

TO: Interested Persons
FROM: Maria Sanchez and Giovanna Morchio
RE: Probing Don't Know Answers—Do we always want to do this?

Methodological research has shown that interviewer behavior and the wording of survey questions can be manipulated in order to lead respondents to articulate opinions about obscure and fictitious issues. The manipulations discussed most frequently in the literature are of two sorts. On the one hand, the question wording read to the respondent excludes the explicit choice of a 'don't know,' 'no opinion' or 'haven't thought much about it' response alternative. This situation encourages respondents to choose an answer from among the substantive response options offered in the question text. As a result, the 'no opinion' type of answers are elicited with much less frequency under this question format than would be the case if the option to bail out had been offered to the respondent as part of the question text.

The second type of manipulation focuses on interviewer behavior. The interviewer is instructed to repeat the question, or to probe the answer given whenever a 'no opinion' type of answer is volunteered by a survey respondent. The interviewer acts as if he or she expected a definite answer from the respondent who frequently complies by redirecting the final choice to one of the substantive response options.

SRC's interviewer training materials emphasize the need for interviewers to always probe a "don't know" (DK) answer at least once, by having the interviewer repeat the question or utter standardized probing phrases such as "Well, what do you think?", "Most of us don't know for sure, but what do you think?", or "What would be your best guess?". After initial training, SRC supervisors reinforce continuously the rule of always probing DK answers at least once. Interviewers are reminded to probe DK answers whenever superiors detect violations in the course of live monitoring sessions or while reviewing completed interviews for evaluation. (Supervisors are able to detect whether answers were probed while reviewing written materials because interviewers are trained to record, in standardized shorthand notation, the type of probe used as a means of documenting which questions were probed and which phrase was used.) Telephone interviewers in the centralized Telephone Interviewing Facility at ISR are reinforced more regularly and more frequently than field interviewers in the use of interviewing techniques. Telephone supervisors have more immediate and direct contact with their interviewers. While field interviewers use the same training materials as telephone interviewers to learn about interviewing techniques, field supervisors are unable to monitor the exchange between respondents and their interviewers in an unobtrusive or timely manner, and they get to review only 10-15 production interviews from each of their interviewers for a typical study.

With this information as background, this memo addresses various issues. First, it provides data to support the notion that probing DK responses may elicit from respondents information of questionable validity (so-called pseudo-opinions). Second, it documents the fact that interviewer training practices with regard to probing have a clear and direct bearing on the nature of the estimates obtained from surveys. Third, it raises the issue of what to do in the future with regard to the training of SRC interviewers, including the possibility of doing nothing while we work on more systematic experimentation dealing with the interaction of types of questions and the appropriate probing of DK answers depending on question type.

The research reported here was carried out by NES staff and will be further reported as part of the existing series of NES working papers on the technical characteristics of the NES data collection.

The 1984 NES Study

The 1984 NES Study provides vivid and serendipitous confirmation of the fact that pseudo-opinions are elicited from survey respondents when interviewers probe DK answers systematically.

All 1984 NES respondents were initially interviewed face-to-face during September-October of the election year. The study design called for a second (post-election) interview with all pre-election study participants but, for a number of reasons, the decision was made to reinterview half of the respondents in the field and the other half from the Telephone Interviewing Facility at ISR. Accordingly, near the end of the pre-election study, the sample was divided into random halves. The face-to-face (FTF) half-sample was reinterviewed in November-December by SRC interviewers who had worked on the pre-election study. The telephone (TEL) half-sample was also reinterviewed during the months of November and December by SRC telephone interviewers.

Some of the pre-election respondents scheduled for a telephone reinterview were interviewed in person at the time of the post-election study. Those handled this way included persons who had refused to disclose their telephone numbers when first interviewed, those living in households without telephones, and all persons identified by the pre-election interviewer as unlikely to tolerate a telephone reinterview for a variety of reasons (too old, poor hearing, ill-health, etc.). (The pre-election coversheet included a question addressed to interviewers to help us identify these special cases.) Non-telephone cases and all cases flagged by interviewers as candidates for a FTF reinterview are excluded from the analysis reported below, without regard to the half-sample into which the cases were subsequently randomized. This procedure results in purified half-samples (FTF/TEL) which, statistically speaking, have comparable demographic characteristics with regard to age, education, income, sex, race, marital status, employment status, region of the country, and population density of the area of residence.

