Journal of Official Statistics
Vol. 8. No. 1, 1992, pp. 5–17

Cognitive Aspects of Surveys: Yesterday, Today, and Tomorrow

Judith M. Tanur[1] *and Stephen E. Fienberg*[2]

Abstract: Since the early 1980s, the movement to study cognitive aspects of surveys has offered a new perspective on an important class of measurement errors, long studied by survey researchers. We present a brief history of the movement. We then examine what we see as some major achievements of the movement so far. These are the establishment of the government laboratories in the United States, with a parallel new sensitivity to cognitive methods and findings; and the reconsideration of older research and conceptualizations under the cognitive rubric. Finally, we stress the importance of appropriate design for embedded experiments to study cognitive aspects of surveys. We conclude with the suggestion of continuous embedded experimentation in on-going surveys.

Key words: Cognitive laboratories; experimental design; generalizability; nonsampling errors; questionnaire design; sample survey design.

1. Introduction

Over the decade of the 1980s, the movement to develop cognitive aspects of survey methodology (CASM) began to offer a new perspective on an important class of measurement errors in surveys. Seeing the answers to survey questions as the end product of such cognitive processes as comprehension, memory, and judgment, the CASM movement has sought to bring to bear insights, theory, and methodology from the cognitive sciences (especially cognitive psychology) to assist our understanding of measurement errors and to aid in their control. Simultaneously, it has offered investigators in the cognitive sciences the survey process as a broader arena for testing their theories than is available in the usual cognitive laboratory. While cognitive psychology and survey research traditionally have had very different goals and very different working methods, recent CASM developments stress the interplay of the methodologies.

Surely the issues that are addressed under the CASM banner are not new. The impact of question meaning, wording, and ordering and the issues of recall and reporting biases have long been the subject of methodological research. Early on, workers in the survey enterprise recognize that sampling error was both controllable and measurable but that nonsampling sources of error were likely to dominate total survey error. Various sources of nonsampling error were documented by Deming (1944) and were the focus of the

[1] Department of Sociology, State University of New York at Stony Brook, Stony Brook, NY 11794-4356, U.S.A.
[2] Office of Vice President (Academic Affairs), York University, 4700 Keele St., North York, Ontario M3J 1P3, Canada.

research in the ensuing decades. Nonsampling errors are usually divided into nonresponse errors and response or measurement errors. Traditional research has emphasized their contributions to the bias component of total survey error (see Hansen, Hurwitz, Marks, and Mauldin 1951), although more recent perspectives stress both fixed and random components (Groves 1989). Among the potential sources of measurement error are the interviewer, the respondent, the survey instrument, and the mode of survey measurement, as well as their interactions (Groves 1987).

Investigation of some of these sources of nonsampling error has been part of the survey research tradition. For example, Neter and Waksberg (1964) describe the recall phenomenon of telescoping in which respondents reporting the dates of purchases place them more recently in time than when they actually took place. The evidence for this phenomenon came from studies that varied the reference period for reporting and observed differences in estimates, rather than from direct measurement of the dates on which the purchases occurred. Similar reports of the phenomenon and measurements of the bias it introduces into survey estimates can be found in Sudman and Bradburn (1973) as well as in the extensive documentation on the development of the National Crime Survey (NCS) in the 1970s. In the NCS, data on victimizations were gathered longitudinally in successive interviews. Thus the bounding provided by information from earlier interviews allowed for the investigation of both forward and backward telescoping both within and between reference periods.

The CASM movement has attempted to approach such known phenomena as telescoping, as well as other issues, from novel points of view, to introduce new methodologies, and to provide a basis for systematic treatment. CASM methodology, in our view, is not a panacea for all the ills that survey measurement is err to. Rather, this methodology provides a different set of diagnostic tools to assess the problems that cause measurement error. As such, CASM has not made the traditional approaches obsolete, but it has offered valuable new insights.

2. A Brief History of the CASM Movement

In the 1970s, the survey research community evinced renewed concern over issues both of falling response rates and of the validity of survey data being increasingly used for academic research and policy decisions. The response on both sides of the Atlantic included the convening of a series of conferences to suggest ways to increase the validity of conclusions derived from survey data. One such conference considered alternatives to survey data collection (Sinaiko and Broedling 1976). Another long term effort dealt with issues of reporting subjective phenomena (Turner and Martin 1984). Others explored issues of cognition that would affect validity of answers given to survey questions (e.g., see Biderman 1980; Moss and Goldstein 1979).

