her models. Thus, she revised existing theories about the influence of one of her key independent variables—religion in this case—and in doing so was able to include additional variables from her survey to improve the statistical fit of the models. Using in-depth interviews in these and other creative ways—before, during, and after survey data collection—is an exciting possibility, the potential of which we are only beginning to realize.

Summary

The initial impetus for survey researchers to use open-ended questions, experiments, and in-depth interviews in conjunction with surveys may have been largely methodological. And there is no question that these are, and continue to be, powerful supplements to surveys that have improved the quality of survey data, and advanced our understanding of survey research methods more generally. But as we look to the future and seek to understand substantive issues that are not only complex, but also more difficult to measure with traditional surveys because of their subtlety and sensitivity, it would be helpful to develop these hybrid data collection efforts that can augment standard survey methods. The result is not only a better understanding of puzzles in the survey data, but also a deeper and richer picture of the phenomenon of interest. That is, experiments, open-ended questions, and in-depth interviews have great potential as methodological tools for survey researchers. They do not displace surveys—rather, as illustrated in this paper, they both depend upon and build upon the results of traditional surveys.

REFERENCES

Allport, Gordon W. 1954. The Nature of Prejudice. New York: Addison-Wesley Publishing Company.

Anderson, Barbara A., Brian D. Silver, and Paul R. Abramson. 1988a. "The Effects of Race of the Interviewer on Measures of Electoral Participation by Blacks in SRC National Election Studies." Public Opinion Quarterly 52: 53-83.

Anderson, Barbara A., Brian D. Silver, and Paul R. Abramson. 1988b. "The Effects of the Race of the Interviewer on Race-Related Attitudes of Black Respondents in SRC/CPS National Election Studies." Public Opinion Quarterly 52: 289-324.

Apostle, RA, Glock CY, Piazza T, Suelzle M. 1983. The Anatomy of Racial Attitudes. Berkeley: University of California Press.

Bobo L. and James Kluegel. 1993. Opposition to Race-Targeting. American Sociological Review 58:443-64.

Bobo, Lawrence, James Kluegel, and Ryan Smith. 1997. "Laissez-Faire Racism: The Crystallization of a 'Kinder, Gentler' Anti-Black Ideology." Racial Attitudes in the 1990s: Continuity and Change, edited by Tuch and Martin. Greenwood, CT: Praeger.

Charles, Camille Z. 2000. "Neighborhood Racial Composition Preferences: Evidence from a Multiethnic Metropolis." Social Problems 47; 379-407.

Converse, Jean M. 1987. Survey Research in the United States: Roots and Emergence. Berkeley: University of California Press.

Davis DW. 1997. Nonrandom Measurement Error and Race of Interviewer Effects Among African Americans. Public Opinion Quarterly 61: 183-207.

Devine, Patricia G. 1989. "Stereotypes and Prejudice: Their Automatic and Controlled Components." Journal of Personality and Social Psychology 56: 5-18.

Emerson, M.O., G. Yancey, and K.J. Chai. 2001. "Does Race Matter in Residential Segregation? Exploring the Preferences of White Americans." *American Sociological Review* 66:922-935.

Hatchett, Shirley and Howard Schuman. 1976. "White Respondents and Race-of-Interviewer Effects." Public Opinion Quarterly 39: 523-528.

Hutchings, Vincent L., Nicholas Valentino, and Lara Rusch. 2001. "What Have you Done for My Group Lately?: The Effects of Subtle Racial Cues in Campaign Ads." Paper presented at the American Association for Public Opinion Research annual meeting. May, 2001.

Hyman, Herbert H. and Paul B. Sheatsley. 1956. "Attitudes Toward Desegregation." Scientific American 195: 35-39.

Hyman, Herbert H. 1964. "Attitudes Toward Desegregation." Scientific American 211: 16-23.

Kinder DR, Sanders L. 1996. Divided by Color: Racial Politics and Democratic Ideals. Chicago: University of Chicago Press

Krysan, Maria. 1998. "Privacy and the Expression of White Racial Attitudes: A Comparison Across Three Contexts." Public Opinion Quarterly 62:506-544.

Krysan, Maria. 2002a. "Whites Who Say They'd Flee: Who Are They and Why Would They Leave?" Demography 39(4): 675-696.

Krysan, Maria. 2002. "Community Undesirability in Black and White: Examining Racial Residential Preferences Through Community Perceptions." Social Problems 49(4): 421-543.

Krysan, Maria and Reynolds Farley. 2002. "The Residential Preferences of Blacks: Do They Explain Persistent Segregation?" Social Forces 80(3): 937-980.

Krysan, Maria and Mick P. Couper. 2002. "Race in the Live and Virtual Interview: Racial Deference, Social Desirability, and Activation Effects in Attitude Surveys." Unpublished manuscript.

Pearce, Lisa D. 2002. "Integrating Survey and Ethnographic Methods for Systematic Anomalous Case Analysis." Sociological Methodology 32(1): 103-132.

Schuman H. 1966. The Random Probe: A Technique for Evaluating the Validity of Closed Questions. American Sociological Review 31: 218-22

Schuman, Howard and Lawrence D. Bobo. 1988. “Survey-Based Experiments on White Racial Attitudes Toward Residential Integration.” American Journal of Sociology 94: 273-299.

Schuman H, Converse J. 1971. The Effects of Black and White Interviewers on Black Responses in 1968. Public Opinion Quarterly 35: 46-68

Schuman H, Steeh C, Bobo L, Krysan M. 1997. Racial Attitudes in American: Trends and Interpretations, Revised Edition. Cambridge: Harvard University Press.

Sears, David O., Jim Sidanius, and Lawrence Bobo. 2000. Racialized Politics: The Debate About Racism in America. Chicago: University of Chicago Press.

Sniderman PM, Piazza T. 1993. The Scar of Race. Cambridge: Harvard University Press.

Sniderman, Paul M. and Douglas B. Grob. 1996. “Innovations in Experimental Design in Attitude Surveys.” Annual Review of Sociology 22: 377-399.

Wittenbrink, Bernd, Charles M. Judd and Bernadette Park. 1997. “Evidence for Racial Prejudice at the Implicit Level and Its Relationship with Questionnaire Measures.” Journal of Personality and Social Psychology 72(2): 262-274.

A Non-Probability Sample of Comments on Studying Social Issues Using Survey Research, In-Depth Interview, and Ethnographic Methods: A Work in Progress

Samuel R. Lucas
University of California-Berkeley[8]

Some analysts have claimed that survey research and qualitative methods have distinctly different epistemological bases (e.g., Neuman 1991). The basis of this claim, however, seems rather thin. The two methodological approaches have certainly been used to study the same or strikingly similar phenomena--the relation between schools, students, and families (e.g., Lareau 2000; Coleman 1988); the tenuous hold low-wage and post-welfare workers have on economic stability (e.g., Newman 1999; Meyer and Cancian 1998); the role of significant others in student success (e.g., MacLeod 1995; Hauser, Tsai, and Sewell 1983); the impact of wealth on racial differences in opportunities (e.g., Conley 1999; Oliver and Shapiro 1995; Phillips, et. al. 1998); the role of racial diversity and social class diversity in the maintenance of secondary school tracking regimes (e.g., Wells and Serna 1996; Lucas and Berends 2002); the complex unfolding of and human response to disaster (e.g., Vaughan 1996; Kreps and Bosworth 1993), and more.

The questionable nature of the unique-basis claim becomes more clear when we consider that researchers working in "both" methodological traditions attempt to advance theoretical insights, to provide important substantive knowledge, and to propose policy responses relevant to the social issues they study. In short, researchers appear to have the same aims despite their different methodological approaches.

Further, and most important, researchers conduct their studies *in the same world*. The epistemologically-relevant imperatives of that world do not change when a researcher moves from one methodological approach to another. Hence, it would seem that, at the very least, social researchers, of whatever persuasion, must resolve the same problems. Although they may have different *technical* strategies for resolving or reducing the impact of those problems, the problems do not dissolve simply because a researcher has a different maintained assumption about their practice.

The observations above are preface to a set of points I posit are general.

Point 1: Survey Researchers may learn from Qualitative Research

Survey researchers specifically, and quantitative researchers in general, should recognize the limits of their work. At the most elementary level, the very existence of error terms in equations implies an incomplete understanding of the phenomenon under study. More fundamental, however, is that the very systematicity of quantitative methods should make clear the existence, if not the particulars, of a vast and perhaps infinite domain of approaches *not* tried, of digressions *not* pursued, of models *not* estimated.

In daily work many tame this infinite domain, in part, by focusing very precisely on a manageable subset of statistical problems. This strategy becomes more visible the more one reads the statistical literature directly, rather than the sociological translation of that literature. Most statistical papers study, say, heteroskedasticity, or measurement error, or varying parameter estimation, and so on. Many of these and other statistical issues open on to gigantic literatures of their own. Interestingly, however, it is the rare statistical treatise that seriously complicates the problems by taking on two or three or four. Unfortunately, it is that complexity that really animates the world of sociological research, more vexingly

[8]All work was conducted with the assistance of the Demography Department of the University of California-Berkeley. I thank Ann L. Mullen for comments on an earlier draft. All errors are those of the author. Please direct all correspondence to Samuel R. Lucas / Sociology Department / University of California-Berkeley / 410 Barrows Hall #1980 / Berkeley, CA 94720-1980 or via e-mail to Lucas@demog.berkeley.edu

because that research is conducted in the absence of solid knowledge about the phenomenon and its complex relation to other phenomena not directly under study.

One approach to this state of affairs is to encourage--or even beg--statisticians to turn their attentions a little more to the complex interplay of these challenges. This is a promising avenue, and one I would recommend.

An alternative approach, however, is to accept that statistical investigation of the complex interplay may be slow in coming and in any case insufficient in the end. If so, what may be needed is a fresher look, a more in-depth read, of the phenomena under study. Qualitative researchers endeavor to obtain that in-depth read, and they do so using methods honed in a variety of challenging research venues. Survey researchers may learn a great deal by routinely bringing a qualitative dimension to their designs.

