Ordinary Questions, Survey Questions, and Policy Questions

HOWARD SCHUMAN

A FUNDAMENTAL paradox of survey research is that we start from the purpose and use the method of ordinary questioning employed in daily life, yet our results are less satisfactory for this purpose than for almost any other. This paradox in turn complicates the use of survey results in relation to policy questions in ways that neither the public nor the policy makers understand.

Let me first explain what I mean by the paradox, then consider its causes, and finally offer an example of one practical way of dealing with it.

In ordinary life a question is typically asked because one person wishes information from another. You might ask an acquaintance how many rooms there are in her new house, or whether she enjoyed a session at this conference, or if she favors the legalization of abortion in America. The normal assumption on both sides of the interaction is that you are interested in her answer in and of itself. Your purpose might vary from wanting advice based on her experience to trying to get a sense of her general values, but in any case you usually evaluate her answers as such and make judgments based on them. Let us call this process *ordinary questioning* and the questions themselves *ordinary questions*.

In survey interviews we ask much the same kinds of questions—that is, their form, their wording, and the manner of their asking are hardly distinguishable from ordinary questions. We do at times use fancy formats, with names like Likert-type and Semantic Differential; but, by and large, the survey questioner cannot depart too far from ordinary questioning, for the essential nature of the survey interview is communication with people who *expect* to hear ordinary questions.

Howard Schuman, Director of the Survey Research Center and a Professor of Sociology at the University of Michigan, was President of AAPOR in 1985–86. This is a slightly revised version of his presidential address at AAPOR's 41st Annual Conference, St. Petersburg Beach, Florida, May 17, 1986. Helpful comments on a previous draft were made by Stanley Presser and Carlota Smith. The address was written while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences. The author is grateful for financial support provided by National Science Foundation #BNS-8011494 and SES-8411371.

Indeed, the major justification we give for entering a respondent's home and taking her time is that we seek information—whether facts or opinions or both. The whole interview is couched in the framework of ordinary questioning. Not surprisingly, the respondent assumes that the interviewer is directly interested in the facts and opinions she gives, just as would an acquaintance who asked the same questions. By this I don't mean that the respondent necessarily assumes a personal interest in her answers, though of course sometimes that is the case. What the respondent does assume, however, is that her answers will be added to those of all others who are interviewed, and tallied up to give totals that are directly interpretable. If opinions are asked for, the survey is seen as a kind of referendum, and the investigator is assumed to be interested in how many favor and how many oppose legalized abortion or whatever else is at issue. If facts are being sought through the survey, the assumption is that the goal is a table showing how many people have what size homes, or whatever.

Yet experienced survey researchers know that it is often just such tables that are most problematic to interpret. Especially when we deal with opinions, but even when we deal with what are called facts, the results for any *single* question—what we call the marginals—are usually too much a function of the way the question is asked to allow for any simple interpretation. The results on abortion depend heavily on the conditions and definitions presupposed in the question wording, and the same is true to an extent even for the question on how many rooms you have in your house. Don't trust the marginals in any absolute sense is one of the first lessons a person should learn when working with survey data.

Instead, what we do is to hold constant the question—or the index, if more than a single question is being considered—and make comparisons across time or other variables. We may not know what 65 percent really means in terms of support for legalized abortion, but we act on the assumption that if the question wording and other survey conditions have been kept constant, we can say, at least within the limits of sampling error, that this represents such and such an increase or decrease from an earlier survey. Or that if 61 percent is the figure for men and there is a quite different figure for women, a sex difference of approximately such and such a size exists. There are obviously assumptions in all this that are problematic, such as whether constancy is really maintained over time or across other variables, and also whether relations are really less affected by question wording than are marginals. But those assumptions are not my concern here: I am interested in much simpler but in a way more basic issues.