Put another way, the differences that emerged while comparing the two half-samples cannot be attributed to biases in the demographic composition of either sample. Nor are differences explained by differential treatment of respondents on the part of interviewers who conducted the pre-election study. At the time this study was taking place, interviewers had no knowledge of the half-sample assignments for individual study participants. Finally, differences in the data collection periods between the FTF and the TEL samples cannot account for the results obtained since the field periods were identical for both studies.

Analysis

A systematic difference was detected early in comparing post-election results obtained from the two samples. The FTF sample had a higher overall incidence of DK answers over a large number of variables referencing various content areas. Nowhere was the difference more striking than in the answers to these four items dealing with factual and verifiable political information matters:

J2. Do you happen to know which party had the most members in the House of Representatives in Washington before the election (this/last) month? (IF NECESSARY: Which one?) [V5751]

<u>TEL</u>		<u>FTF</u>		
<u>N</u>		<u>N</u>		
132	17%	76	9%	1. REPUBLICANS
502	63%	455	56%	5. DEMOCRATS
162	20%	286	35%	8. DON'T KNOW
6	1%	1	*	9. NA

J3. As a result of the election (this/last) month, which party will now have the most members in the House of Representatives? [V5752]

<u>TEL</u>		<u>FTF</u>		
<u>N</u>		<u>N</u>		
211	26%	159	19%	1. REPUBLICANS
471	59%	418	51%	5. DEMOCRATS
114	14%	238	29%	8. DON'T KNOW
6	1%	3	*	9. NA

J4. Do you happen to know which party had the most members in the U.S. Senate before the elections (this/last) month? (IF NECESSARY: Which one?) [V5753]

<u>TEL</u>		<u>FTF</u>		
<u>N</u>		<u>N</u>		
302	38%	233	29%	1. REPUBLICANS
278	35%	199	24%	5. DEMOCRATS
213	27%	384	47%	8. DON'T KNOW
9	1%	2	*	9. NA

J5. As a result of the election (this/last) month, which party will now have the most members in the U.S. Senate? [V5754]

<u>TEL</u>		<u>FTF</u>		
<u>N</u>		<u>N</u>		
376	47%	286	35%	1. REPUBLICANS
273	34%	191	23%	5. DEMOCRATS
145	18%	339	41%	8. DON'T KNOW
8	1%	2	*	9. NA

The puzzling initial conclusion appeared to be that, for some unknown reason, the TEL respondents were more informed about politics than their counterparts reinterviewed in the field. However, this conclusion was not supported by other survey results and, in particular, by the answers to another post-election factual item in which respondents were asked to name the House candidates running in their districts and to provide each politician's partisan affiliation. For this question, about the same proportion of respondents (45% TEL, 46% FTF) indicated they could give the information requested, but FTF respondents had a higher overall number of correct mentions than TEL respondents.

Having ruled out earlier other extraneous sources of variation having to do with half-sample demographic composition, differential treatment of pre-election respondents and noncomparable field periods, we looked at the interviewers themselves as a possible explanation. And here there was a large known difference between the field and the telephone interviewing staffs. Virtually all of the field interviewers were newly-hired for the pre-election study either to replace terminated interviewers in self-representing areas, or in order to staff the SRC 1980 National Sample primary areas where SRC had never interviewed before. In contrast, the telephone interviewers were winding down from a very busy year in which the demands for interviewing had run extremely high. So, the two staffs of interviewers differed significantly with regard to interviewing experience, and it is a well-known fact that less-experienced interviewers have a tendency to accept DK answers without probing. Further, the centralized setting of the Telephone Facility will correct sooner for this tendency since the interviewing techniques of telephone interviewers are reinforced with greater regularity than those of field interviewers by virtue of their proximity to the supervisory staff. For these reasons alone, we would have expected fewer DK answers in the survey results returned by telephone interviewers compared to field interviewers. However, the magnitude of the observed differences, especially as regards the questions above, did surprise us.

DK Probing — Does it elicit mere guesses or valid information?