Of course, those who have labored in the field *doing* survey research know that survey research already has a healthy qualitative component. Pre-testing of instruments often entails drawing in a small set of potential respondents, having them experience the instrument, and then engaging them, either one-on-one or in a focus-group type setting, in a discussion of their experience. Although gains from such efforts may not be documented in peer-reviewed publications on a routine basis, the practice has been sufficiently useful that it is a routine part of questionnaire design in many of the best survey research organizations.

The public in general, and many scholars specifically, remain oblivious to this almost routine merging of qualitative and quantitative methods. The rapprochement of research designers may be greatly furthered by publicizing this aspect of survey research work, by breaking down the belief some may have that some research approaches are antagonistic toward others.

Point 2: Qualitative Researchers may learn from Survey Research

The achilles heel of qualitative research is sampling. Although some make assertions about how the aims of qualitative research differ and thus probability sampling is not necessary (e.g., Miles and Huberman 1994), the introductory comments should make clear my skepticism about this claim to a unique aim. I am not sure about the basis of these claims--whether it is disdain for the challenge probability sampling would impose, misunderstanding about the gains that probability samples provide, or something else. But the claim is in error.

Absent probability sampling, we do not know whether we have identified some common feature of the population, or some idiosyncratic niche. Some may hear this as implying that qualitative research aims to generalize to the population, and might respond by contending that qualitative researchers often aim to generalize theoretically. But, when one seeks to generalize theoretically, what one is saying one wants to do is to draw inferences about the relationships among some theoretically-specified entities (henceforth called first-order inferences), and to link those first-order inferences to broader inferences about how the wider world works (henceforth called second-order inferences). What I must convey here is that certainly one may be able to draw first-order inferences by studying a non-probability sample. Those inferences pertain only to the sample one has, at best. The problem is that, if the sample is a non-probability sample, there is *no* basis for making any second-order inferences about the relationships one has studied.

An example may make the point. If one decides to study gender and academia, one might interview six female professors to whom one has access, and six male professors to whom one has access. One might discern that the female professors are quite a bit more professional in their demeanor, while the male professors are more casual. One has, therefore, discerned a relation between gender and demeanor. One might regard this as theoretically interesting, and might probe further. In doing so one might learn that the women have to deal with stereotypes students bring to class and colleagues bring to faculty meeting. One might learn that these stereotypes lead these persons to expect women to be nurturant and emotionally available. The women may therefore, in order to protect themselves from an unremitting hoard of colleagues and students expecting emotional availability, adopt strategies designed to highlight the professional aspect of the student-teacher and collegial relation. The men, in contrast,

freed from the constraints of that particular set of expectations, are at liberty to be casual, for the boundaries are far less likely to be crossed, and thus they are far less likely to become overwhelmed by lines out the door of students seeking emotional comfort rather than intellectual support.

In the best of circumstances one may draw all of the above first-order inferences on the basis of an in-depth qualitative study of a non-probability sample of male and female faculty. However, no matter how true the inferences turn out to be in the wider world, there is no basis for making any kind of generalization beyond the twelve faculty studied in the project. In short, second-order inferences about the relationship between gender and demeanor are indefensible.

Many researchers may not understand this fact, and part of the blame for the misunderstanding lies at the feet of those who teach sociological research methods to graduate and undergraduate students. To simplify matters we all-too-often demonstrate sampling theory with the example of the mean. Apparently, what many take away from these demonstrations is that if one wants to estimate an average, one should use probability samples, but if one is focused on relationships, probability samples may not be needed. This is a fundamental misunderstanding of the point teachers are trying to make. We can certainly improve our teaching by replacing some of the examples we already use with others that demonstrate the gains probability samples provide for the study of relationships of interest.

Fortunately, we need not wait for the entire discipline to be replaced by new cohorts who have been exposed to more nuanced training concerning sampling. Instead, we can capitalize on the realization that probability sampling is a routine element of survey research design. By merging a qualitative component to the survey research project, qualitative research can gain insights into just what is lost, and what is gained, by using probability samples. Further, these gains will be less costly, because the cost is borne in large measure by the survey component; only the subsampling for the qualitative design need be passed on to the qualitative component. The gains in knowledge will accrue as long as researchers use probability sampling to select which survey respondents will be studied qualitatively. Any other approach to respondent selection will forfeit the gains that come from combining qualitative and quantitative approaches.

Point 3: Improving Our Understanding of Social Issues Depends Crucially on *Methodological* Learning

A key unstated assumption of the foregoing is that understanding requires a solid defensible foundation on which to stand. This is very similar to another oft-unstated assumption of sociologists' work, namely, that social policy will be better if it is based on a solid defensible foundation of empirical facts and whatever theories are consistent with those facts (e.g., Duncan 1969). Some may regard the former assumption as a conceit academics have the luxury of adopting, and the latter as evidence of a frightening amount of ignorance of how power works in society.

But, as a scholar, it would seem impossible to do any research at all if one did not believe the second assumption, that knowledge increases either the chance social policy will become consistent with human flourishing, or increases human flourishing itself directly. One does not have to believe knowledge *decides* matters of policy to believe knowledge can influence the matters of policy.

The pathways through which knowledge can influence policy are many, and going into those pathways will take us far afield. But, to provide one example, voters base decisions often on taken-for-granted assumptions, many times read debates in light of those assumptions, and are unlikely to support any policy that violates those assumptions. Although voters do not set policy, they do influence who is placed in an office charged with setting policy. If sociological research can, over time, change what is taken for granted, it can change the background against which all policy prescriptions are read. In this one way, one way of many ways, knowledge may influence policy.

If that is to happen, however, one must consider the first assumption: understanding requires a solid defensible foundation on which to stand. This assumption becomes evident in the peer review process very clearly for journals. No matter how much one might believe the claims of the analyst, if those claims are not based on methods widely-regarded as appropriate, the claim is unlikely to see the light of

day in the journal. On the off-chance that the work is published, it is likely to be refuted either by comments to the journal or in later work that, in painstaking and sometimes painful detail, takes apart and reveals the weaknesses of the earlier work.

The reasons for this state of affairs are many, but for our purposes here, the main implication is that the *process* of investigation is the *foundation* for the claims. The analogy with the current correct functioning of the U.S. legal system is apt. If the police enter someone's home and find they have stolen your television set, yet they enter without probable cause, the evidence will (or should) be tossed out of court. If the police then obtain a warrant, and enter the person's office, and find they have stolen your computer, the evidence will (or should) be admitted. In both cases the person has committed a crime, but the process of evidence generation is acceptable in one case, and unacceptable in another. The analogy with sociological investigation is not exact, but it is close enough, for in law evidence is tossed out when inappropriate methods are used in order to preserve the rights of all of us; in sociology, evidence is tossed out when inappropriate methods are used to preserve the possibility that we will discern the true state of affairs. Given that as scholars we do not know what the true state of affairs is, we rely on methods that have a strong tendency to correctly reveal the state of affairs to us.

Hence, by bringing together different methods, it is quite likely that our ability to discern the hidden workings of the social world more effectively will increase. Bringing the different methods together can only improve our substantive knowledge, to the extent that the best of each approach is not lost. Seen in this way, the coming together of qualitative and quantitative approaches is a promising development for the investigation of substantively important social issues.

Concluding Remarks

There are probably many other lessons that each research tradition can learn from the other. The above should be regarded as examples, and, what is more, they should be regarded as simple examples. The examples are simple because there is so much misattribution, misunderstanding, and mistrust on the part of many practitioners of the various methods, that simple, robust, and easily demonstrated lessons each can learn from the other are required at the outset. Such simple lessons will increase mutual understanding and perhaps open the door to more complex, and perhaps more emotionally-difficult, lessons.

The long-term aim, of course, is to one day find that the qualitative/quantitative distinction has dissolved, as all scholars come to recognize the inescapable nature of the epistemological challenges we all face when we seek to go into the world and come out with something we can call knowledge. This day may be far-off, but it also may be closer than the walls that separate us may lead us to believe. This may be possible because, at a deep level, there is something quite optimistic and honorable about all of our efforts to wade into the messy mass of social interaction, social structure, and social power, and return with a story or two to help those we left behind when we began our investigation. We would do well to deepen our respect for all those who, from the same human starting point of ignorance and hope, but drawing on different technical tools, join in this thankless task. This deeper understanding, appreciation, indeed, empathy, is the ultimate aim of all the lessons the different but not really so different research approaches can teach and can learn from each other. And with such empathy, the walls of our understanding will tumble down, and we will better placed to study, describe, and understand the complex social world within which we all live.

References

Coleman, James S. 1988. "Social Capital in the Creation of Human Capital." *American Journal of Sociology* 94: S94-S120.

Conley, Dalton. 1999. *Being Black, Living in the Red: Race, Wealth, and Social Policy in America*. Berkeley, CA: University of California Press.

Duncan, Otis Dudley. 1969. “Inheritance of Poverty or Inheritance of Race?” pp. 85-110 in *On Understanding Poverty: Perspectives from the Social Sciences*, edited by Daniel P. Moynihan. New York, NY: Basic Books, Inc.

Hauser, Robert M., Shu-Ling Tsai, and William E. Sewell. 1983. "A Model of the Stratification Process with Response Error in Social and Psychological Variables." *Sociology of Education* 56: 20-46.

Kreps, Gary A., and Susan Lovegren Bosworth. 1993. "Disaster, Organizing, and Role Enactment: A Structural Approach." *American Journal of Sociology* 428-463.

Lareau, Annette. 2000. *Home Advantage, second edition*. Lanham, MD: Rowman and Littlefield.

Lucas, Samuel R., and Mark Berends. 2002. "Sociodemographic Diversity, Correlated Achievement, and De Facto Tracking." *Sociology of Education* 75: 328-348.

MacLeod, Jay. 1995. *Ain't No Makin' It: Aspirations and Attainment in a Low Income Neighborhood*. Boulder, CO: Westview.

