Articles

Combining Cognitive and Statistical Approaches to Survey Design

STEPHEN E. FIENBERG AND JUDITH M. TANUR

Sample surveys provide data for academic research, government policy-making, the media, and business. Statistical research aims to improve survey data by reducing extraneous sources of variability and thus increasing accuracy. Researchers have begun to use paradigms adapted from the cognitive sciences to study those sources of variability associated with the processes that the respondent undertakes in understanding questions, remembering, judging and estimating, and formulating answers. To generalize laboratory-based findings, researchers must begin to embed designed experiments that vary the questionnaire content into sample surveys of broad populations. Issues associated with the design of and statistical inference from such embedded experiments are examined and illustrated with an example on the effects of context questions on responses in attitude surveys.

SAMPLE SURVEYS IN WHICH A RANDOMLY SELECTED PART OF a population answers questions have become an extremely important data-gathering device in the last half century (*1*). Social scientists, government agencies, the media, and other commercial interests often use such surveys to estimate characteristics of the population from which the sample was drawn, for example, the percentage in favor of a particular candidate, the unemployment rate, or the relation between age and criminal victimization. If samples are drawn with the help of appropriate probability methods, these estimates are usually very accurate, but survey statisticians and social scientists continue to labor to improve accuracy.

Inaccuracies in sample surveys arise from two broad sources, sampling error and nonsampling error. Sampling error is attributable to the fact that the characteristics of any randomly chosen sample will differ, by chance alone, from the characteristics of any other randomly chosen sample and thus from the characteristics of the population as a whole. The properties of random sampling error can be studied with the use of statistical theory, and this form of error decreases as the sample size increases (*2*). Nonsampling errors take two forms. The first encompasses all the things that can go wrong when human beings act and interact, such as misunderstood questions, failures of memory, and misstated answers. The second form is more specific to surveys—nonresponse and refusals, and coding errors (*3*). Nonsampling errors cannot be estimated directly by statistical theory and do not necessarily decrease as sample size increases. They often have both random and fixed components.

S. E. Fienberg is dean of the College of Humanities and Social Sciences and a professor in the Department of Statistics and the Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213. J. M. Tanur is a professor in the Department of Sociology, State University of New York, Stony Brook, NY 11794.

Survey researchers have evolved a highly developed art of questionnaire design and interview procedures to reduce nonsampling errors and have carried out many studies to test aspects of that art (*4*). But until recent years research on understanding the survey interview situation has been relatively unsystematic: an experiment to answer a specific question here or a series of studies to answer a set of interrelated questions there. Recently, however, survey researchers have recognized that among nonsampling errors are those occasioned by the cognitive processes that respondents are required to exercise in the survey interview situation. Respondents must often recall events and make judgments or estimates, and they always face issues of comprehension of the questions asked—their meaning to respondents as well as their meaning to interviewers. Survey researchers are now beginning to draw directly on the concepts of cognitive psychology and the expertise of cognitive psychologists to investigate more systematically these issues of nonsampling error (*5, 6*).

This new movement to study cognitive aspects of surveys has already generated a good deal of research (for example, *5–7*), including the experiment we discuss on the effects of the questionnaire context on responses to attitude questions. To test generalizations of laboratory results from cognitive psychology in actual questionnaires used in survey practice, researchers need to design statistical experiments that are embedded in surveys. The research to accomplish this aim embodies opportunities and pitfalls that will be the subject of the remainder of this article.

Importance and Problems of Embedded Experiments

The most common embedded experimental design is the split-ballot experiment, perhaps better called a split-sample experiment, which randomly administers alternate questionnaires or other variations in procedure to subsets of the sample. If the subsets are independent and structured with the same sample design features, we can compare the distributions of answers or the estimated relations between variables to explore the effects of varying the questionnaire or other procedure (*8*). But such designs are not the most powerful available; in particular, they do not take advantage of the fact that an interviewer carries out multiple interviews (which may be more similar to each other than to those carried out by another interviewer) nor do they take advantage of the sample design that often involves clustering, that is, interviewing multiple respondents from each randomly chosen geographic area. Respondents from the same cluster are expected to be more similar to one another than to respondents from other geographic areas. Split-ballot designs "average over" these features rather than control for them or for other potentially interesting explanatory features.