The CASM movement in the United States had its roots in the 1980 conference organized by Albert Biderman for the Bureau of Social Science Research (Biderman 1980). Funded by the Bureau of Justice Statistics, it brought together statisticians, cognitive psychologists, and survey researchers to focus on the National Crime Survey. While the many participants found the cross-disciplinary prospects exhilarating, no institutionalized structure resulted, and the only publication resulting from that conference of which we are aware was the work of Loftus and Marburger (1983) on

improving respondents' dating of events by landmarking. The conference did, however, stimulate through its participants and others a variety of activities.

In 1983, the Committee on National Statistics of the National Research Council, with funding from the National Science Foundation (NSF), organized an Advanced Research Seminar on Cognitive Aspects of Survey Methodology and thereby coined the acronym CASM with the express notion that the purpose of the Seminar was to build a bridge over deep interdisciplinary chasms. In a week-long retreat, statisticians, survey researchers, cognitive psychologists, anthropologists, and agency administrators addressed not only the problems arising in surveys and how cognitive theories and methods might be applied toward their solutions, but also how surveys might be used by those in the cognitive sciences to expand beyond the walls of their laboratories. By design the Seminar included a mix of people from academia and from government agencies and focussed on the National Health Interview Survey (NHIS). Jabine, Straf, Tanur, and Tourangeau (1984) give a report of the Seminar and outline some of the research proposals that originated there.

In our view, the most important U.S. institutional arrangements traceable to the CASM Seminar are a series of government laboratories set up to explore cognitive aspects of surveys. Largely because the NHIS was taken as the focus of the original CASM Seminar and because personnel from the National Center for Health Statistics (NCHS), especially Monroe Sirken, were enthusiastic participants in the Seminar and active advocates thereafter, the first such laboratory was established at NCHS. It too was co-funded with NSF. The Bureau of Labor Statistics and the Census Bureau have since established their own laboratories for pretesting questionnaires, investigating redesign options, and carrying out basic research (Dippo and Herrmann 1991; Martin 1991; Sirken 1991).

Another important institutional arrangement that had its roots in the CASM Seminar was the Social Science Research Council's (SSRC) Committee on Cognition and Survey Research, an academic/government agency partnership that stimulated research in the interdiscipline through its meetings and workshops. An account of the activities of the SSRC Committee appears in Tanur (1992). Other results are more indirect. For example, see the excellent discussions of CASM phenomena in Groves (1989) and the interactional analysis in Suchman and Jordan (1990) that was supported in part by the SSRC Committee. A parallel development addressing many of the same issues has been centered at ZUMA (Zentrum für Umfragen, Methoden, und Analysen). ZUMA activities include many research studies, two international conferences, and a newsletter on cognitive aspects of surveys (see, for example, Hippler, Schwarz, and Sudman 1987).

3. Some Key CASM Contributions

Our purpose here is not to attempt an exhaustive review of the substantive results of the CASM movement. A fine review has been prepared recently by Jobe and Mingay (1991) and interesting evaluations of progress have been carried out by Aborn (1989, 1991), himself closely associated with the early development of the movement. We will, however, point to what we see as three major contributions of the movement: (i) the effects of the establishment of the cognitive laboratories in the U.S. government agencies; (ii) the role played by cognitive psychologists taking theories generated in the analysis of surveys into the laboratory for testing and then back into the field; and

(iii) a new set of interdisciplinary collaborations between cognitive psychologists and survey researchers that is yielding new models for both disciplines.

First, we consider the importation of tools from the cognitive laboratory into the U.S. government survey enterprise. This has led to a new level of awareness among government survey methodologists of the cognitive difficulties occasioned by the tasks required of survey respondents. Government survey enterprises have a long history of methodological care and experimentation – but traditional methodological survey experiments were carried out primarily in the form of full-scale field tests, such as those described by DeMaio (1983). The cognitive laboratories in the government agencies now use such tools as "think aloud protocols" and cognitive interviewing with small numbers of subjects to do early pretesting and to secure insight into redesign options, sometimes even options favored by previous field testing. For illustrations, see Tucker, Miller, Vitrano, and Doddy (1989). When pretesting supplements for the NHIS, the laboratory at NCHS routinely invites subjects with the kinds of medical problems germane to the supplement into the laboratory. Such subjects provide responses to the draft questionnaire, but also report their thought processes aloud as they are answering. Such think aloud protocols give clear evidence of mismatches between survey designers and respondents on concepts and vocabulary. Of course, field tests of innovations are crucial before a change is made in an operational survey, but this new approach of going back and forth between the laboratory and the field seems to add flexibility and perhaps reduce costs.