Meyer, Daniel R., and Maria Cancian. 1998. "Economic Well-Being Following an Exit from AFDC." *Journal of Marriage and the Family* 60: 479-492.

Miles, Matthew B., and A. Michael Huberman. 1994. “Sampling: Bounding the Collection of Data.” Pp. 27-34 in *Qualitative Data Analysis: An Expanded Sourcebook* (2nd Ed.). Thousand Oaks, CA: Sage.

Newman, Katherine S. 1999. *No Shame in My Game: The Working Poor in the Inner City*. New York: Knopf/Russell Sage.

Neuman, W. Lawrence. 1991. *Social Research Methods: Qualitative and Quantitative Approaches*. Boston: Allyn and Bacon.

Oliver, Melvin, and Thomas Shapiro. 1995. *Black Wealth/White Wealth: A New Perspective on Racial Inequality*. New York, NY: Routledge.

Phillips, Meredith, Jeanne Brooks-Gunn, Greg J. Duncan, Pamela Klebanov, and Jonathan Crane. 1998. "Family Background, Parenting Practices, and the Black-White Test Score Gap," pp. 103-145 in *The Black-White Test Score Gap*, edited by Christopher Jencks and Meredith Phillips. Washington, DC: Brookings.

Wells, Amy Stuart, and Irene Serna. 1996. "The Politics of Culture: Understanding Local Political Resistance to Detracking in Racially Mixed Schools." *Harvard Educational Review* 66:93-118.

Vaughan, Diane. 1996. *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. Chicago, IL: University of Chicago Press.

Comments

Peter Marsden
March 29, 2003

I have been asked to speak about basic issues for survey research and "frontier" issues for the future. My comments come primarily from reflections on my involvement in the General Social Survey (GSS) project–a repeated cross-section survey–in a variety of roles since 1972.

Though much of this workshop is about innovation and change, I want to begin my remarks by sounding the decidedly unfashionable note that there is value to stability as well as change. The long-term objectives of the GSS project are to track aggregate change in the social indicators it measures over periods of time, and to ascertain the degree to which such change can be explained by cohort turnover and other demographic changes. The project has learned that most changes take place slowly, so that fulfilling these objectives depends on making observations over extended periods. By fixing such design features as the wording of questions and the context in which they are posed, the target population, and the mode of administration, the GSS makes it more likely that conclusions about change reflect the phenomena of interest rather than design features.

Additional virtues of stability include

1) ease of use for the thousands of professionals and students who make use of repeated surveys on a continuing basis, but may not track technical details closely;
2) the value of having a stable benchmark survey against which the results of surveys incorporating various innovations can be measured, or which can serve as a point of reference for studies with more limited target populations. (I note that this particular virtue of the GSS design was highlighted in remarks by AAPOR President Mark Schulman at a 30th Anniversary event held by the Eastern Sociological Society last month.)

That said, there are a variety of innovations that could and should–costs permitting–be introduced. One set of them has to do with the increasing ethnic diversity of the U.S. population, a topic of very widespread interest throughout sociology and other social sciences. To reflect this, continuing surveys like the GSS need to expand their target populations beyond the English-speaking. The clearest-cut case is that of Hispanic-Americans, about a quarter of whom speak Spanish only and are outside the GSS target population. Inclusion of non-English-speaking Asian-Americans would be more costly because of their smaller numbers and multiple languages. Mere expansion of the target population would be a modest step, however, since non-English-speaking Hispanics and Asians are sufficiently few within the U.S. that years of data collection would be required before we could make precise group-specific estimates for these groups. Occasional oversamples–of Hispanics, African-Americans (oversampled in the GSS during 1982 and 1987), or both–would yield great insight into these subgroups.

The GSS has long believed that external data on characteristics of the places where its respondents live could enhance the value of its individual-level data. Some such data are available through Census sources and have been linked to the GSS by many investigators, taking appropriate precautions to protect the confidentiality of respondents. The project receives approximately one inquiry per week for the necessary areal identifiers, but has never had the resources to assemble and distribute them as part of its public-release data files. Similar data are available on the industries within which respondents and their spouses are employed, and if linked to the individual survey could be useful in understanding survey content on stratification and work.

In addition to linking data from external sources, more specific data on the contexts that surround survey respondents may be assembled either as part of a survey itself, or in auxiliary studies. The GSS has measured aspects of the local social networks of its respondents in 1985, 1987, 1988, and 1998. Auxiliary studies gather contextual data that is more targeted than those available via Census sources. In 1991 and 2002, auxiliary National Organization Studies have surveyed the employers of GSS respondents, yielding data on the workplace rather than the more aggregated industry context. In 1998, a similar National Congregations Study assembled data on the places where GSS respondents attend religious services. Data from these auxiliary studies hold dual interest–partly as context for understanding factors that influence responses to the GSS, but partly in their own right for researchers interested in studying groups. Indeed, representative sample surveys like the GSS have proven to be an efficient device for drawing probability-proportional-to-size samples of non-listed groups with which GSS respondents are affiliated. GSS affiliates are convinced that similar auxiliary studies will yield insight into both neighborhoods and voluntary groups/associations.

Selecting topics to be included repeatedly in continuing surveys must place bets on what topics are going to retain their currency over long periods of time. Introducing new topics, or revising the content of sets of questions that measure existing topics, must be done with care to avoid compromising the utility of existing series. We believe that this is best accomplished through consultation with substantive specialists at all stages, as was done in designating the GSS "Core" in the early 1970s and has been done in the revision of its items on race relations and gender role attitudes in the 1990s. Both the standing Board of Overseers and *ad hoc* expert committees have contributed to such revisions.

Clearly the continuing ethnic diversification of the U.S. suggests that there may need to be a corresponding diversification of what the GSS measures about race and ethnic relations–to extend these measurements beyond questions about whites and African-Americans. We have ample basis for doing this in the topical modules on aspects of race and ethnic relations that have appeared in the GSS from time to time. Two fully new topics suggest themselves as subjects that perhaps should be monitored: the increasingly widespread use of information technology in aspects of American life, and popular views about genetic engineering and other scientific advances that may engender public controversy–though the pace of technological change in the information technology area may make it difficult to follow the GSS practice of stable question wording here.

Innovations in data collection methods will play some part in the future of the GSS, notwithstanding the project's reluctance to shift from face-to-face interviewing toward telephone administration (which is becoming decreasingly representative, and in any case cannot accommodate an interview as long as that in the GSS) or Internet administration (about which we harbor concerns having to do with representativeness). Recent GSSs have made use of factorial vignettes on health and mental health, among other experiments within surveys. At least one topical module has asked respondents to reply to questions after listening to tape-recorded accounts. With CAPI administration, numerous other possibilities including video and other visual stimuli exist and will likely be explored. One challenge in the use of such material lies in documenting the data gathered through these innovative techniques such that they are understandable and accessible to a large and broad community of users.

Measurement Error in Different Types of Surveys

Randall J. Olsen
March 2003

Let's start with first principles on measurement error: it is ubiquitous. If we can't deal with measurement error we are in the wrong business. But what does it mean to "deal with" measurement error? In this short paper, "deal with" means two things – minimize measurement error and then deal with the error that remains with statistical tools. The survey industry frequently focuses on factors such as recall error, conditioning and question order effects, and seam problems in event histories as the drivers of measurement error. I would like to briefly make the case that these are the least of our problems – problems that arise at any of three phases in the survey measurement process. My artificial taxonomy of phases is 1) conceptualizing the measure, 2) verbalizing the measure and 3) administering the measure.

Surveys generate measurement error in a variety of ways. Roughly speaking, there are as many different kinds of measurement error as there are ways of posing questions. We can reduce the dimensionality of the space of measurement error by listing some of the major ways data collection can go awry.

For the most part, measurement error infects cross-sectional, recurring and longitudinal surveys with similar virulence. However, longitudinal surveys offer more opportunities to deal with the problem.

Sources of Measurement Error

Data collection is an indirect conversation between the scientist and the respondent. We have to start with the ability of the scientist to formulate a conception of what phenomenon we are interested in. One of the more insidious sources of measurement error has nothing to do with cognition or recall. If we fail to grasp the heart of the social science problem we are investigating error is inevitable. Consider two of the more difficult domains for data collectors – income and assets. Income modules measure the flow of purchasing power into an individual, family or household over some well-defined period of time. Asset modules measure the value of financial instruments and real property held by individuals, families and households at one particular time. Without the discipline of a framework that distinguishes the flow rate of purchasing power into a consumption unit from the measure of a consumption unit's accumulated stock of real goods and productive assets (or financial instruments that represent claims on real goods and assets), questions about the economic status of a consumption unit can degenerate into a formless mass of interrogatories that leave conceptual gaps that the social scientist will be hard-pressed to deal with.

Measurement without Theory. Measurement error is inevitable without a firm grasp on what is being measured. To be sure, if the agenda of the survey is to count the number of television sets in a household, conceptual problems play a minor role. When we speak of "measurement error" we presuppose that we are dealing with something that is, in fact, measurable. My wife is unable to answer which of the dogs she has owned she loved the most. This does not augur well for an attempt to "measure" on a five-point scale how much she loved each dog. Mathematicians attach a set of very particular conditions to the concept of measure, among them the ability to place the outcome in correspondence with the real number line. Any attempt to appropriate and apply the term "measurement error" to a concept that defies measurement may end in unhappiness.

Surveys provide a stimulus and record a response. We can always record that response even if we are dealing with a reaction that does not conform to the rigorous framework of "measurement". All the social sciences deal with attitudes at some point, and psychologists have provided ample evidence the responses to attitudinal questions reflect variations in many dimensions of context. In a polar case where

there are non-attitudes, as Converse hypothesized, we may only be measuring the context. In such a situation the term "measurement error" seems strangely out of place. Put another way, if a person's "true" attitude, when expressed on a five point scale, does not have a "true" value, but has probability mass spread over all five points, measurement becomes a very complex matter, and if we are recording a single response to that five point scale, the term "measurement error" becomes meaningless.