Parallels between randomized experiments and sample surveys can

be exploited in the design of embedded experiments to achieve greater precision for comparisons of interest (*9*). For example, in a randomized block experiment (*10*) relatively homogeneous experimental units are grouped together into larger collections known as "blocks." All treatments are randomly assigned within blocks. This device of "blocking" decreases heterogeneity among experimental units and thereby increases experimental precision. In the context of experiments embedded in surveys, interviewers can be used as blocks to measure and control for variability among interviewers. The classic formulation of this approach is Mahalanobis's interpenetrating networks of samples (IPNS) (*8, 11*). For example, in a survey of the economic conditions of factory workers in an industrial area of India, Mahalanobis divided the area into subareas and arranged for the selection of five independent random samples within each subarea. Each of five interviewers worked in all subareas. This IPNS design thus provided five independent estimates of the economic conditions and, as a consequence, allowed for an evaluation of the response variation associated with interviewers.

Another strategy to improve precision is to use clusters as blocks, giving members of a cluster different variants of the questionnaires or other procedures and thus taking advantage of within-cluster similarity to highlight differences between the procedures. Because a cluster is usually assigned intact to an interviewer (in order to minimize travel and other costs), blocking on clusters usually involves some degree of blocking on interviewers as well (*8, 9*).

There are operational problems created by the implementation of embedded designs that take advantage of the sample structure. For example, the added complexity of blocking on interviewers, that is, grouping into a block all units assigned to a given interviewer, runs counter to the philosophy of simplification typical of everyday survey practice. Indeed, some have argued that using alternative questionnaires for different respondents is too complicated for interviewers and consequently generates other kinds of nonsampling errors. Nevertheless, blocking on interviewers has been carried out; a recent example is the experiment by Tourangeau and Rasinski analyzed below. An earlier example comes from the methodological study of Durbin and Stuart (*12*), who designed a $3^3 \times 4 \times 2$ factorial experiment with a completely cross-classified structure involving three survey organizations, three types of questionnaires, three interview areas in London, four ages of respondents, and both sexes. Further, within one of the survey organizations they completely cross-classified age of interviewer and sex of interviewer. Each interviewer, although confined to only one district, handled all three questionnaires in approximately equal numbers with an approximate balance of age groups and sex groups of respondents. The finding of this study was that, although most of the independent variables had little effect on response rate, inexperienced student interviewers had statistically significant lower response rates than experienced interviewers. Commenting on the purported difficulty of carrying out such investigations, Durbin and Stuart remarked (*12*, p. 184):

> Although highly elaborate designs are often used in other sciences, it is not unnatural that in a field in which the experimental material is composed of human beings, the tendency should have been towards simplicity of layout. In our own experience, however, the extra amount of organization necessitated by the design we used proved to be a good deal less troublesome than had been expected.

It is crucial that careful preparation of forms and instructions to interviewers precede any attempt to have interviewers vary their behavior according to an experimental design. One way to accomplish this is through the use of Computer Assisted Telephone Interviewing (CATI), in which questionnaire variations can be programmed to appear on the interviewer's computer screen as he or she talks to the respondent. Thus the advent of telephone interviewing not only makes the use of clustering less crucial (interviewers can much more easily dial across the country than they can travel to do in-person interviewing), thus avoiding the confounding of interviewers with clusters, but also facilitates blocking on interviewers.

Another sort of problem that may arise, however, can only be addressed by withholding information from interviewers. Rosenthal and his colleagues have found that, without any intention of influencing the outcome, those conducting experiments tend to get results congruent with their expectations (*13*). Analogously, an interviewer who administers several differing forms of a questionnaire to respondents may well expect, and thus get, differing responses. The harmful effects of such expectations can be mitigated if the interviewers are kept uninformed of the direction the researcher theorizes the differences to take.

Inferences from Embedded Experiments

The embedding of statistically designed experiments within sample surveys raises issues of inference that have been rarely discussed in published sources. Despite the formal parallels in structure, there is a fundamental inferential distinction between experimental and survey contexts. Randomized statistical experiments are designed to ensure internal validity, that is, the defensibility of the cause-effect relation between the treatment and the outcome within the experiment itself. On the other hand, sample surveys use probability sampling to ensure that results will have external validity, that is, results that can be generalized from the specific sample to the population from which it was drawn. We have been able to discern at least three possible perspectives for statistical inference in embedded experiments (*8*).

1) One can use the standard experiment paradigm, which relies largely on internal validity based on randomization and local control (for example, the device of blocking) and on the assumption that the unique effects of experimental units and the treatment effects can be expressed in a simple additive form, without interaction (*10*). Then inferences focus on within-experiment treatment differences.

2) One can use the standard sampling paradigm, which, for a two-treatment experiment embedded in a survey, relies largely on external validity and generalizes 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. Then inferences focus on the mean difference or the difference in the means of the two populations.