Both Jobe and Mingay (1991) and Aborn (1989) point to ways in which the movement has benefited the cognitive sciences, and cognitive psychology in particular. They note, among other things, the unanticipated ability of cognitive psychologists to take theories generated in the analysis of surveys and survey-based experiments into the laboratory for testing, as well as their more anticipated ability to test laboratory-generated theories in the field. So far the results seem as often to disconfirm survey-based theories in the laboratory as to confirm them. For example, Bradburn, Rips, and Shevell (1987) hypothesized that telescoping is the result of clarity of memory arising from the vividness or salience of an event. Such clarity, they argue, would mislead a respondent who is using an availability heuristic (a mental shortcut that makes instances easily summoned to memory seem more frequent) to gauge the regency of an event. This hypothesis was not supported in a laboratory experiment by Thompson, Skowronski, and Lee (1988) in which misdating was not related to memorableness as rated by subjects at the time of recall. This "refutation" points to a system of mechanisms more complicated than those previously envisaged. In a series of studies that tend more to cumulation, the laboratory work of Loftus and Fathi (1985) and survey results of Loftus, Smith, Klinger, and Fiedler (1992) suggested that recall was more efficient in a backwards direction (most recent event first) than in a forward direction. Additional experiments that included the alternative of free recall, however, and the work of Jobe, White, Kelley, Mingay, Sanchez, and Loftus (1990) on the National Health Interview Survey/National Medical Expenditures Survey Linkage Field Test, found that free recall of doctor's visits was at least as accurate as recall in which the respondents were instructed as to order. Thus, there appears to be consistent evidence that free recall is at least as good as any imposed order. While these examples suggest that research has just begun to yield some

cumulation by confirmation, the CASM movement does seem to provide continuity in the research, both by furnishing a structure and by bringing together researchers who might otherwise not have talked to one another.

This "talking to each other," it seems to us, has resulted in the third major contribution of the CASM movement. As old problems that have plagued the field of survey research are explained to and explored by investigators with training in the cognitive sciences, the new perspectives such "outsiders" bring to bear inspire new models. We offer two examples.

The first deals with the issue of telescoping referred to above. New work by Huttenlocher, Hedges, and Bradburn (1990) represents an ideal use of a survey to test cognitive theories and to shed light on an issue relevant to surveys themselves. To explore how people report time, these investigators subsampled respondents to the General Social Survey (GSS) and telephoned them with a series of follow-up questions some time after the GSS interview. When asked the date on which the GSS interview had occurred, few respondents were able to answer correctly, but when asked the number of days that had elapsed since the interview a majority of respondents were able to answer. This finding supports an interpretation that temporal memories are stored as event sequences rather than as calendars. Further, when asked simply "When did the interview take place?" almost three-quarters of the respondents answered in terms of elapsed time. The proportion of reports in terms of weeks decreased and the proportion of reports in terms of months increased with actual elapsed time.

Finally, Huttenlocher et al. (1990) explored two answering stategies that resulted in downward biases in response to the question "How many days ago did the interview take place?" First, respondents seemed to impose limits on themselves, as if thinking that about two months or 60 days was the longest time they could reasonably be expected to estimate by days. This self-imposed reference period gave rise to telescoping from inexact temporal memory similar to that found in earlier work by Huttenlocher, Hedges, and Prohaska (1988) when the reference period was imposed by the experimenter. Second, respondents tended to over-use certain culturally prototypic values – 7, 10, 14, 21, 30, and 60 days. Note that the distances between these prototypic values increase with the size of the values. Thus, if respondents tend to round to these values, they will introduce a net underestimate of elapsed time, further contributing to telescoping.

The Huttenlocher et al. (1990) approach draws on insights from the psychological literature to construct a model for reporting errors that takes into account effects due to bounding as well as effects associated with rounding to various types of prototypic values. The heaping-up of reported events at boundaries of reference periods as well as at selected special values is not a new observation. It has been the focus of much previous research. The new perspective that Huttenlocher et al. (1990) bring offers an integrated interpretation of the previously observed phenomena and a framework in which further discussions about them can take place. The emphasis in this work is not the traditional survey approach of measuring bias and error – these are already known to exist and be substantial; rather, it is on identifying fundamental mechanisms with the aim of eventually altering the survey instrument and administration to eliminate or at least reduce the distortion due to that source of error.