Survey research is a big field and I must be brief. Consequently I will despair of dealing with survey topics that have no natural metric and for which we cannot readily speak of a real-valued function that represents the distance of a response from its true underlying conceptual value.

Even when the designer has a firm grasp on what it is they mean to measure, there is the risk that in recurring or longitudinal surveys a new designer will step in to make "improvements" that destroy a scale, create identification problems or alter a key construct. I have seen parts of longitudinal surveys analytically destroyed by repeated efforts to "improve" matters. This is not a business for short attention spans.

We are Dealing with People Here. The second step in the process is to translate the conceptual goals of the designer into question wording. This can be very tricky. Often the concepts we are after are difficult to communicate to the respondent. This gives rise to a difficult compromise between asking a question the great bulk of respondents can understand, and asking a question that is useful to the social scientist. In economics, questions about time preference or attitudes about risk can involve hypothetical choices between complex alternatives. Some respondents cannot deal with these questions, but simpler questions may be analytically useless. The very complexity of what we seek to measure can preordain us to measurement error. Let me give another example. An interview that will remain nameless posed a number of questions to respondents about whether they cohabited. Comments that came back indicated many respondents had no idea what the word "cohabit" meant. However, their weak vocabularies were easily matched by the social scientists who wrote the question who, in discussions, were pretty much unable to articulate what they meant by cohabit. Their inability to put a conceptual fence around the word "cohabit" mooted the inability of the respondents to understand the word and hence state whether they are inside or outside that particular rhetorical fence. Even if the people writing the question came up with an easier word than "cohabit" there is sufficiently wide variation among social scientists just exactly what "cohabit" involves that there is a real danger that different users will interpret the same data differently. Is cohabit a euphemism for being a sexual partner? For pooling economic resources and sharing household duties in a particular way? We don't always think of measurement error in terms of conceptual confusion about some domain of human activity, but as we study more complex and subtle behaviors, what we like to think of as objective questions can become subjective after we are done interpreting the data.

So, after talking about two of my arbitrarily chosen three steps in the survey process, we have focused primarily on conceptual problems. This is my overly cute way of saying that in surveys the biggest part of measurement error lies at our feet because before we go to the field we have not considered with sufficient care why we are collecting the data, how we intend to model human behavior and what sorts of statistical techniques we intend to use. Having repeated cross-sections or a longitudinal survey does not cure these problems.

Cognition. Third, the survey is put into the hands of the interviewer to conduct a conversation with the respondent. At this point the measurement error problems become more conventional. The interviewer-respondent dyad uses the written questions and jointly measure the concept contained therein. It is here where cognitive mis-understandings dominate the measurement error process, although we should not downplay the ability of an interviewer or respondent to transmogrify a simple question on topic A into a new question on topic B.

What to Do About Measurement Error

It is at this third point that what we commonly call measurement error comes into play. Provided our conceptual ducks are in a line, we can apply contributions from a variety of disciplines over the last half-century to ameliorate the problem. I will group these palliatives into three categories: reduce the error, calibrate the error, and model the error.

Technique. The social psychologists have made perhaps the biggest contribution to error reduction. They have identified many ways in which the interviewer/respondent conversation can generate error due to poor question wording and context effects. Their insights have been invaluable even when we ask fairly straight-forward factual questions. The way people think about answering and the limits on respondents' cognitive powers impose severe limits on our questions if we hope to reduce response error. I am a neophyte here, so I will defer to the social psychologists and move on.

Calibration. The next approach to dealing with measurement error is to calibrate the error. Calibration is important because most of the response error patterns identified in the literature do not simply result in the addition of an error component with a mean of zero, but generate systematic over- or under-estimates. External data resources can help us understand the aggregate degree to which respondents under- or over-report. For example, based on aggregate financial data, we know that survey respondents typically under-report their holdings of assets. By inflating reported asset amounts by some fraction, we can produce asset data that match the aggregates but still leave us with matching data on covariates such as demographics, location, income occupation and work history that are unavailable in flow of funds and national income accounts data. Another variant of this strategy is to match to records data. Studies of welfare recipiency run up against the fact that respondents are sometimes reluctant to admit they received welfare payments, and when they do, recall them poorly. The ability to supplement survey data with records data yields databases of unusual strength. Matching is not always a superior strategy however. Among criminologists there is considerable lore that people are willing to admit to more arrests and crimes than official records reveal. However, records matches on arrests and so forth appear to be the exceptions that prove the rule that matching is an important adjunct to survey work. This leads to a suggestion for NSF: privacy legislation has made records matches exceedingly difficult and survey researchers are not nearly the threat to confidentiality posed by others. A less punitive approach to would-be record matching would have a substantial payoff even if it were only done to help us calibrate our measures.

Modeling. Finally, statistical measurement error models have a long tradition in some of the social sciences. Economics has a long tradition of explicitly modeling error components in measures and then determining the statistical implications. The sociological literature also features extensive work on models where and underlying theoretical construct is made manifest through a set of observable variables. One of the more frequently used techniques for dealing with measurement error is known as instrumental variables whose pedigree goes back to the Cowles Commission and two-stage least squares and the k-class estimator. A common feature of these techniques is that one can deal with the measurement error problem if one has multiple variables that measure the true construct with error. The finite sample statistical properties of these estimators are not as straight-forward as least squares. However, many social scientists use large secondary data bases and, for well conditioned data, large sample properties are enough to support statistical inference. Instrumental variable and multiple indicator approaches to dealing with measurement error are better developed for regressions and regression-like models. When surveys only seek to recover mundane measures like the population mean, we can run into trouble. When we cannot calibrate or model, it may be the best we can do is stick with stable measures over time and emphasize the *change* in our measures over time rather than their level, hoping that the measurement error problems are at least stable. This presupposes these stable measures are conceptually sound.

Longitudinal versus Repeated Cross-sections versus Cross Sectional

Getting to the point at long last, how do measurement error considerations play into the choice of cross-sectional, repeated cross sections and longitudinal data? First, for pure database utility, we are almost certainly better off with 2,000 variables on 10,000 persons over a two year period than 1,000 variables on 10,000 separate persons in each of two years for a total of 20,000 people. Recall that standard errors fall with the square root of sample size. In terms of model complexity, if we restrain ourselves to five variables per regression, the longitudinal data will support 2,000 choose 5 regressions, whereas two repeated cross sections will support 1,000 choose 5 regressions, or about $1/30^{th}$ as many. More importantly, for studying measurement error, having repeated measures on the same construct is important given the premium attached to multiple measures, each subject to error, in constructing Instrumental Variable or LISREL models. In addition, because a conceptually sound model that tells us what we should measure is so important, large numbers of variables trump large numbers of observations in formulating a complete model. If we must despair of measurement without error, dealing with a stable population of respondents will reduce one source of variability to repeated measures, and if the questions we pose to respondents are somewhat complex, we may find that greater time in sample yields the benefit of a well-trained respondent who is better prepared to answer our questions; perhaps we should speak of time in sample bias reduction.

It's Mostly Our Fault

Given my emphasis on the principal investigator as the principle culprit in measurement error, design needs to be forward-looking to the analysis stage. The sorts of databases being considered here should not be built to produce eye-catching statistics for USA Today. If one is interested in univariate descriptive statistics, the CPS, Census or even a one-off telephone survey will do the job. There is a lot of Gallup Poll-class stuff out there. Major social science surveys need to think through from beginning to end just exactly how we learn from data.

NSF Workshop on Repeated Cross-Sectional and Longitudinal Surveys
March 28-29

Frank Stafford and Bob Schoeni
Institute for Social Research, University of Michigan

Session: Survey Frontiers

We have been asked to provide some brief comments on "survey frontiers," with focus on economic issues. We describe a few frontiers on two domains: those involving survey content and sample, and those involving measurement and distribution of data.

Innovation in measurement and distribution of data

Event History Calendar (EHC)

Both the scientific need for information on past life events and budget constraints have stimulated a need for data collection methods that improve the accuracy of reports of historical information. Consistent with a life course perspective, researchers seeking to understand choices and outcomes made today need information from the distant past. While longitudinal data reduce this need, new sample members of panel surveys do not typically enter the panel at birth, therefore the need still exists to obtain information on their earlier life events, and perhaps the events of their parents. Moreover, some surveys are interviewing sample members less frequently – every two years instead of annually – but yet want high quality data on all important events at all times during the longer interval between assessments. EHC is one method that has been found to enhance the accuracy of reports of historical information.

Easy to use customized extracting and codebook generation via the internet

Technology exists to significantly reduce the resources that are need to take complex panel survey and create customized data extracts for each individuals researcher's specific needs on a given project. A complaint of panel data is that the fixed costs of learning how to use the data is sometimes high, but innovation in distribution can and has reduced this burden, stimulating more research. An added benefit of this approach is that it can much more easily be used for educational purposes in, for example, undergraduate courses that seek to enhance new scientific endeavor among young students.

Linking administrative data to survey data

A wealth of information is available through administrative records such as Medicare utilization, Social Security earnings, and NDI cause of death. Over the past several years there have been significant efforts to link such information to survey data. While raising confidentiality concerns, the scientific advantages of having such data are significant. Continued efforts on this front are worth pursuing.

A few other innovations that have taken place in recent years that need to continue to be developed and used include internet interviewing, use of mixed modes to reduce costs, using of unfolding brackets, and more closely linking the sections of instruments that obtain information on income and assets to reduce miss-reporting.

Innovation in content and sample

International data

Continued globalization has stimulated a need to better describe the differences across countries and understand their underlying causes. Almost all developed countries, as well as many developing countries, now have panel surveys of families or individuals. Harmonization across surveys should be explored further to enhance our ability to learn from experiences across countries. Indeed, many panel surveys in other countries have used panel surveys in the US as a model, which means that in many cases harmonization with existing panel surveys across the world is already underway.

Children

Our nation has relatively little survey data on the situation of children, including the factors that influence their choices and outcomes. Including this information within the broader context of a panel survey of the families and individuals with whom children spend their lives is of high priority. Panel survey efforts such as the NLSY, Fragile Families, and the PSID Child Development Supplement address some of this gap.