The difference between what the respondent expects to be done with

her answers—what we might call the referendum point of view—and what the sophisticated survey investigator expects to do—the analytic point of view—is a very large difference. The respondent in a national survey believes that the investigator is going to add up all the results, item by item, and tell the nation what Americans think. But the investigator knows that such a presentation is usually problematic at best and can be dangerously misleading at worst. Thus there is an important difference in the goals of the two main groups of actors in the survey process—the respondents and the investigators—and among other difficulties this difference can lead to ethical problems having to do with informed consent. "Your Opinion Counts," says the blurb, but more often than not your opinion does not, and indeed should not, count in the simple referendum sense that respondents assume.

Moreover, to make matters even more awkward, policy makers often have the same point of view as respondents: they want to know how many people favor and how many oppose the issue that they see themselves as confronting. Yet it may be neither possible nor desirable for the survey to ask exactly the question the policy maker has in mind, and in any case such a question is likely to be only one of a number of possible questions that might be asked on the issue.

Because of the expectations of both the public and the policy makers, referendum-type assertions based on surveys are constantly being made, and all of us feel pressure to continue to do so. Here are several examples, all from good organizations. One recent poll reports that 83 percent of Americans agree that the manned space shuttle program should continue, and that 70 percent say civilians should be included as astronauts—no comparative analysis being offered. As a second example, an article in *Science,* one of the preeminent publications in America, reports that two-thirds of the respondents in a national sample said they would be willing to take a polygraph test, yet neither comparative analysis nor even question wording is provided. And as a final example, a survey that I had supervised was reported by my university's news bureau as finding that 61 percent of Michigan residents favored the release of dogs and cats by public pounds for use in medical experiments, with a "margin of error" of 3.5 percent. This last example is one I will come back to later.

Problems with the Referendum Point of View

At this point it is useful to say something about why the referendum point of view—that is, the emphasis on marginals—is not a good one in most surveys. One explanation sometimes given for the undependability of marginals is that answers in surveys are frequently affected by

small, even trivial, changes in question wording. I think that this explanation is not the main one. At the very least it is exaggerated, and it may even be theoretically nonsensical.

In the book *Questions and Answers in Attitude Surveys,* written with Stanley Presser, we tried in a number of split-ballot experiments to shift marginals by means of small or even not so small changes in wording and were usually either unsuccessful or produced only minor shifts. For example, in one experiment we substituted the phrase *end pregnancy* for the word *abortion,* having been told that the former was a softer way of putting the matter and likely to yield more support for abortion. Yet the alteration produced no change at all in marginals. In another instance, we removed what seemed to be clearly loaded wording from an alternative to a question on work values and, to our surprise, again found no effect. All in all, our deliberate attempts to shift marginals by variations in tone of wording or phrasing were noteworthy more often for the negligible than for the large effects produced.

These negative results are supported by the theoretical difficulty of defining exactly what is meant by a small change in wording. Does *small* refer to the size of the word changed? Or to its grammatical character? Or to what? Consider the difference between the words *a* and *the*. Such articles in English tend to be treated as unimportant words, yet it is really no surprise that substitution of one for the other can have large effects. For example, if we say that the secretary of state shook hands with the wife of the Egyptian ambassador, this implies something quite different than if we say that the secretary of state shook hands with *a* wife of the Egyptian ambassador. The shift from *the* to *a* moves us from monogamy to polygamy. It would be odd to say that such a change in wording is small, just because articles are less conspicuous in English than nouns or verbs.

No, it is not usually tiny changes in wording that make marginals so untrustworthy, but several other factors about questions. I will take time to stress only two. First, respondents feel enormously constrained to stay within the framework of a question. They will almost always use one of the two or three alternatives given by the interviewer, rather than offering a substitute of their own. Given another alternative or a different frame of reference, however, their answers—and the marginals—might shift substantially.

Let me give one recent example. Using an open version of the standard question on the Most Important Problem facing the United States, Jon Krosnick, Jack Ludwig, and I identified the main issues mentioned by Americans at that point in time—for example, unemployment and inflation were then the leading issues. Then we constructed a new

version of the same question that presented just four alternatives, *none* of which had been mentioned spontaneously as most important by more than 1 percent of the population—for example, "quality of the public schools." We also added to the question the explicit option that the respondent need not stick to the four alternatives mentioned, but could give anything else that he or she regarded as a more important issue. Despite this option, 72 percent of our national sample chose one of the four listed alternatives. In other words, even when told explicitly that they were not limited to the choices we presented, most respondents stayed within the bounds of exactly those choices. Here, as elsewhere, marginals reflect, first and foremost, our decisions as investigators, and only within that framework do they reflect the choices of most respondents.