A second example concerns the long-standing issue of the correspondence between attitudes and behavior (see, for example, Deutscher (1973) and the discussion in Schuman and Presser (1981)). Recent work by Russell Fazio and his various colleagues sheds some new light on this old problem. They argue that the more likely an individual's attitude is to be activated from memory when he/she encounters an attitude object, the more likely he/she is to act in accordance with that attitude, at least when social desirability is not involved. Further, they argue that one can measure the accessibility of an attitude by the latency of its self-report: the more rapid the response the more accessible the attitude.

Several studies seem to support this chain of reasoning. Fazio and Williams (1986) found that self-reported attitudes towards Ronald Reagan measured in the summer of 1984 correlated better with ratings of Reagan's performance in a televised debate in the fall among subjects with short response latencies than among those with longer latencies. These differences in strength of correlation held up when the predicted behavior was the vote in the November election. In a study by Houston and Fazio (1989), the latency of response to an attitude question regarding capital punishment similarly differentiated groups with high and low correlations between such attitudes and judgments of the quality of research ostensibly supporting (or not supporting) the efficacy of capital punishment. Fazio, Powell, and Williams (1989) linked latency of reported evaluations of such mundane objects as raisins, gum, and candy to the predictability of subjects' choosing these objects as a reward for participating in the experiment. Dovidio and Fazio (1992) suggest that response latency can be measured in survey interviews using computer assisted personal interviewing (CAPI) capabilities, as well as in telephone surveys using computer assisted telephone interviewing (CATI) technology or touchtone data entry (TDE) to measure accessibility of attitudes in operational surveys.

The research of Fazio and his colleagues draws upon previous psychological work on the distinction between spontaneous and deliberate behaviors and uses these ideas to shape our understanding of the link between the expression of attitudes and subsequent behavior. While there is a substantial survey research literature that looks at the link between attitudes and behavior, sometimes mediated by social desirability, this previous literature did not give an integrated perspective rooted in the accessibility of memory.

4. Embedded Experiments: Some Concepts and Implementations

Random sampling and randomized experimentation have common bases in statistical theory (Fienberg and Tanur 1987). Many examples of their combination, as formal experiments embedded in sample designs, can be found in the survey research literature over the past 50 years. Such embedded experiments are most typically used to compare alternative aspects of survey methodology (e.g., questionnaire content, training methods, or collection techniques). They use either pilot surveys or methods test panels. In this section, we summarize some key concepts associated with the domain of embedded experiments and stress their use in the context of on-going surveys.

Perhaps the most commonly-used design for an embedded experiment is that of the split-ballot or split-sample experiment, which randomly administers alternate questionnaires or other variations in procedures to subsets of the sample in a survey.

For example, Abelson, Loftus, and Greenwald (1992) describe a series of split-ballot studies using the 1987 and 1989 National Election Survey Pilot Study in which two alternative forms of vote self-report questions were used.

The formal theory of statistical experimentation stresses the role of experimental control in sharpening the contrasts among estimated treatment effects. Such devices as "blocking" systematically extract large components from the estimated error variance associated with a basic design, thereby improving the precision of the comparisons in question. The sample design of most national surveys involves forms of clustering and/or multiple interviewers. The experimental strategy of improving precision through the use of sample clusters as blocks for a split-ballot experiment takes advantage of the fact that respondents in the same cluster are expected to be more similar to one another than to respondents from other clusters to increase the precision of comparisons between alternative survey procedures. Because a cluster is usually assigned intact to a single interviewer, blocking on clusters also involve a form of blocking on interviewers. Such blocking would be especially useful since many studies have found inter-interviewer variability to be a major source of measurement error. Few embedded experiments use this sort of blocking, replicating the set of experimental treatments randomly within each cluster and hence within interviewers, and the argument usually advanced is that using different forms of an interview renders an interviewer's task too difficult. Designers of such embedded experiments are also sometimes concerned that unsupervised field interviewers will, if they are aware of the existence of alternative survey procedures, use the most convenient one regardless of instructions. Both these objections loose sight of the early history of embedded experiments (e.g., at the U.S. Bureau of the Census) in which such blocking was common and successful, often to the surprise of the critics who had said it could not be done (see the discussion in Fienberg and Tanur 1988). See also the recent work by Rothgeb, Polivka, Creighton, and Cohany (1991) that reports on research in which every interviewer administered three different versions of the Current Population Survey (CPS).