Organizational information

We have relatively little data that allow us to examine the situation and change of key organizations as well as the consequences of the organizations and the decisions they make on individuals within those organizations. One can fill this gap by sampling organizations – firms, for example – or one can interview organizations that population-based panel members interact with. This latter approach has been used in some surveys, including the PSID and HRS, where firms have been contacted and interviewed to obtain information on pension options provided by the firm. Efforts such as this could be expanded to cover other aspects of organizations.

Interdisciplinary domains

A greater degree of integration of disciplinary approaches is needed, including data to support such efforts. Economic data needs to be combined with geographic/spatial information, health, including clinical measurements, and technology. This calls for a broader, lengthier survey with links to a variety of contextual data sets and sometimes expensive health measurements.

Special Issues for Over-Time Designs
March 28-29, 2003

Marianne Stewart
School of Social Sciences,
University of Texas at Dallas
e-mail: mstewart@utdallas.edu

An Overview

My perspective on special issues for over-time designs is based on two sources. One is engagement in the primary data collection associated with the 1974-79-80 Canadian National Election Studies, the 1992 and beyond Party Support in Britain project, and the 2001 British Election Study (BES). Another is involvement in secondary data analysis of the key determinants of presidential approval in the United States and party support in Britain and Canada using aggregate-level time-series data mainly collected by the Gallup Organization. However, the principal reference here is the 2001 BES.

The British Election Studies (BES) project celebrates its fortieth birthday this year. The first BES survey (1963) was conducted by David Butler (Nuffield College, Oxford) and Donald Stokes (University of Michigan) as part of a baseline study for the 1964 post-election survey. Since then, BES surveys have been undertaken after all general elections, thereby providing a series of 11 consecutive data collections. The project has been housed at two universities in the UK: Nuffield until 1974, Essex until 1983, Nuffield until 2001, and Essex again in 2001. The principal investigators of the 2001 BES are David Sanders and Paul Whiteley, University of Essex, and Harold Clarke and Marianne Stewart, University of Texas at Dallas. BES funding typically has been provided by the Economics and Social Research Council (ESRC) of the UK, which is the British equivalent of the National Science Foundation (NSF).

BES Design Features

Post-Election, In-Person Interviews: Since the early 1960s, the principal BES surveys have used in-person interviews (now CAPI), with fieldwork starting right after a general election. The 2001 BES also used post-election, in-person interviews to maintain design continuity as stipulated by the ESRC call for proposals.

Large Sample Sizes: BES tend to have relatively large sample sizes. For example, the 1992, 1997, and 2001 post-election Ns are 3534, 3615, and 3219, respectively. These sample sizes:

1) enable the analysis of the regional bases (England, Scotland, Wales) of political attitudes and behavior;
2) facilitate multi-level modeling of contextual effects on voting and local political activities, including party constituency campaigning;
3) permit, as appropriate, the use of panel designs for the purposes of tracking individual-level change and disentangling causal processes.

Multi-Wave Panels: The BES core datasets have been the post-election, cross-sectional surveys. However, several datasets also have had significant panel components. Initially, the panel design was motivated by Butler and Stokes' interest in political change, and, subsequently, this theme has informed much work on electoral choice in Britain. Most BES panels involve reinterviews with respondents from previous BES post-election surveys, but several also have been conducted during inter-election periods. The latter include 1963-64-66-69-70; February 1974-October 1974-75-79; 1983-86-87; and pre-election 2001--post-election 2001--2002/03.

In the 1990s, the amount of relevant panel data was augmented by initiation of the British Election Panel Studies (BEPS) at Nuffield's Center for Research on Economic and Social Trends (CREST). The BEPS design involves annual reinterviews with large national panels. The first BEPS panel covered the 1992-97 period (1992-94-95-96-97) and the second one covered 1997-2001 (1997-98-99-2000-01). Since the BEPS surveys include a number of key questions from the BES, the former provide significant additional power for addressing important topics about the short- and longer-term dynamics of party support. Other longitudinal studies that have carried some politically relevant questions (e.g., party support) are the British Social Attitudes surveys since the early 1980s and the British Household Panel Study (similar to the PSID) which is a long-wave panel focusing on UK households.

The 2001 BES

The 2001 BES was designed to include multi-wave panels for the purpose of addressing four important questions: the decision (not) to vote, the electoral choice that is made, the factors that explain the election outcome, and the role of elections in democratic political life in contemporary Britain. The design elements are:

1) Pre-Election Campaign Baseline Survey, Post-Election Survey, and Pre-Post Election Panel: The pre-election (CAPI) survey (N=3209) was conducted in the two months before the official campaign began. A panel (N=2315; 72.1%) of the pre-election survey respondents was re-interviewed after the election as part of the traditional post-election (CAPI) survey. "Top-up" interviews bring the total N for the post-election survey to 3035. Survey work by National Opinion Polling (NOP).

2) Rolling Cross-Sectional Campaign Survey With Pre-Post Election Panel: Over the course of the (approximately 28-day) official campaign, interviews were conducted by telephone (CATI) with 4810 respondents (average daily N ~ 160). A panel of 3751 respondents was re-interviewed immediately after the election. Survey work by The Gallup Organisation (UK).

3) Mid-Term Panel: Participants in the 2001 post-election (CAPI) survey were reinterviewed by telephone in late 2002-early 2003, producing a panel of 1157 respondents and additional top-up interviews of 850, for a total of 2007. Survey work by NOP.

4) Internet Survey: A post-election survey was conducted on a no-cost basis, producing a N of ~7000. Survey work by You.Gov.

5) Media Content Analysis. A content analysis was undertaken of major British newspapers during the campaign period to track variation in thematic coverage during the election campaign. Survey work at University of Essex.

6) Vote Validation Study: Respondents' reported voting was checked using official records. The records for England and Wales are available to the public at the Lord Chancellor's Office in London and those for Scotland at local sheriffs' offices. Checking at University of Essex.

Linkages to Other Surveys: Key questions in the 2001 BES such as party identification, economic evaluations and emotional reactions, party leader performance, vote intention, and other forms of political activity have been asked every month in a set of CATI surveys conducted by Gallup on behalf the principal investigators. Some of the questions have been asked since January 1992, with funding from NSF (the Party Support in Britain project), and others on attitudes towards democracy and participation were added in 2000, with support from ESRC (the Attitudes Towards Democracy and Participation

project). The data (N=127,500, ~130 repeated cross-sections) provide the ability to study factors driving the aggregate dynamics of party support.

Relative (Dis)Advantages of Repeated Cross-Section and Panel Designs

Advantages of Repeated Cross-Section Designs:

1) Most political attitudinal and behavioral research is observational, not experimental. Accordingly, microclimates or laboratory settings with variable administrations of measurements can rarely, if ever, be created. Rather, we often must wait for the world to emit an event whose signal can be measured by our instruments for the purpose of assessing its impact on a political outcome variable of interest. Repeated cross-sections can be very useful, particularly for capturing effects that reverberate at the time that a politically relevant decision must be made, and for capturing quick upward or downward movements characterized by, for example, the strong pulse-decay effect that tends to occur in public opinion. In these respects, the Party Support in Britain project was able to capture the effect of the 1992 currency crisis in the UK on governing Conservative and opposition party support, an event that not only cost the Conservative Party control of the government in 1997 but also has continued to echo 10 years later with the ongoing ebb of public confidence in Conservative economic management. More recently, the same project, with its Attitudes Towards Democracy and Participation companion, picked up the effect of the then-impending war in Iraq by recording strong downward movements in not only economic evaluations, particularly expectations, but also key party support variables, including support for Labour and approval of its record in government.
2) Unfortunately, panel designs are cost-prohibitive "Voyager-like" instruments for this purpose. They are particularly costly in terms of the time that they can take to get into the field, thereby overshooting the effect of the event, or the amount of time that they can to take to get out of the field, thereby capturing the effect, if any, when it may be little more than faint noise. In sum, panel designs may be broadbanded.
3) Cost-effectiveness of administration is particularly achieved when the data collection is subcontracted to an experienced and highly reputable survey organization that is positioned to undertake telephone, mailback, and/or internet interviewing.
4) Since there is a general lack of ambiguity in unraveling the temporal sequence of causal processes over time, theoretical claims can be tested for at least weak exogeneity, rather than simply stimulated as in the case of single, cross-sectional designs.
5) Problems of panel recruitment, panel retention and attrition (some original respondents could not be reinterviewed), and panel conditioning (the original interview experience affects a subsequent one) are less severe.

Advantages of Panel Designs:

1) As is well known, panel designs can be well-suited for the purpose of developing and testing (rather than inferring) models of individual-level change in attitudes and behaviors of the same individuals over different points in time. This is very important since it is hypothesized, and often the case, that aggregate-level patterns of stability or change may not be replicated at the individual level. For example, "big-bang" movement in party support, often associated with partisan realignment, may very well be interpreted as the product of more slowly paced, but highly discernible, "dealignments of degree" that have transformed the UK electorate.
2) Panel data enable the development and testing of measurement error in survey responses and, thus, the readjustments that may be necessary for reliable estimation of structural models. But, in this regard, it may be argued that a truly effective panel design requires truly large Ns. This may especially be the case with mixed Markov modeling of the non-zero versus zero

probabilities of moving or staying with a particular attitude or behavior for different groups, while taking account of random measurement error. For example, if an individual has three choices at four time points (panel waves), then there are 3x3x3x3 = 81 possibilities for that individual or cells into which s/he could place. In the event that we also are interested in generalizing or making inferences and in avoiding the sparse cell problem, we may require ~100 cases per cell and, thus, 8100 respondents which again is very costly.

3) Both repeated cross-sections and panel designs spare researchers the distorted results in attitudes and behaviors that typically are produced by post-hoc rationalization or recall in survey questions. In the case of the Party Support in Britain Project, for example, the monthly monitoring of (1992) vote recall between 1992 and 1997 would suggest a Labour landslide in 1992, which, of course, was not the case.