3) One can conceptualize a population of experiments, of which the present embedded experiment is a unit or a sample of units, and thus capitalize on the internal validity created by the design of the present embedded experiment as well as the external validity created by the generalization from the present experiment to the conceptual population of experiments. Then inferences focus on treatment differences in a broader context than simply the present embedded experiment.

Because these three approaches focus on the same experimentally observed quantities but deal with possible inferences differently, they can potentially lead to different conclusions.

Consider, for example, an experiment to compare four different versions of a question on household income, with clusters that are part of a multistage area probability sampling design where each interviewer is assigned a cluster of four households to survey. Within a cluster, the four versions of the question are randomly assigned to households. The key response variable of interest is "reported household income" in dollars, typically transformed to a

logarithmic scale. We have a randomized block design embedded in the clusters of a complex sample survey design.

In the first inference approach, we use a randomization analysis for the randomized block design (*10*) or an analysis of variance (ANOVA) model with fixed effects for both interviewers and questions, and a normally distributed error term. This analysis holds the survey design as fixed and focuses internally within clusters or interviewers on the differences in effects for the questions, thereby adjusting for the differential effects of interviewers.

In the second inference approach, we divide the data into four subsets corresponding to the four versions of the question. We would then treat each subset as a sample from a population, where the sampling design is the same as that for the entire survey, but without the final stage of clustering. In each we would estimate the average household income of the population and the corresponding standard error. Finally, we would compare the estimated population averages (although to do so properly we would need some estimate of the correlations among the four estimates induced by the within-cluster intraclass correlation). This is the proper analysis for a standard split-ballot experiment, but more typically survey researchers ignore the correlation among the estimates. The inference here is external to the experiment and relies on the probability mechanism used to generate the sample. There is no natural way here to adjust for interviewer effects while still retaining an inference mechanism tied solely to the sample-selection probability mechanism. To deal with interviewers and their effects here, we need to consider them to be randomly selected from a fixed population of interviewers. This would then lead to something equivalent to the third approach.

For the third inference approach, we have a sample of size one from a superpopulation of embedded randomized block experiments. One way to handle the inference problem is to treat the interviewers as a sample from a population of interviewers; this leads to a mixed-effects ANOVA model with interviewer effects treated as a random component and question effects treated as fixed components (*14*). The formal analysis of the model here is related to, but different from, the one used in the first approach.

What is going on in the mixed-effects ANOVA model is a generalization of the treatment effect differences to the superpopulation of experiments from the present embedded experiment. The way we achieve this generalization is through representation of the interviewers as having been drawn from a superpopulation of interviewers corresponding to the conceptualized superpopulation of experiments. Thus the distinction between the first and third approaches is not simply one involving the difference between fixed and random effects in an ANOVA model but more importantly involves the level of applicability of the treatment effects.

What differences might we expect among the inferences associated with the use of the three approaches in an actual experiment? If there really are differences among interviewers, then the second approach may differ appreciably from the other two and thus would be wrong. The third approach differs from the first approach primarily through the inclusion of an extra component of variation associated with the estimated treatment effects corresponding to the "interviewer-treatment" interaction (*15*). Thus in the third approach an estimated difference in treatment effects will appear to be less precise than in the first approach. This is as it should be, because we need to pay an extra amount for the ability to generalize beyond the embedded experiment at hand. As a consequence, the mixed-effects model approach should yield "statistically significant" differences less frequently than the fixed-effects approach. The choice between the first and third approaches must depend on the intended applicability of the results.

Table 1. Estimated logit effects for comparison of neutral context with treatment combinations for abortion and welfare responses. (In logit models as fitted in GLIM, all ANOVA-like parameters are identified by setting the first level equal to zero. Then, the "constant" parameter corresponds to the log-odds of the expected counts for the neutral context and for interviewer 1. Furthermore, each of the four treatment effects represents the contrast of those effects with the neutral context, and the interviewer effects represent comparisons of each other interviewer with interviewer 1. This form of parametrization is often referred to as effects coding.)

Parameter	Abortion	Welfare
Constant	−.24 (.31)*	.09 (.31)
Rights/fraud, blocked	.21 (.33)	−.44 (.33)
Rights/fraud, scattered	.67 (.33)	−.29 (.33)
Values/government, blocked	.12 (.33)	−.96 (.35)
Values/government, scattered	Aliased†	Aliased
Interviewer 2	.28 (.33)	1.02 (.34)
Interviewer 3	−.44 (.32)	−.23 (.33)
Interviewer 4	−.14 (.34)	.01 (.34)
Log-likelihood ratio statistic‡	3.6	2.6
Degrees of freedom	9	9

*Estimated standard errors for the corresponding estimated coefficients are given in parentheses. †Because no cases fell into this marginal category, this effect is not estimable. GLIM refers to this as extrinsic aliasing. ‡This is the log-likelihood ratio goodness-of-fit statistic for the model applied to the cross-classification of the binary response variable by treatment by interviewer. Because 4 treatment by interviewer combinations had nonzero observations, we have lost 4 of the expected degrees of freedom, yielding 9 degrees of freedom.