All too often, even when different treatments are assigned to respondents within clusters, the traditional mode for reporting survey results "averages over" these features rather than controlling for them in an analysis-of-variance-like fashion. For example, the Abelson et al. (1992) study referred to above reports an analysis that treats the two half-samples used for the different question forms as if they were independent.

We can carry the concept of embedded experiments one layer deeper than that used in a simple split-ballot experiment. Factorial designs can also be embedded in surveys. An excellent example is provided by a set of experiments (Branch, Jobe, and Kovar 1989) proposed by researchers at Boston University and the National Center for Health Statistics (NCHS) to be embedded in the fifth wave of the Massachusetts Health Care Panel Survey, a longitudinal survey that began with 1629 respondents aged 60 or over in 1974–75. There are two sub-experiments planned; one on the accuracy of reports of activities of daily living and the other on recall of previous states of health, of occupation, or of places of residence. Both use as a factor two versions of the questionnaire. One version will be the standard one, while the other will be developed in the NCHS cognitive laboratory and will, for the recall task, include the use of a personal time

line and additional retrieval cues, techniques developed in the laboratory to improve recall. In the experiment to investigate the quality of data on activities of daily living there are two other factors: mode of interview (in person versus telephone) and respondent type, with data for each subject living in a multi-person household being provided both by the subject him/herself and by a proxy respondent. The recall experiment will use responses provided at earlier waves of the survey as validation data, so that the interval between waves will be a within-subjects factor. In addition to the traditional versus cognitively-designed questionnaire factor, this experiment will also use as a factor living arrangements (living alone versus living with others) in order to check on the generalizability of the results. Subjects living in multi-person households who are interviewed in person in the first experiment will also participate in the second experiment, participation order to be counterbalanced. Tourangeau and Rasinski (1988) report yet another example of an embedded factorial experiment in their study of the effects of the context of the questionnaire on attitudes reported in a pilot survey (also see the reanalysis in Fienberg and Tanur 1989).

We need not restrict ourselves to factorial designs in randomized blocks to make effective use of embedded experiments. The probability structure underlying a split-plot experimental design is identical to that underlying a cluster sample (see Fienberg and Tanur 1987). In the split-plot experiment one can apply treatments at either the whole-plot or the sub-plot level, each of which has its own error term (or randomization structure). Similarly, there are two sources of variability in a cluster sample, between clusters and within clusters. By linking the experimental and sampling structures we can consider two different sets of treatments, one applied to subsets of clusters and the other within clusters as suggested above. Such a design is useful when the first set of treatments (say, alternative collection techniques) applied at the whole-plot level (between clusters) requires less precision than the treatments (say alternative questionnaires) applied at the sub-plot level (within clusters).

Randomized statistical experiments are designed to ensure internal validity, that is, to demonstrate a cause-and-effect relation between treatments and outcome within the experiment itself. Sample surveys, on the other hand, are usually designed for external validity, that is, to ensure generalizability of the results from the sample to the population from which it was drawn. The intertwining of the two paradigms in embedded experiments leads us to consider three possible perspectives on the inferences drawn from an embedded experiment. We couch our exposition in terms of a two-treatment experiment, but of course the notions generalize to more than two treatments.

If one uses the standard experimental paradigm, relying on internal validity and the assumption that the unique effects of experimental units and treatment effects can be expressed in simple additive form, without interaction, then inference focuses on within-experiment treatment differences. This is the simplest form of inference and its simplicity perhaps helps explain why experimenters have been slow to adopt designs more complicated than the traditional split ballot, or to tailor analyses to fit more complicated design features introduced to provide local control. Alternatively, one can use the standard sampling paradigm, relying on external validity and generalizing the observations for each of the treatments to separate but paired populations of values. Each unit or individual in the original

population from which the sample was drawn is conceived to have a pair of values, one for each treatment. But only one of these is observable, depending on which treatment is given. Inference focuses on the mean difference or the difference in the means of the two populations, depending on the presence or absence of blocking. Finally, one can conceptualize a population of embedded experiments, of which the present embedded experiment is a unit or sample of units, thus capitalizing on both the internal validity of the current experiment and the external validity implicit in the generalization from the current experiment to the superpopulation.