Disadvantages of Panel Designs:

1) Panel attrition (again, the inability to reinterview original respondents in a subsequent wave due to refusal, address change, etc.) is associated with several problems, one of which is sample or selection bias. That is, respondents who survive to wave 2 may not be the same as those originally included in wave 1, with the former having a variety of skewed social (e.g., more educated) and political (e.g., more interested) characteristics. As is well known, efforts have been made to take account of attrition when using panel data, with Heckman selection bias and related (e.g, Achen, 1987) models. Of two simultaneously estimated models, one may be discrete choice, i.e., how people got into the panel, that is panelists and non-panelists, whereas the other is a dependent variable of interest in a logit, probit, or alternative setup. The discrete choice model will include characteristics of respondents (e.g., age, gender, etc.) and characteristics of the interview (such as length of interview or whether it is conducted at night or on the weekend, thus dampening the probability of being in the panel). Selection bias models have been applied not only in economics but also in education research (e.g., GRE as predictor of graduate school success) and international relations (dyadic relationships and conflict).
2) Panel conditioning – not unrelated to a time-in-sample problem – refers to the initial interviewing experience affecting the reinterview, thereby posing threats to inference of attitudinal and behavioral stability or change in the general population. Panel conditioning may be exacerbated by mode, notably when random panel assembly in internet mode is not designed for sufficient retiring of existing panelists or reseeding or topping-up of the panel. As is well known, internet polls are susceptible to sampling bias and, in the case of the UK, their respondents simply do not come off the electoral register or postal address file (for face or mailback interviewing) or RDD (for telephone).

A variety of procedures has been developed for dealing with both panel attrition and, to some extent, panel conditioning. One is reseeding or topping up, that is, retiring panelists and replenishing the panel with new respondents who, in turn, constitute fresh recruits for the next wave (but, on statistical estimation of the size of such panel effects when panel data are embedded in cross-sectional designs, see Bartels, 1999). Another is control over characteristics of the interview, notably its length and timing. A third is introduction of fresh questions on relevant topics of interest being interspersed with important questions used to track cross-time change, but without introducing question-ordering effects. This may reduce respondents' working memory of the content and ordering of previous instruments. A fifth is, of course, "fee for service," that is, the use of a variety of incentives from small cash payments to token items to charitable donations to lotteries. The ability to develop these and other procedures further not only means more continuity and innovation in important datasets designed to capture public attitudes and behaviors. It also serves to increase the inferential benefits of panel data over the inferential costs of panel effects (Bartels, 1999), where the latter bedevil some types of questions more than others.

The Challenge of Rising Nonresponse Rates

Roger Tourangeau
Joint Program in Survey Methodology, University of Maryland
Survey Research Center, University of Michigan

One of the great problems faced by surveys generally these days but particularly by surveys repeated over time is the problem of maintaining survey quality in the face of rising nonresponse rates. Many, many variables affect response rates (including the mode of data collection, the length of the data collection period, the number and timing of the attempts to contact members of the sample, whether incentives are offered and, if so, their amount, the topic and sponsorship of the survey, the length of the questionnaire, and the data collection organization, just to name a few of the key ones), but a wide range of evidence encompassing surveys with many different methodological features suggests that response rates are generally falling both in the United States and throughout the developed world. The industry has responded to this crisis by redoubling its efforts to stem the tide, through increasing use of incentives and other means to maintain response rates. Nonetheless, most observers agree that the trends in response rates are downward and the trends in data collection costs are upward, creating a double bind for survey research organizations and the organizations that fund surveys. Surveys have to try much harder just to stay in the same place. Bob Groves's paper will focus on the implications of this trend for survey costs; my presentation will focus on the evidence for and implications of falling response rates themselves.

Trends in nonresponse rates. As Groves and Couper (1998) argue, it is useful to distinguish among three main forms of nonresponse — nonresponse due to noncontact with the sample unit, nonresponse due to the sample unit's refusal to cooperate, and nonresponse due to the sample unit's inability to participate (e.g., inability to complete the interview in English). There are good grounds for concern on all three fronts.

Consider, first, the mounting difficulties of contacting potential respondents. With households telephone surveys, Caller-ID, call blocking, and answering machines have created formidable and widespread barriers to contact in household surveys (see Oldendick and Link, 1999; Tuckel and O'Neill, 1996, for some recent results). These technologies have caught on quickly — the majority of American households now have answering machines, Caller-ID, or both, and research indicates substantial numbers of households use them to screen out unwanted calls. Many survey professionals report anecdotally that telephone response rates have plummeted over the last decade or so. Although the published evidence for this dramatic downtown is not so clear, Steeh and her colleagues (Steeh, Kirgis, Cannon, and DeWitt, 2001) show that it takes an increasing number of attempts to reach households selected for the Survey of Consumer Attitudes (SCA) and the Georgia State Poll. The average number of call attempts for the SCA peaked at around 12 in 1999, the last year Steeh and her colleagues examined; in the mid-90s, the figure was close to six calls. Again, although the evidence is largely anecdotal, many survey researchers believe it is also much more difficult than it used to be to contact potential respondents in face-to-face surveys; more people live in gated communities, institutions, doormen buildings, or other settings that shield residents from unwanted visitors than lived in such settings ten or 15 years ago.

Steeh and her coworkers (2001) argue that any recent fall in telephone response rates probably reflects greater difficulty in reaching telephone households rather than greater reluctance to participate, but the opposite seems to be the case for face-to-face surveys. Nonresponse due to refusal to take part in surveys seems to be rising for household surveys all over the developed world (Groves and Couper, 1998; de Leeuw and de Heer, 2001). With some Federal surveys in the United States, the problem of rising nonresponse rates is somewhat obscured, since the surveys make an increased number of callbacks and, as a result, fewer sample cases are never contacted. Even in these surveys, though, the refusal rates appear to be rising (see, for example, Atrostic, Bates, Burt, and Silberstein, 2001), driving up the overall nonresponse rates. Figure 1 below displays the total nonresponse rates and refusal rates for two major Federal surveys—the Consumer Expenditure Survey and the National Health Interview Survey (NHIS).

Both showing increasing overall nonresponse rates and, in both cases, the overall increase mainly reflects a rise in the rate of refusals. The NHIS data also suggest that noncontact may be increasing in that survey.

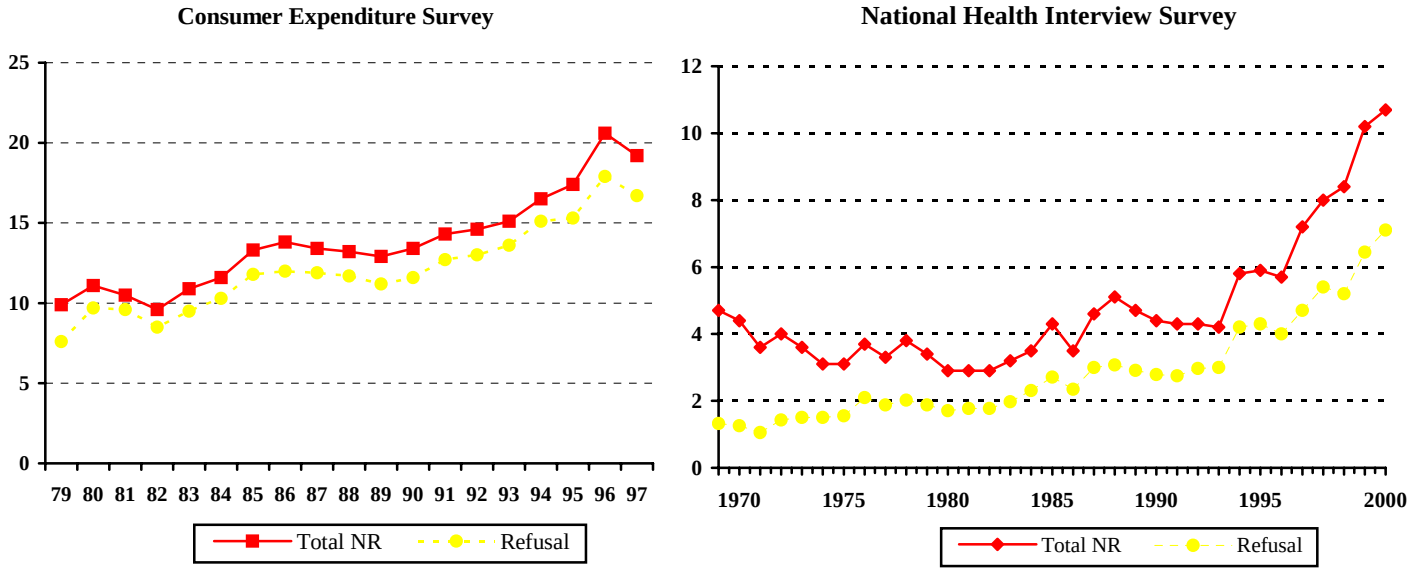


Figure 1. The total nonresponse and refusal rates for the Consumer Expenditure Survey (left) and the National Health Interview Survey (right). Source: U.S. Census Bureau (via Bob Groves).

Still another cause for concern involves long-term shifts in the demographic makeup of the U.S. population, specifically the rising percentage of foreign-born and elderly Americans. In 2002, 11.5% of the U.S. population was foreign-born; according to Long Form data from Census 2000, 8.1% of the population over age five reported that they speak English less than "very well." Many surveys now field both Spanish and English questionnaires, but only two-thirds of those who are less than completely fluent in English are Spanish speakers. Similarly, the rising proportion of the population who are over 65 years of age suggests that an increasing percentage of sample persons will be unable to complete interviews because they have hearing problems or other physical infirmities that prevent them from participating.

These mounting difficulties in reaching people, persuading them to take part, and enabling them to complete the interview add up to rising rates of nonresponse; the problem affects both telephone and in-person surveys, but it's probably worse for telephone surveys — currently, very few telephone surveys achieve response rates higher than 60 percent. The problem affects both repeated cross-sections, like the Consumer Expenditure Survey and the NHIS and panel designs, like the Current Population Survey. Panel attrition — losses due to nonresponse in the second or later waves of a longitudinal surveys — compound the effects of nonresponse in the first wave, especially in surveys that make no attempt to include initial nonrespondents in later rounds.