Context Effects in Attitude Surveys: An Example

Among the cues respondents use to interpret what a question really "means" and to decide what sort of answer they are expected to give is the preceding set of questions. Cognitive psychologists explain that this context establishes a "schema," a mental framework that the respondent uses to organize or understand the meaning of a particular question. This happens with questions of fact and also with those that ask for subjective evaluations. For example, questions about health care activities during a 6-month period followed by similar questions covering a 2-month period seem to encourage respondents to segregate the periods (*16*). This result is reminiscent of the way that questioning about happiness in a specific life domain (such as marriage) preceding questioning about general happiness seems to encourage respondents to segregate marital happiness from happiness in other life domains (*17*).

A fertile field of research is the exploration of how these so-called context effects work (*18*). For example, Tourangeau and Rasinski (*19*) carried out an experiment to study context effects in attitude surveys. They presented each respondent with 4 issues at differing levels of familiarity (abortion, welfare, aspects of banking legislation, and proposed immigration legislation), using 4 different orders of presentation of the target issues (balanced in the form of a Latin square), 2 versions of the context questions used in advance of the target question (positive or negative), and 2 methods of structuring the context questions ("scattered" across issues with all context questions preceding the target questions or "massed" by issue with context questions followed by the linked target question). This yielded 16 versions of the questionnaire to which the investigators added 2 additional versions with neutral context questions, for a total of 18 versions. The responses of interest consisted of answers (favor versus oppose or agree versus disagree) to the four target issues (plus possible "don't know" responses). The notion was that, for familiar issues, scattered context questions would "prime" beliefs that would be retrieved and applied in answering the target question, whereas, for unfamiliar issues, massed context questions would help the respondents to interpret the target question better than would scattered ones (*20*).

Each interviewer used (approximately) a simple random sample of respondents from telephone banks listed in the Chicago directory. The interviewers received the questionnaires in batches of 18 and worked their way through a batch as they reached respondents willing to be interviewed (there was a 35% combined rate of refusal and nonresponse). There were originally 4 interviewers, each of whom was scheduled to carry out 5 batches of 18 interviews. In fact, some of these interviews were actually carried out by a fifth interviewer, so that for the 4 core interviewers we do not quite have 5 replications of the full experimental design. For the 4 core interviewers there were 353 completed interviews (of which 39 were in the neutral condition). Our analysis will be based only on these interviews by the core interviewers, ignoring the issues of nonresponse.

We can consider the 4 interviewers as blocks and, for the full experiment, within each block we have up to 5 replications of an 18-treatment experiment, where 16 of the treatments represent a $4 \times 2 \times 2$ factorial design. The outcomes for a given interviewer by treatment combination can be cross-classified according to the four dichotomous target response variables. Because of this categorical response structure, Tourangeau and Rasinski analyzed the "effects" measurable by this overall design using logit models. The interest here is in binary responses, and thus we focus on the probabilities associated with the two possible outcomes or the ratio of the probabilities, known as the odds. Logit models relate the logarithm of the odds for the binary responses variable to an additive model with components corresponding to the effects of explanatory variables and their interactions (*21*). For simplicity, our analysis will consider only the issues of abortion and welfare, and we shall consider them separately (*22*). Hence for each issue we have four treatment combinations (positive versus negative contexts, by scattered versus massed structures).

Table 2. Estimated logit effects associated with four models for the log-odds of a positive response to the abortion target question.

Parameter	Model			
	i	ii	iii	iv
Constant	−.65 (.38)	−.74 (.33)	−1.80 (.71)	−1.79 (.66)
Context	.71 (.44)	.71 (.43)	1.39 (.81)	1.23 (.80)
Mode (scattered)	.57 (.34)	.55 (.33)	.52 (.34)	.49 (.33)
All favorable responses	1.01 (.35)	1.02 (.34)		
Three or four favorable responses			1.93 (.67)	1.89 (.66)
Context by mode	−.72 (.47)	−.71 (.47)	−.65 (.47)	−.65 (.47)
Context by all favorable responses	−1.26 (.48)	−1.24 (.47)		
Context by three or four favorable responses			−1.61 (.81)	−1.41 (.80)
Interviewer 2	.25 (.33)		.34 (.33)	
Interviewer 3	−.47 (.33)		−.39 (.33)	
Interviewer 4	−.13 (.34)		−.03 (.33)	
Log-likelihood ratio statistic	14.1	19.0	25.45	30.3
Degrees of freedom	23	26	23	26

Table 3. Estimated logit effects associated with four models for the log-odds of a positive response to the welfare target question.