It is our contention that inferences from an embedded experiment ought to be done on the basis of a model that incorporates appropriate features of the sampling design and their ramifications throughout the embedded experiment. Although the features of the sample design are often used to achieve local control in the experimental design, these features are rarely introduced in the analysis stage. Unfortunately, while the tradition of split-ballot experiments is widespread in the survey research community, the formal incorporation of interlocking features of their design is rarely part of reported analyses (see the discussion in Fienberg and Tanur 1988). For example, the OMB Statistical Policy Working Paper on Approaches to Developing Questionnaires (DeMaio 1983) has a major section on split-sample testing which is totally silent on this issue and which reports two examples of analysis in which design features are ignored.

For example, Butcher and Eldridge (1990) drew samples from the same 40 postal sectors to test whether a seven-day or a one-day travel diary provided higher quality data. While they made an attempt to balance the skill levels of the interviewers across the treatments, they did not introduce the design feature of having the same interviewers carry out both sorts of interviews. Further, despite the fact that the postal sectors were used as blocks, Butcher and Eldridge used an analysis that treated the two samples as independent and found relatively small differences favoring the one-day diary. Had they analyzed the data taking advantage of the blocking introduced at the design stage, the increased precision of the estimates might well have allowed for stronger statistical conclusions. Then the choice between the one-day and seven-day diaries could have been made on the basis of the importance of the substantive differences instead of on the seeming lack of statistically significant differences.

5. Looking Ahead

Most of the examples we have described in the previous section are of full-scale field tests of alternative survey tools emanating from cognitively-inspired studies, in which the entire sample is used for the embedded experiment. In a sense, carrying out such experiments is often viewed as a luxury for those involved in the day-to-day work of survey taking or alternatively as the final test of a long range methodological research program. But cognitively-based research, like that in most other areas, tends to produce improvements in small increments and, if we are to take full advantage of the cumulation of such increments, survey organizations need to plan for regular experimentation in the context of on-going surveys. What we are recommending is the "reservation" of a subsample in on-going surveys to be used for embedded experiments. Using 10% of the households in the CPS at least once or twice a year for carefully designed experiments would occasion little degradation of the accuracy of the CPS

(with a total of over 60,000 households reporting per month) but it would provide an ample sample size for well-controlled experiments linked to proposals for methodological improvements coming out of the cognitive laboratories of the Bureau of the Census and the Bureau of Labor Statistics.

Most survey organizations today do carry out field tests on partial samples when introducing new survey tools, especially for on-going surveys, whether of a repeated cross-section or a longitudinal variety. In the 1960s and 1970s, for example, the Bureau of the Census and the Bureau of Labor Statistics (BLS) had a separate methods test sample that they used as an ongoing vehicle for studying possible changes in the Current Population Survey. We offer two recent examples:

i. BLS and NORC carried out a comparison of Computer Assisted Personal Interviewing (CAPI) with traditional pencil-and-paper interviewing for one-half of round 12 of the National Longitudinal Study of Youth (Bradburn et al. 1991). In this study interviewers were the units of interest and half of them in round 12 were randomly assigned to experimental (CAPI) or control (traditional) groups, with stratification on the basis of race/ethnicity, urban status, number of cases assigned, etc.

ii. BLS has for close to 20 years been investigating the effect of diary format on the quality of data collected in the Consumer Expenditure Survey (CES). Both field and cognitive laboratory research have indicated that an important dimension affecting quality and respondent burden is the specificity of the categories used to describe items purchased. During 1991 and 1992, 20% of the sample for the CES will receive a nonspecific diary while the remaining 80% will receive the pre-existing production diary. BLS anticipates deciding in 1993 which of these forms is more efficacious and subsequently using only that form (BLS 1990).

But the notion of setting aside a parallel sample or even a subsample (typically much smaller than one-half or even 20% of the basic sample) is not currently a regular practice. If it were, the major efforts required to mount *de novo* the embedded experiments documented in these examples would be avoided. Setting aside a subsample for embedded experimentation is admittedly expensive, not only in the cost of the sample units expended on the experiment but also in terms of the cost of the control and management (see e.g., the discussion of the difficulties surmounted by Bradburn et al. 1991). Nonetheless, we believe that it is in the long-run cheaper than the price of ignorance that results from not having done a controlled experiment at all.

The watchwords of the industrial quality movement in the U.S. and Japan have been "continuous improvement" through process control and careful experimentation. We believe that those involved in the survey domain should have similar goals. The development of a mechanism for carrying out embedded experiments involving proposed survey improvements on a regular basis would do much to move us in the right direction. The cognitive laboratory provides a key link in this improvement process by offering a setting and access to theories for identifying alternatives for field testing.