Attitudinal and opinion survey questions do not usually have right or wrong answers. These types of questions offer less desirable testing ground to explore the effects of DK probing than factual information items. Fortunately the NES quartet displaying large differences in the proportion of DKs across samples consists of questions with 'true' or 'false' answers. These questions are especially useful to analyze the quality of the information elicited from respondents after probing DK answers.

The four items can be combined to produce a new variable indexing the number of correct mentions each respondent gave out of a possible total of four. Since each question presents a dichotomous choice (T or F), one can argue that by chance alone the combination of 2T and 2F items would be the most frequent occurrence when someone answers by merely guessing for each of the four questions. To be precise, the probability of this event is .375. Next in line as outcomes predicted by chance alone are the combinations '3T-1F' and '1T-3F'. Each of these outcomes has an expected probability of 0.25. The most infrequent outcomes obtained by guessing are represented by the combinations '4T-0F' and '0T-4F', with an expected probability of .063 each.

If DK probing elicited real information among TEL respondents, then a couple of things would follow. First, the joint distribution represented by the index would show a higher proportion of TEL respondents than FTF respondents in cell outcomes corresponding to combinations on the high end of the political information index (all correct mentions; 3 out of 4 items correct). Conversely, proportionately fewer TEL than FTF respondents would be represented in the low information end of the index (all items incorrect).

On the other hand, if DK probing elicited mere guesses, the sample index proportions for the TEL and FTF samples would be fairly comparable for most cells except two: the one corresponding to the most favored outcome predicted by chance (2 correct mentions out of 4 possible), and the cell corresponding to zero correct mentions represented by 4 DKs. Further, these two cells should compensate each other across the two samples. The TEL index distribution should show a higher proportion of respondents than the FTF distribution in the outcome category "2 out of 4" where most of the telephone "guessers" would

land, and the difference in these proportions should be mirrored in a comparable disproportionate bulging in the FTF proportion for the cell corresponding to the 4 DK answers which field interviewers did not probe systematically. Beyond this, we would predict proportionately more TEL than FTF respondents in the next most likely outcome categories, the cells corresponding to the patterns '3T-1F' and '1T-3F'. Finally, small differences in the sample proportions would be predicted for the least likely outcomes, '4T-0F' and '0T-1F'.

The results presented below support the hypothesis that DK probing elicits "mere guesses", at least as far as these factual items are concerned. The empirical results displayed in the table validate the predictions derived from theoretical considerations. The relative probabilities of various index outcomes in the FTF and TEL samples match results predicted under conditions involving answer guessing on the part of uninformed TEL respondents.

Further validation can be provided that DK probing elicited information of questionable validity from among uninformed TEL respondents in the 1984 NES Study. Respondents in both samples were asked to state their level of interest in the 1984 political campaigns—they were asked whether they were very interested, somewhat interested or not much interested in following the election campaigns in 1984. Generally speaking, subjective political interest should predict fairly accurately a person's ability to provide informed answers about factual questions on the subject of politics. In particular, interest in the 1984 campaigns should be a strong predictor of the degree of response accuracy for the 4 NES factual items since the latter deal with the balance of partisan power in Congress both before and after the 1984 elections. This predictive association of the relationship between political interest and the accuracy of factual information is stronger among FTF than TEL respondents. 46% of the FTF sample respondents who described themselves as not much interested in the campaigns answered DK to all four political information items while the comparable figure among the uninterested TEL respondents is 24%. Among persons somewhat interested in the campaigns, 21% of the FTF respondents and 9% of the TEL respondents gave 4 DK answers. For this same group, 13% of the FTF participants answered two of the four items correctly, while 35% of the somewhat interested TEL respondents manifest that same level of accuracy. These large differences disappear when the TEL and FTF subgroups comprised of those very much interested in the campaigns are compared. Four correct answers were given by 30% of the very much interested FTF respondents and by 37% of the comparable TEL group.