The impact of nonresponse. So the response rate picture is bleak. It is an article of faith among survey researchers that lower response rates mean higher nonresponse *bias*, but several studies suggest that the relationship between the two is more complicated than most of us thought. The bias due to nonresponse in a mean or proportion depends both on the rate of nonresponse and the covariation between the response propensity (the probability that a member of the sample will respond) and the substantive variable of interest:

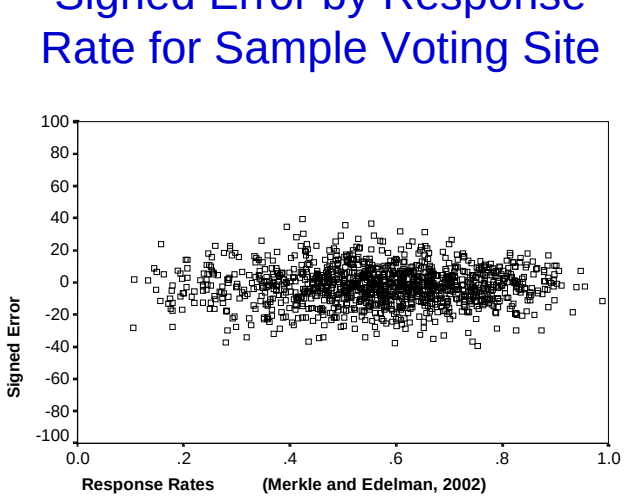
$$Bias = Cov(r, y) - (1 - \bar{r})\bar{Y}$$

in which r is the response propensity, y the substantive variable, $1-\bar{r}$ is the expected nonresponse rate, $\bar{Y}$ is the overall population mean or proportion. When nonresponse is viewed as deterministic (you're either always a respondent or always a nonrespondent given a fixed survey design), the bias formula reduces to:

$$Bias = (1 - \bar{r})(\bar{Y}_r - \bar{Y}_n)$$

in which $\bar{Y}_r$ is the mean for the respondents and $\bar{Y}_n$, the mean for the nonrespondents. The latter formula indicates that when the nonresponse rate is zero or when the respondents don't differ on average from the nonrespondents, there is no bias. The bias formulas are more complicated for analytic statistics, like subgroup differences or regression coefficients.

Three recent studies show that sometimes there is little relationship between nonresponse rates and nonresponse error (Curtin, Presser, and Singer, 2000; Keeter, Miller, Kohut, Groves, and Presser, 2000; Merkle and Edelman, 2002). Each of these studies has its limitations, but taken together they present a strong case that larger nonresponse rates don't necessary signal larger biases. Keeter and his colleagues contrasted two telephone survey protocols that differed in the length of their field periods and the number of callbacks they included in a randomized experiment; the two experimental groups differed markedly in response rates (60.6 versus 36.0 percent), but only 14 of 91 variables showed significant differences by experimental group and most of these were quite small. Merkle and Edelman tracked response rates and overall error in exit polls at sample precincts across four election years. The results (displayed in the figure) indicate no correlation between the two. They argue that the factors that produce nonresponse in exit polls (e.g., whether official allowed the interviewers to stand near the entrance to the polls) are completely unrelated to voting preferences. This study is important because the population value being estimated — the vote at the precinct — is known. Curtin and his colleagues looked at the Index of Consumer Sentiment from the SCA, a national telephone study. Using call record data, they simulated the impact of ending efforts to interview sample cases after, say, 5 or 10 callbacks. Again, there were small effects from these large (simulated) changes in the data collection protocol. These results do not indicate that nonresponse is never a problem, but they do suggest that, when the variables of interest are unrelated to the factors that produce nonresponse (as in the exit polls), falling response rates may not be a major source of worry.



Next steps. Of course, the big issue is when response rates matter and when they don't. So far, the research has only looked at descriptive statistics like means and totals, and it's not entirely clear whether the findings on nonresponse rates and nonresponse errors would apply to analytic statistics as well. But we clearly need both theoretical development and empirical research to help discover the general conditions under which nonresponse matters and when it doesn't. In addition, we need to explore how nonresponse affects estimates (for example, trend estimates) from recurring surveys. In looking at differences across waves of a repeated cross-section or panel survey, it is the *differential* nonresponse bias that matters — the difference between the nonresponse biases in each wave. This means that recurring

surveys may be more robust against nonresponse bias than cross-sectional surveys are. But it would be nice to see that confirmed empirically. Finally, there are still many unresolved questions about the effectiveness (and the cost-effectiveness) of many standard survey procedures for reducing nonresponse rates. If the data suggest that we do need to reduce nonresponse rates, we need to know a lot more about how to do that at the lowest cost.

References

Atrostic, B.K., Bates, N., Burt, G., & Silberstein, A. (2001). Nonresponse in U.S. government household surveys: Consistent measures, recent trends, and new insights. *Journal of Official Statistics*, *17*, 209-226.

Curtin, R., Presser, S., & Singer, E. (2000). The effects of response rate changes on the Index of Consumer Sentiment. *Public Opinion Quarterly*, *64*, 413-428.

deLeeuw, E., & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison." In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey Nonresponse* (pp. 41-54). New York: John Wiley and Sons.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household surveys*. New York: John Wiley.

Keeter, S., Kohut, A., Miller, C., Groves, R., & Presser, S. (2000) Consequences of reducing nonresponse in a large national telephone survey. *Public Opinion Quarterly*, *64*, 125-148.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213-236.

Merkel, D., & Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey Nonresponse* (pp. 243-258). New York: John Wiley and Sons.

Oldendick, R., & Link, M.W. (1999). Call-screening: What problems does it pose for survey researchers. Paper presented at the International Conference on Survey Nonresponse, Portland, Oregon, October 28-31.

Steeh, C., Kirgis, N., Cannon, B., & DeWitt, J. (2001). Are they really as bad as they seem? Nonresponse rates at the end of the twentieth century. *Journal of Official Statistics*, *17*, 227-247.

Tuckel, P., & O'Neill, H. (1996). Screened out. *Marketing Research*, *8*, 34-43.

Tracking Trends and Contexts:
Problems and Strengths with the NSF surveys

Reeve Vanneman
University of Maryland

I want to raise two issues about trends in academic surveys since the inception of the NSF three main social science surveys (the GSS, NES, and PSID):

- the trend away from repeated cross-sections to panel studies and
- the trend towards more contextualized data.

The lack of a sustained panel study for political science and sociology permanently handicaps these discipline's ability to address adequately the causal questions that are central to scientific inquiry. Nevertheless, the GSS and NES continue to make essential contributions to their disciplines. The long time series now available and the recent further development of contextual measures linked to the surveys deserve continued support.

The best way to maintain the advantages of the GSS and the NES in supporting time series and contextual research and to fulfill the need for panel data is to develop a fourth NSF supported survey. While the PSID has fulfilled well the research function of panel data for economics, some see a need for a new panel sample that does not tie research to the same 5000 families (and their descendants). Thus, **a new fourth panel survey** offers the opportunity to create a genuinely inter-disciplinary social science resource for an era in which inter-disciplinary research is bound to grow.

Repeated Cross-Sections: Tracking Trends over Time (& Space)

History. The NES time series begins in 1948; the PSID in 1968; the GSS in 1972. By any standard, these are old timers of survey research. Among widely used data files, they are surpassed only by the Census itself (the IPUMS begins in 1850) and the CPS (electronic data begin in 1962). (Intermittent public opinion data are available from the Roper Center since 1935, but these are not as widely used.). Rough contemporaries would be the NLS of mature and young women which began in 1967 and 1968); the NLSY which begins in 1979; and the Eurobarometers which began in 1970.

The longevity of the GSS, NES, and PSID is their great advantage. As for people, age confers benefits and costs. With time, the surveys have accumulated great stores of knowledge and broad perspectives on the field. However, the arteries harden and the muscles atrophy, so we are less nimble than we used to be. Questions naturally arise whether we can keep up with the fast changing environment and whether new ideas and cutting edge methods have passed us by. In many ways these surveys are still remarkably nimble. Despite the imperative to maintain continuity, each still nurtures experimental ideas and new topics.

Panels vs. repeated cross sections. The PSID, of course, is a panel study, but since it is close to a representative sample of the population, it is sometimes analyzed for its time series aspects much like true repeated cross-sections such as the GSS, NES, CPS, etc. Of the major social science data files that have been funded since 1970, virtually all have been panel studies:

- the AddHealth,
- most of the Family Life Surveys (IFLS, MFLS),
- the HRS (Health and Retirement Surveys),
- High School and Beyond and its NCES successors,
- the early NLS surveys, NLSY79 and NLSY97,

- the NSFH (National Survey of Families and Households),
- the SIPP (Survey of Income and Program Participation);
- and others too numerous to mention.

The trend away from repeated cross-sections to panel studies reflects the methodological advantages of panel studies (e.g., fixed effect designs). The trend certainly raises the question whether repeated cross section designs are now obsolete.

Tracking changes over time. Their longevity makes the GSS, NES, and PSID more valuable with each wave of data. Each of these surveys can trace changes over the last three to five decades on an impressive range of topics. Most often we are interested in simple changes in the *level* of some variable, for instance, political alienation, religiosity, or inter-generational wealth transfers. Sometimes we are also interested in changes in the *variance*, for instance, in income inequality or opinion polarization. Perhaps less often, but perhaps more importantly for social science, we are interested in changes in the *relationships* among variables, for instance, between earnings and race, sex or schooling, or between class position and presidential voting, or between parental and own occupation. These surveys provide the main research data base for asking these questions about national changes over time.