6. Bibliography

Abelson, R.P., Loftus, E.F., and Greenwald, A.G. (1992). Attempts to Improve the Accuracy of Self-Reports of Voting. In Tanur, J.M. (ed.) Questions about Questions: Inquiries into the Cognitive Bases of Surveys. New York: Russell

Sage Foundation, 138–153.

Aborn, M. (1989), Is CASM Bridging the Chasm? Evaluation of an Experiment in Cross-Disciplinary Survey Research. Paper presented at the American Statistical Association 1990 Winter Conference, San Diego, CA, January 4–6, 1990.

Aborn, M. (1991). Discussion to Session on Cognitive Laboratories. Seminar on Quality of Federal Data. Statistical Working Paper 20, Federal Committee on Statistical Methodology, Vol. 2, 281–287, Washington DC: Office of Management and Budget.

Biderman, A. (1980). Report of a Workshop on Applying Cognitive Psychology to Recall Problems of the National Crime Survey. Washington, DC: Bureau of Social Science Research.

Bradburn, N.M., Frankel, M., Hunt, E., Ingels, J., Schoua-Glusberg, A., Wojcik, M., and Pergamit, M. (1991). A Comparison of Computer-Assisted Personal Interviews with Personal Interviews in the National Longitudinal Survey of Labor Market Behavior – Youth Cohort. Proceedings of U.S. Census Bureau's Seventh Annual Research Conference, 389–397.

Bradburn, N.M., Rips, L.J., and Shevell, S.K. (1987). Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. Science, 236, 157–161.

Branch, L.G., Jobe, J.G., and Kovar, M.G. (1989). Cognitive Aspects of Surveying the Oldest Old: Wave Z. National Institutes of Health Research Proposal.

Bureau of Labor Statistics (BLS) (1990). Consumer Expenditure Survey. Office of Management and Budget Clearance Package, 11–12.

Butcher, R. and Eldridge, J. (1990). The Use of Diaries in Data Collection. The Statistician, 39, 25–41.

DeMaio, T.J. (ed.) (1983). Approaches to Developing Questionnaires. Statistical Policy Working Paper 10, Federal Committee on Statistical Methodology, Washington, DC: Office of Management and Budget.

Deming, W.E. (1944). On Errors in Surveys. American Sociological Review, 9, 359–369.

Deutscher, I. (1973). What We Say/What We Do. Glenview, IL: Scott, Foresman.

Dippo, C. S. and Herrmann, D. (1991). The Bureau of Labor Statistics' Collection Procedures Research Laboratory: Accomplishments and Future Directions. Seminar on Quality of Federal Data. Statistical Working Paper 20, Federal Committee on Statistical Methodology, Vol. 2, 253–267, Washington, DC: Office of Management and Budget.

Dovidio, J.F. and Fazio, R.H. (1992). New Technologies for the Direct and Indirect Assessment of Attitudes. In Tanur, J.M. (ed.) Questions about Questions: Inquiries into the Cognitive Bases of Surveys, New York: Russell Sage Foundation, 204–237.

Fazio, R.H., Powell, M.C., and Williams, C.J. (1989). The Role of Attitude Accessibility in the Attitude-to-Behavior Process. Journal of Consumer Research, 16, 280–288.

Fazio, R.H. and Williams, C.J. (1986). Attitude Accessibility as a Moderator of the Attitude-Perception and Attitude-Behavior Relations: An Investigation of the 1984 Presidential Election. Journal of Personality and Social Psychology, 51, 505–514.

Fienberg, S.E. and Tanur, J.M. (1987). Experimental and Sampling Structures: Parallels Diverging and Meeting. International Statistical Review, 55, 75–96.

Fienberg, S.E. and Tanur, J.M. (1988). From the Inside out and the Outside in: Combining Experimental and Sampling

Structures. Canadian Journal of Statistics, 19, 135–151.

Fienberg, S.E. and Tanur, J.M. (1989). Combining Cognitive and Statistical Approaches to Survey Design. Science, 243, 1017–1022.

Groves, R.M. (1987). Survey Data Quality. Public Opinion Quarterly, 51 (Supplement), s156–s172.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: Wiley.

Hansen, M.H., Hurwitz, W.N., Marks, E.S., and Mauldin, W.P. (1951). Response Errors in Surveys. Journal of the American Statistical Association, 46, 147–190.

Hippler, H.J., Schwarz, N., and Sudman, S. (eds.) (1987). Social Information Processing and Survey Methodology. New York: Springer-Verlag.