Parameter	Model			
	i	ii	iii	iv
Constant	.07 (.36)	.35 (.26)	.25 (.38)	.38 (.31)
Context	−1.72 (.44)	−1.56 (.42)	−2.59 (.70)	−2.30 (.68)
Mode (scattered)	.12 (.35)	.09 (.34)	.09 (.36)	.08 (.33)
All favorable responses	−2.02 (.42)	−1.71 (.39)		
Three or four favorable responses			−.98 (.35)	−.88 (.33)
Context by mode	.85 (.50)	.79 (.48)	.90 (.49)	.87 (.47)
Context by all favorable responses	3.00 (.57)	2.65 (.53)		
Context by three or four favorable responses			2.63 (.71)	2.30 (.68)
Interviewer 2	1.42 (.37)		1.13 (.35)	
Interviewer 3	−.01 (.35)		−.29 (.33)	
Interviewer 4	.09 (.35)		−.06 (.34)	
Log-likelihood ratio statistic	23.6	45.9	19.1	39.4
Degrees of freedom	23	26	22*	25*

*The values of degrees of freedom for columns iii and iv are reduced by 1 relative to those of columns i and ii, respectively, because of the presence of a zero marginal total.

The context-setting questions for the abortion issue dealt either with women's rights (coded 0) or with traditional values (coded 1). Those for the welfare issue dealt with fraud and waste in government programs (coded 0) or with governmental responsibility to provide services (coded 1). The context-setting questions were designed so that respondents would find them easy to agree with; indeed, 53% and 43% of the respondents agreed with all the abortion and welfare context items, respectively. The massed versus scattered (or mode) variable was coded 0 for massed, 1 for scattered, and the responses on the target questions were coded 0 for disagree and 1 for agree.

Our analysis of the results of this embedded experiment (*23*) began with an examination of separate logit models for predicting the log-odds of a positive response on the target questions ("abortion" and "welfare") that contrasted the four basic treatment combinations with the neutral context, while controlling for interviewer effects and their possible interactions with the treatment comparisons. Table 1 contains the estimated coefficients for interviewer effects and for the comparisons between the neutral context and each of the four treatments. When we compare the size of the estimated treatment effects with the size of their estimated standard errors, for "abortion" we see the suggestion of an effect due to the administration of the "women's rights" context in the scattered mode relative to a neutral context and a possible effect of interviewer 3 (versus the remaining interviewers). For "welfare," we see the suggestion of a somewhat different effect, due to the administration of the "fraud-and-waste" context in the massed mode relative to a neutral context and a relatively strong effect associated with interviewer 2 (versus the remaining interviewers). These treatment effects are puzzling, for there is no reason a priori to expect that these two out of eight manipulations should differ from the neutral treatment. We are comforted by the fact that they are only marginally statistically significant, at best, and merely serve to reassure us that the experimental manipulations did have some effect; we dissect the treatment effects more specifically below. Adding in the effects of the treatment by interviewer interactions (not shown in Table 1) yields substantively and statistically insignificant estimated effects. This result strongly suggests that going from the fixed-effects analyses of Table 1 to mixed-effects models will yield essentially no differences in conclusions. We now turn to a direct comparison of treatment effects.

We begin with the abortion issue and the logit model analyses reported in Table 2. The models are grouped in pairs, with the first pair showing a 0 to 3 versus 4 split on the number of favorable responses to the context questions and the second pair showing 0 to 2 versus 3 or 4 split. The two equations in each pair differ by the inclusion in the first of the estimated effects due to interviewers.

There appears to be little demonstrable effect due to interviewer on the log-odds of a positive response to the abortion question, and thus the first and second inference modes give similar results. Moreover, looking across all four models (Table 2), we see at best modest effects related to mode and to context by mode. The action in these models is associated with a context by response-to-context interaction.