INDEX CATEGORIES BASED ON 4 ITEMS

TEL HALF-SAMPLE N %		NUMBER CORRECT MENTIONS	NUMBER INCORRECT MENTIONS	NUMBER FINAL DK MENTIONS	FTF HALF-SAMPLE N %	
178	22.2	4	0	0	170	20.8
49	6.1	3	1	0	19	2.3
232	28.9	2	2	0	123	15.0
44	5.5	1	3	0	21	2.6
18	2.2	0	4	0	21	2.6
15	1.9	3	0	1	21	2.6
48	6.0	2	1	1	36	4.4
24	3.0	1	2	1	18	2.2
4	0.5	0	3	1	8	1.0
28	3.5	2	0	2	67	8.2
38	4.7	1	1	2	61	7.5
9	1.1	0	2	2	18	2.2
17	2.1	1	0	3	39	4.8
9	1.1	0	1	3	19	2.3
78	9.7	0	0	4	174	21.3
11	1.4	NOT ASCERTAINED			3	0.4
802	100.0				818	100.0

CONCLUSION

These results raise many questions:

- 1- Are attitudinal/opinion items in the same category as factual items with regard to the effects of probing DK answers? That is, would DK probes also elicit merely "guesses" for this other general class of survey questions, or do NES results generalize only to questions with right and wrong answers?
- 2- What about factual questions which are not 'test' items such as the NES information questions, but estimates the respondent provides about himself/herself, his/her income or health behavior, or some such class of personal questions which also have right and wrong answers even though investigators will probably not have access to verification sources? Do we get better 'guesses' when DK answers are probed for this type of question than for external factual items?
- 3- Are there different types of DK answers? Anyone who has ever spent time monitoring interviews may already suspect that there are at least two qualitatively different types of DK answers. One means "I haven't the foggiest" while the other signifies "I have some idea, but I am unsure or undecided". Both answers are articulated by respondents with the same words, and it is the intonation or the emphasis, or additional peripheral phases which give away the true meaning. Do we probe the unsure/undecided variety only? If so, how do we teach interviewers, especially trainees, to distinguish between the two meanings?
- 4- Should investigators continue to evaluate survey data quality, as they frequently do, by counting how many DKs were recorded by interviewers throughout a questionnaire administration? Do we believe that a low overall incidence of DK answers is indicative of better quality data? More pointedly, the higher incidence of DK answers in the NES FTF half-sample is not an indicator of lower quality of interviewing data on the part of newly-trained SRC field interviewers. On another often used measure of interviewing data quality—the number of substantive mentions elicited by interviewers for open-ended questions in a study—there is either no difference between the TEL and the FTF samples, or the difference favors the field interviewers. NES post-election questions asked in open-ended format are: likes and dislikes of House candidates running in the respondent's district of residence, and identification of the most important problems facing the country these days. The likes/dislikes set reveals no significant differences in the number of mentions obtained by field and telephone interviewers, while the FTF sample yields more important problems mentions than the TEL sample. The NES pre-election study had included many open-ended questions and study-specific training was designed for field interviewers working on that study. This training made provisions for drilling interviewers on the handling of open-ended questions through the use of scripted interviews with answers that required probing. Training makes a difference; the location of interviewers is less relevant.
- 5- What should the SRC Field Section do, or should it do something, to standardize the behavior of field and telephone interviewers in this area of DK probing? A difference exists now and the issue is whether the training of either group of interviewers should be modified to achieve comparability in this specific area.
- 6- Can we do something unobtrusively in conjunction with regular production data collection to gather additional information on this matter over the immediate future? We have ideas

and would like to hear those of others.

- 7- How should we in the Field Section alert principal investigators to this potentially troublesome problem? Which questionnaire construction conventions should we adopt to alert interviewers about the presence of questions they should not probe?
- 8- The benefits of a centralized interviewing setting have been frequently praised. The liabilities have not been discussed nearly as much. For better or worse, interviewing practices are propagated faster and reinforced more strongly in a centralized setting. When production problems are detected, as in the case of DK probing, the effects are likely to be stronger than the ones resulting in a decentralized setting. The situation is similar to the one we face with CATI. The computer-assisted technology controls for much of the random measurement error potentially introduced by interviewers in paper-and-pencil administrations, when questions that should have been asked are skipped and inappropriate question sequences are administered. On the other hand, this same CATI control mechanism is quite capable of generating large systematic measurement errors when program bugs are diagnosed in a CATI application after the data collection has ended. Control and centralization entail costs and benefits as everything else in life.