Maintaining the research infrastructure to answer these change questions is seen as essential by the research community. This advantage creates an irresistible conservative bias in the funding and design of new waves of these studies. But the need for continuity creates also a fundamental tension between science's usual bias towards new ideas versus the need to preserve the past in order to study change. Among the many issues raised by this tension, I would like to highlight three advantages and two problems:

Advantage: public knowledge. We all bemoan the superficiality with which survey data are reported in the media and consumed by the public. Rarely do we find even a simple cross-tabulation reported, and it is my impression that such analyses may even be declining over time. But the one comparison made more frequently than any other is the comparison over time. When the public is interested in anything other than the current level of some variable, it is usually in how those levels have changed over time. This common curiosity in how America has changed is a wedge that we should use to encourage better use of social science knowledge. At the minimum, we need to appreciate that these time series data are one of our most valuable products from the public's point of view and an important argument for continued public support of the social sciences.

A related advantage is the widespread usefulness of these surveys for *teaching* purposes. Because cross-sectional designs are simple (if less powerful), they are often the first choice of introductory methods courses. The number of undergraduates who have first become interested in social science research because of these surveys is uncountable.

Advantage: a macro perspective. We usually think of survey data as micro-data to be analyzed to evaluate micro models. The divide within the social sciences between micro and macro perspectives is not easily overcome, but these time series surveys are one tool available to bridge that gap. On the most basic level, we can analyze to what degree national changes in some outcome are a product of national changes in the micro-determinants of that outcome. It is my impression that most of those analyses are rather disappointing; compositional changes rarely account for much of the change in some outcome. These disappointments should be a challenge to our micro models. We need to understand better that macro level changes are not a simple aggregation of micro level differences. All of the social sciences have wrestled with these micro-macro discontinuities to some extent. But analysts of these repeated cross-sections face those issues more directly and thus these data are a constant goad to think through these problems better.

Cross-national comparisons. The repeated cross-sections share this macro perspective with another trend in survey research that has not made our agenda: the repetition of survey items across national boundaries. The Luxembourg Income Studies, the International Social Survey Program, and the Comparative Studies of Electoral Systems have raised our awareness of the possibilities of going beyond U.S. borders to test our models. While the comparative thrust is not limited to these efforts, the GSS, NES, and PSID have clearly inspired much of this work both directly through personal contacts and indirectly as outstanding exemplars. That is a big plus for these surveys.

Advantage: interdisciplinary communication. The 21st century will be a century of interdisciplinary co-operation. We have seen this happen already in the biological and physical sciences; it is bound to happen eventually in the social sciences. While each of these surveys is usually seen as primarily the preserve of one of the disciplines; in fact, they are already an avenue of interdisciplinary exchange. We need to encourage this further. When economists or political scientists explore a "sociological" data set, it is usually the GSS. Sociologists use the economists' PSID (and NLSY) and the political scientists' NES with great regularity. When researchers use the data, there is a need to find out how the data have been used before. This forces researchers to read outside their usual disciplinary boundaries. More reading can (sometimes) lower the barriers of language and implicit assumptions that separate the disciplines' approaches to similar topics. (Although sometimes, alas, it seems to just reinforces mutual contempt.) We don't need to abandon those long held assumptions; but if we are interested in the same data, we can learn to communicate better. When biologists, computer scientists, chemists, and physicists combine to unravel the human genome, they don't abandon their disciplinary strengths; but they do learn how to talk to one another and take advantage of each others' expertise. The social sciences could benefit from similar learning. In their governance, each of the supervisory boards already include members of other disciplines. We need to encourage this co-operation further.

Problems: new issues. Not all issues that seem important to us in the 21st century have been systematically included in these time series. Some concerns that now seem fairly central for understanding contemporary America (e.g., immigration and the internet), were non-existent or not so central when these surveys began. This constant problem of balancing continuity with receptivity to new issues is a frequent issue discussed in the oversight boards. The broader representation brought in by these boards seems to have brought a wider inclusivity of issues, although there are still complaints heard of lags between issues arising and their inclusion in these surveys.

Problems: new methods. The usual mantra of repeated cross sections is "If you want to measure change, don't change the measures." This is a powerful conservative force in all these surveys. Some have bowed to its imperatives more than others. PIs and oversight boards constantly face these issues and seem (to me at least) to try to make decisions reflecting the best interests of the wider research community. There are no simple answers here, and I do not want to second-guess these decisions. Perhaps it would be interesting to stimulate more conversations across the three surveys to see the differences in priorities.

The one issue I do want to raise is how this useful conservative bias against changing question wording can seep into other aspects of survey design where the benefits of continuity are not so obvious. I am especially concerned by what seems to me the resistance of the GSS to adopting any attempt at panel design. The NES has periodically included some panel designs over a few elections, but there is no sustained panel study in political science. Another example of conservatism that interferes with good science might be the reluctance to use bilingual interviews in an increasingly immigrant and multicultural society. Perhaps others can identify other issues where the necessary conservatism on question wording has been inappropriately applied to other areas of survey design.

Survey Frontiers: Embedding Respondents in Contexts

One of the most promising developments in survey research has been the growth in the contextual data that have been "attached" to individual respondents. These contexts may be employing firms, neighborhoods, religious congregations, or even informal social networks. This trend towards more contextualized survey data can be seen in all our disciplines. The ease and expansion of estimating multilevel (random coefficient) models has encouraged this growth in our data collection so we are likely to see even more of these multilevel data files in the future.

The trend deserves support but also some more systematic review. The advances we have made in the last decade have been dramatic but rather disconnected. There is much in common among the different contextualized data efforts, but also much that is different. Can they learn more from each other about the strengths and problems of contextual data? Where will this trend lead surveys in the future? Which contexts need to be regularly included in our surveys? What other kinds of contextual data need to be collected beyond attaching already existing administrative or census data to the surveys? What are the best solutions to protecting privacy with the more identifiable individual data when they are embedded in contexts?

Three disparate examples will illustrate the range of developments that are occurring quite independently of each other but which share the common feature of embedding survey responses in larger social contexts:

1. Under the leadership of **John Abowd , John Haltiwanger, and Julia Lane,** an NSF and Census supported project is matching individuals in many of our most important surveys to their social security (or state-level unemployment) records and from that to firm level data that employ these individuals. The most commonly cited advantage of linking data to administrative records is the accurate individual time-series that can be added to surveys. But the addition of contextual data about the employing firm is at least as important an advantage. The potential of these linked contextual data are enormous. Because the social security (or state unemployment) records include virtually everybody in a firm, we will also have a rather complete profile of earnings dispersions within the firm. The firm IDs also enable a link to other census collected data about the firm. Abowd has used the combination of longitudinal individual earnings histories and firm-level earnings records to separate individual and firm effects on earnings. But this just scratches the surface of the potential of these data once they become more widely accessible. It may not be too much of an overstatement to say that much of the future advances in our knowledge about earnings, job mobility, organizational growth and change, and a wide variety of other topics will depend on these linked data.

2. **Robert Sampson** and his colleagues have followed up William Julius Wilson's ideas about neighborhood effects to collect survey data in Chicago that take advantage of more concentrated clustering than we might usually consider prudent. This clustering enables Sampson to aggregate individual perceptions of "collective efficacy" – the degree to which residents expect neighbors to engage in various collective efforts to maintain the neighborhood – up to the neighborhood level and show that these perceptions are related to neighborhood crime rates. The explosive growth in interest in neighborhood effects thanks to Wilson, and in social capital thanks to Robert Putnam, means that such contextual data will become ever more important in future research. Sampson has shown what can be done beyond just attaching existing census data to individual surveys. But Wilson and Putnam also emphasize the importance of local *institutions* – schools, churches, neighborhood associations, local governments – in shaping people's lives. As yet, we are not systematically collecting data on these local institutions in concert with our individual survey data collection.

3. **Paul Beck, Russell Dalton, and Robert Huckfeldt** have added contextual data from individual networks, local media and political party organizations to the usual individual survey interviews in order to examine how social contexts influence voting choices. Their emphasis on contexts reinvigorates a social influence tradition of political research from the 1940s and 1950s. This earlier model had been eclipsed by a more radically individualized model of politics that was perhaps more congenial to our survey methods. Like neighborhood research, these political scientists used county-level data to measure some of these influences: they combined content analyses of local newspapers and interviews with county political party officials to embed the survey interviews in their local political contexts. Perhaps even more innovatively, they elicited network data from their survey respondents and then went and interviewed a sample of these "discussion partners" to measure the concordance in political attitudes between their respondents and their networks.

There are probably many other important examples of how surveys are becoming much more contextualized than was true when the GSS, NES, and PSID were begun. The **GSS** has twice collected organizational data from the employers of their sample respondents; and another year surveys of the sample's religious congregations were studied. While these initiatives began as a clever way to sample firms and religious congregations, together with the regular GSS population sample, they provide many opportunities for contextual research. The **AddHealth** study of adolescents, for instance, blanketed some of their sampling units so that they could collect more complete social and sexual network data in order to analyze patterns of reciprocity in friendship choices and reporting of sexual partners. The **Demographic and Health Surveys** in developing countries now routinely collect community level as well as individual level data. And *families* are themselves one of the most important contexts in which individuals are embedded so that survey designs which collect data on the entire family unit (e.g., the **PSID**) might be considered in a sense the pioneers of this contextualized movement.

The trend towards more contextualized survey data is exciting because it presents a more realistic picture of social life than the atomized individuals that used to dominate our survey designs. We used to think of clustering as a necessary but unfortunate cost in our sampling frames; now it is an advantage to be exploited. We used to develop statistical models in which individuals were supposed to be independent observations; now we are more interested in the linkages between them. This amounts to a revolution in the way we do research. But it is, I think, a largely unplanned and spontaneous series of insights that have occurred in disparate places in the social sciences. Much could be gained by thinking through the issues raised by these many separate efforts – especially in the futures we envision for our main surveys.

The growth of contextualized survey designs needs to be maintained and augmented. The repeated time series designs may be better at capturing spatial contextual effects because they can sample more densely within sample frame clusters. With geographic mobility, panel respondents inevitably disperse throughout the country. The cross-sectional surveys' advantage of local clustering argues that both types of designs need to be supported: the repeated cross sections in order to maintain the national time series and to explore contextual questions; the panel studies in order to answer causal questions.