On the answer to the abortion question, model ii appears to provide a far better fit to the data than model iv, a fact borne out by more detailed analyses, and thus we work with it to interpret this interaction. Focusing on those who responded unfavorably to at least one of the context questions, we find the odds of responding positively to the abortion question for those who were given the "traditional values" context are 2.03 ($e^{0.71}$) times as great as the odds for those given the women's rights context. But when we examine those who responded favorably to all four context questions, the odds are only 0.59 ($e^{0.71-1.24}$) as great. Answering negatively to traditional values questions seems to prime positive feelings about abortion, whereas answering negatively to questions on women's rights seems to prime negative feelings. On the other hand, responding positively to women's rights questions seems to prime positive responses to the abortion question, whereas answering positively to traditional values questions seems to prime negative attitudes toward abortion. We may speculate that the impact of women's rights context questions is influenced much more by the respondents' agreement or disagreement than is the impact of the traditional values context questions (Fig. 1 displays the marginal proportions). An alternate explanation is that the women's rights questions are, a priori, more closely related to the abortion question than are the traditional values questions.

The analyses in Table 2 are for fixed-effect models, and additional analyses with interaction terms involving interviewers strongly suggest that a mixed-effects model, with a random component for interviewers, would leave all our conclusions virtually unchanged. Here, then, we have a happy agreement across all three modes of statistical inference.

Next we consider the welfare issue and the logit model results in Table 3. For this issue, interviewer 2 has an estimated effect that distinguishes her from the other three interviewers. This highly statistically significant difference (*24*) focuses our attention on the results for models i and iii, where we see modestly sharpened effects for the inference approach that formally adjusts for the effects of interviewers. Both models i and iii fit their corresponding cross-classifications well, and it is worth considering both. In comparison with model iii, model i shows a heightened interaction between context and favorable responses; the high level of favorable response is defined in model i as agreement with all 4 context questions, whereas in model iii it corresponds with agreement with at least 3. Recall that the first level of context for the welfare issue involved questions on fraud and waste in government programs, which should lead respondents who agree with the context questions to oppose welfare, whereas the first level of context for the abortion issue, questions on women's rights, was chosen to lead respondents who agreed with the context questions to favor abortion. Thus for models i and iii in Table 3 we should expect the effects of context, favorable response, and their interaction to exhibit signs opposite from those for the corresponding effects in Table 2. We do indeed observe this expected reversal, together with magnitudes larger than those in Table 2. Thus we conclude that the context manipulation is more effective on the welfare issue than on the abortion issue.

Finally, we consider again the issue of the advisability of a random-effects component in our model for interviewers. We carried out a fixed-effects model analysis, adding nine additional interaction parameters linking interviewers with context, favorable response, and their interaction. There is some hint here that our inferences might shift a little if we switched to a random-effects component for interviewers (*25*). But we expect the first and third modes of inference to be quite similar and much sharper than the second mode, which ignores the differences associated with interviewer 2.

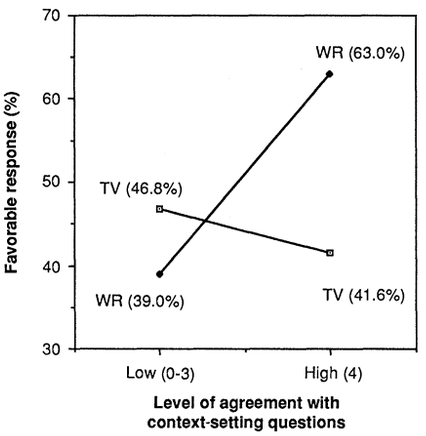

Fig. 1. Marginal proportions of respondents giving favorable answers to the abortion question; WR, women's rights; TV, traditional values. The crossover of response levels for WR and TV in going from low to high level of agreement is illustrative of an interaction between context and level of favorable response to context-setting questions.

Summary and Conclusions

In this demonstration analysis we have uncovered some interesting findings, both substantive and methodological. Both context manipulations seem to affect responses to the target question only in interaction with the respondents' responses to the context-setting questions. If respondents disagree with any of the context-setting questions, they are much more likely to endorse the abortion target question in the traditional values context than in the women's rights context. Respondents who agree with all the context-setting questions are less likely to endorse the abortion target question in the traditional values context than in the women's rights context. The context manipulation seems to have greater impact on the welfare target question than on the abortion target. We note especially the effect on the welfare target question of the interaction between context and the endorsement of the context-setting questions that is stronger than the corresponding interaction for the abortion target question.

For the abortion target there were no strong interviewer effects, so that there was no gain in this case in going from a model that excluded interviewer effects (our second mode of inference) to one that included terms for interviewers as fixed effects (our first inference mode). Because interactions involving interviewers were negligible, it was safe to assume that treating interviewers as random (our third inference mode) would give results similar to those found by the other modes. For the welfare target question, on the other hand, interviewer effects were present, so our first inference mode, including interviewers as fixed effects, was preferable to our second in that it sharpened the inferences we could make about the other factors in the model. The absence of higher order interactions with interviewers led us to believe that treating interviewers as random would not materially change the inferences.