Houston, D.A. and Fazio, R.H. (1989). Biased Processing as a Function of Attitude Accessibility: Making Objective Judgments Subjectively. Social Cognition, 7, 51–66.

Huttenlocher, J., Hedges, L.V., and Bradburn, N.M. (1990). Reports of Elapsed Time: Bounding and Rounding Processes in Estimation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 196–213.

Huttenlocher, J., Hedges, L.V., and Prohaska, V. (1988). Hierarchical Organization in Ordered Domains: Estimating the Dates of Events. Psychological Review, 95, 471–484.

Jabine, T., Straf, M., Tanur, J.M., and Tourangeau, R. (eds.) (1984). Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. Washington, DC: National Academy Press.

Jobe, J.B. and Mingay, D.J. (1991). Cognition and Survey Measurement: History and Overview. Applied Cognitive Psychology, 5, 175–193.

Jobe, J.B., White, A.A., Kelley, C.L., Mingay, D.J., Sanchez, M.J., and Loftus, E.F. (1990). Recall Strategies and Memory for Health Care Visits. Milbank Memorial Fund Quarterly/Health and Society, 68, 171–189.

Loftus, E.F. and Fathi, D. (1985). Retrieving Multiple Autobiographical Memories. Social Cognition, 3, 280–295.

Loftus, E.F. and Marburger, W. (1983). Since the Eruption of Mt. St. Helens. Has Anyone Beaten You Up? Improving the Accuracy of Retrospective Reports with Landmark Events. Memory and Cognition, 11, 114–120.

Loftus, E.F., Smith, K.D., Klinger, M.R., and Fiedler, J. (1992). Memory and Mismemory for Health Events. In Tanur, J.M. (ed.). Questions about Questions: Inquiries into the Cognitive Bases of Surveys. New York: Russell Sage Foundation, 102–137.

Martin, E. (1991). Discussion to Session on Cognitive Laboratories. Seminar on Quality of Federal Data. Statistical Working Paper 20, Federal Committee on Statistical Methodology, Vol. 2, 278–280, Washington, DC: Office of Management and Budget.

Moss, L. and Goldstein, H. (eds.) (1979). The Recall Method in Social Surveys. London: NFER Publishing Co.

Neter, J. and Waksberg, J. (1964). A Study of Response Errors in Expenditure Data from Household Interviews. Journal of the American Statistical Association, 59, 18–55.

Rothgeb, J.M., Polivka, A.E., Creighton, K.P., and Cohany, S.R. (1991). Development of the Proposed Revised Current Population Survey Questionnaire. Paper presented at the Joint Statistical Meetings of the American Statistical Association, Atlanta, GA, August 1991.

Schuman, H. and Presser, S. (1981). Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context. New York: Academic Press.

Sinaiko, H.W. and Broedling, L.A. (eds.) (1976). Perspectives on Attitude Assessment Surveys and Their Alternatives. Champaign, IL: Pendleton.

Sirken, M.G. (1991). The Role of a Cognitive Laboratory in a Statistical Agency. Seminar on Quality of Federal Data. Statistical Working Paper 20, Federal Committee on Statistical Methodology, Vol. 2, 268–277, Washington, DC: Office of Management and Budget.

Suchman, L. and Jordan, B. (1990). Interactional Troubles in Face-to-Face Survey Interviews (with discussion). Journal of the American Statistical Association, 85, 232–253.

Sudman, S. and Bradburn, N.M. (1973). Effects of Time and Memory Factors on Responses in Surveys. Journal of the American Statistical Association, 63, 805–815.

Tanur, J.M. (ed.) (1992). Questions about Questions: Inquiries into the Cognitive Bases of Surveys. New York: Russell Sage Foundation.

Thompson, C.P., Skowronski, J.J., and Lee, D.J. (1988). Telescoping in Dating Naturally Occurring Events. Memory and Cognition, 16, 461–468.

Tourangeau, R. and Rasinski, K. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. Psychological Bulletin, 103, 299–314.

Tucker, C., Miller, L., Vitrano, F., and Doddy, J. (1989). Cognitive Issues and Research on the Consumer Expenditure Diary Survey. Paper presented at the annual conference of the American Association for Public Opinion Research, St. Petersburg, FL.

Turner, C. and Martin, E.A. (eds.) (1984). Surveying Subjective Phenomena (2 volumes). New York: Russell Sage Foundation.

Received November 1990
Revised October 1991