We conclude that, although sometimes our three modes of inference agree, there are times when, perhaps unpredictably, they do not. Because data from surveys are used for so many important purposes, efforts to make them as accurate as possible are crucial. These efforts include the emphasis on reducing nonsampling error, the use of the paradigms of the cognitive sciences, and the embedding of methodological experiments in surveys. Our findings sug-

gest that they should also include care in the embedding design and the tailoring of analyses both to the design and for the population to which one wishes to generalize.

Doing a careful probability-based sample survey or a randomized experiment (in the social sciences or elsewhere) is a demanding activity, both intellectually and operationally. The rewards, however, we believe to be commensurate with the effort expended. Not only does a well-designed and executed survey or experiment provide the basis for statistical inference about the phenomena of interest but it also legitimizes the results. That is, the objectivity achieved instills in colleagues who examine procedures and results a confidence that is otherwise hard to achieve. The more scientists meet the standards of experimental control, full randomization, and fully achieved probability sampling (including avoidance of experimental attrition and nonresponse in surveys), the more others are willing to take their work seriously. And to the extent that such standards are relaxed, the work is appropriately taken less seriously.

Correctly designing, implementing, and analyzing an experiment embedded in a survey in the form advocated in this article are even more demanding tasks than doing a survey or a large-scale experiment. Informal discussions with those who have done embedded experiments suggest that the effort required often exceeds the sum of the efforts required to implement the survey and the experiment separately but is clearly less than their product. We argue that the extra effort invested in designing, conducting, and analyzing such embedded experiments will repay the investors and the social sciences sufficient rewards in the form of increased precision, generalizability, and credibility of results to warrant the cost.

REFERENCES AND NOTES

1. J. M. Converse, *Survey Research in the United States* (Univ. of California Press, Berkeley, 1987); N. M. Bradburn and S. Sudman, *Polls and Surveys: Their Use and Meaning* (Jossey-Bass, San Francisco, 1988).
2. W. G. Cochran, *Sampling Techniques* (Wiley, New York, ed. 3, 1977); L. Kish, *Survey Sampling* (Wiley, New York, 1965); F. Yates, *Sampling Methods for Censuses and Surveys* (Macmillan, New York, ed. 4, 1981).
3. F. Mosteller, in *International Encyclopedia of Statistics*, W. H. Kruskal and J. M. Tanur, Eds. (Free Press, New York, 1978), vol. 1, p. 208.
4. S. L. Payne, *The Art of Asking Questions* (Princeton Univ. Press, Princeton, NJ, 1951) is representative of the art. S. Sudman and N. M. Bradburn, *Response Effects in Surveys: A Review and Synthesis* (Aldine, Chicago, 1974); N. M. Bradburn, S. Sudman, and Associates, *Improving Interview Method and Questionnaire Design* (Jossey-Bass, San Francisco, 1979); and H. Schuman and S. Presser, *Questions and Answers in Attitude Surveys* (Academic Press, New York, 1981) are illustrative of the scientific approach.
5. T. B. Jabine, M. Straf, J. M. Tanur, R. Tourangeau, Eds., *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines* (National Academy Press, Washington, DC, 1984); S. E. Fienberg, E. F. Loftus, J. M. Tanur, *Milbank Mem. Fund Q.* **63**, 547 (1985).
6. N. M. Bradburn, L. J. Rips, S. K. Shevell, *Science* **236**, 157 (1987).
7. D. C. Falthi, J. Schooler, E. F. Loftus, *Proc. Soc. Stat. Sect.* (American Statistical Association, Washington, DC, 1984), p. 19; H.-J. Hippler, N. Schwarz, S. Sudman, Eds. *Social Information Processing and Survey Methodology* (Springer-Verlag, New York, 1987); J. T. Lessler and M. G. Sirken, *Milbank Mem. Fund Q.* **63**, 565 (1985); N. Schwarz, paper presented at the ninth meeting of the Working Party on Employment and Unemployment Statistics, Organization for Economic Cooperation and Development, Paris (April 1987) [OECD-Document MAS/WP 7(87)4]; A. A. White and M. L. Berk, *Proc. Soc. Stat. Sect.* (American Statistical Association, Washington, DC, 1987), p. 66.
8. S. E. Fienberg and J. M. Tanur, *Can. J. Stat.* **19**, 135 (1988).
9. ———, *Int. Stat. Rev.* **55**, 75 (1987).
10. R. A. Fisher, *The Design of Experiments* (Oliver and Boyd, Edinburgh, 1935), chap. 4; O. Kempthore, *The Design and Analysis of Experiments* (Wiley, New York, 1952), chap. 9.
11. P. C. Mahalanobis, *J. R. Stat. Soc.* **109**, 325 (1946).
12. J. Durbin and A. Stuart, *J. R. Stat. Soc. Ser. A* **114**, 163 (1951).
13. See, for example, R. Rosenthal and D. B. Rubin, *Behav. Brain Sci.* **3**, 410 (1979).
14. For general approaches to mixed-effects ANOVA models, see: O. Kempthore and M. Wilk, *J. Am. Stat. Assoc.* **27**, 950 (1955); H. Scheffé, *Ann. Math. Stat.* **27**, 23 (1956). For the use of such models in the survey context, see H. O. Hartley and J. N. K. Rao, in *Survey Sampling and Measurement*, N. K. Namboordiri, Ed. (Academic Press, New York, 1978), chap. 4, p. 35.
15. See H. Scheffé, *The Analysis of Variance* (Wiley, New York, 1959), chap. 8, for a detailed exposition of estimation in mixed-effects ANOVA models and for the related variance formulae.
16. I. Crespi and J. W. Swinehart, paper presented at the meeting of the American Association for Public Opinion Research, 1982.
17. These effects are discussed widely in the literature of survey research. See, for example, S. Sudman and N. M. Bradburn, *Asking Questions: A Practical Guide to Questionnaire Design* (Jossey-Bass, San Francisco, 1982).
18. For example, G. F. Bishop, R. W. Oldendick, A. J. Tuchfarber, *Public Opinion Q.* **48**, 510 (1984); H. Schuman and S. Presser, *Questions and Answers in Attitude Surveys* (Academic Press, New York, 1981).
19. R. Tourangeau and K. A. Rasinski, "Context effects in attitude surveys" (unpublished manuscript, 1987).
20. The concept of priming used here is based on the cognitive science heuristic of reliance upon information that is most accessible or available [see A. Tversky and D. Kahneman, *Science* **185**, 1124 (1974); *ibid.* **211**, 453 (1981)]. In this embedded experiment the context questions help to frame the issue in one of two schemas that we expect will influence the response item. Massed context items were hypothesized to inspire counterarguments and hence activate the opposite schema. For a good illustration of priming effects in a series of political science experiments, see S. Iyengar and D. R. Kinder, *News That Matters* (Univ. of Chicago Press, Chicago, 1987), chaps. 7 to 11.
21. See S. E. Fienberg, *The Analysis of Cross-Classified Categorical Data* (MIT Press, Cambridge, MA, ed. 2, 1980); P. McCullagh and J. Nelder, *Generalized Linear Models* (Chapman & Hall, New York, 1982).
22. Except for the elaborate structure exploring interviewer effects and their possible interactions, the analysis we carry out here is similar in spirit to that done by Tourangeau and Rasinski. We made several analysis decisions, however, that differed from theirs—in particular, different cutoffs designating high and low agreement with the context questions. Thus our analysis does not track theirs exactly, and our conclusions differ somewhat from theirs.
23. In these and subsequent logit analyses we used the GLIM statistical computer program [see R. J. Baker and J. A. Nelder, *The GLIM System: Release 3* (Numerical Algorithms Group, Oxford, 1978]. GLIM is a program developed by the Royal Statistical Society for fitting generalized linear interactive models, of which logit models are a special case.
24. This comparison is reinforced by a comparison of the likelihood ratio statistics for models i and ii and for models iii and iv. Although the signs of the remaining estimated coefficients in columns ii and iv are the same as in columns i and iii, respectively, the magnitudes of the estimated effects for context, for favorable responses, and for their interaction all shrink noticeably.
25. Actually implementing such a mixed model logit analysis is relatively complex. See the detailed approach in D. A. Anderson and M. Aitkin, *J. R. Stat. Soc. Ser. B* **47**, 203 (1985) and its subsequent implementation in L. Stokes, *J. Am. Stat. Assoc.* **83**, 623 (1988). In the present circumstance, where we have only a small number of interviewers and no surprisingly large interviewer-related interactions, the anticipated changes from a fixed-effects analysis are small.

26. We thank R. Tourangeau and K. Rasinski for providing the unpublished data from their embedded experiment and computer printouts from their analyses, and Y. Ding for computational assistance. W. Kruskal, J. Lessler, and H. Schuman and the referees provided us with valuable comments on an earlier draft. Supported in part by NSF grant SES-8701606 to Carnegie Mellon University and under grant SES-8701816 to the State University of New York at Stony Brook.