A second and related factor is that most public issues are discussed at a rather general level, as though there is just a single issue and it has just two sides. But what is called the abortion issue, to take one example, consists of a large number of different issues having to do with the reasons for abortion, the trimester involved, and so forth. Except at the extremes, exactly which of these particular issues is posed and with what alternatives makes a considerable difference in the marginal results, and thus any single set of marginals is likely to be misleading if taken to summarize views on abortion as a general issue. And the same is true for most other serious issues. Thus, it is genuine change in question content that is responsible for most important shifts in marginals, not minor variations in wording.

These two factors are, I believe, among the most important ones in accounting for the difficulty in interpreting marginals, but they are certainly not the only ones. There are numerous issues where the public is so ignorant that the marginals have little meaning and less staying power. There are marginals that are partly due to question or response order effects, probably because the meaning of the question varies from one order to another. And there are marginals that are quite misleading if extrapolated beyond the present moment, because they fail to take into account the impact that national leadership or likely events can have on opinion on a wide range of issues.

Furthermore, beyond problems having to do with the questions we ask, we always need comparative data to make good sense out of results. Suppose that at the end of this conference each of us answered a simple yes/no question about whether the conference had been a valuable experience, with the yield being 60 percent yes's and 40 percent no's. Leaving aside all the problems of question wording and answering, the 60 percent could only be sensibly interpreted against a backdrop of results on other conferences. If the average for yes's over

the past several years had been 40 percent, the organizers this year could well feel proud of their success; if the average over past years had been 80 percent, this year's organizers might well hang their heads in shame. We are all aware of the fundamental need for this type of comparison, yet it is all too easy to forget about the difficulty of interpreting absolute percentages when we feel the urge to learn about public reactions to a unique event.

From all these considerations it becomes clear that the presentation of univariate percentages is fraught with danger, and that satisfying the public hunger for results in the form of referenda should usually be, if not avoided, then put forth with cautions and in subdued tones—unless, of course, there is a simple official referendum that the survey is in fact attempting to predict word for word. I don't want to deny that the public does have opinions on certain major issues, and that all of us at times want to know what those are even when no comparison points exist, but rather that any attempt to summarize those opinions by presenting one or two univariate distributions is quite apt to be misleading.

It is worth asking whether this conclusion applies also to what I have called ordinary questioning in a two-person conversation. In part, I think it does. Answers to ordinary questions, if not followed up, can sometimes be just as misleading as marginals, and there is also probably a good deal more distortion due to social pressure in ordinary questioning than in the survey interview. Still, ordinary questioning in genuine conversation often does allow a fair amount of credence to be given to answers as such, because the questioner is able to adapt questions to the individual answering, to probe deeply, and to continually synthesize the answers received. None of this can be permitted to survey interviewers for fear of both bias and interviewer variance, and that is, of course, the price we pay for the undoubted advantages of standardization. Thus in ordinary questioning one can end up with an integrated sense of what the answerer thinks about a general issue, which is seldom possible in the context of the survey interview.

Three Strategies

This brings me to the last point I want to raise: What can survey investigators do to take account of the inherent ambiguity of marginals? There seem to be three main strategies, assuming that past time points are not available for comparison. One is to include in all important surveys—and who among us wishes to acknowledge doing unimportant surveys—at least some open-ended questions that allow investigators and readers to appreciate, so far as possible, the frames of

reference and alternatives that come most spontaneously to respondents. I don't mean to claim that one can ever avoid providing some initial frame in asking a question, nor that open questions are without serious limitations of their own. But they do offer one way for an investigator to avoid mistaking a mirror of his own mind for a window into the minds of others. Of course, open questions can be especially useful as part of an initial pilot study, as Lazarsfeld proposed long ago, but the need for an adequate number of cases and a serious look at the results suggests that at least some open questions be retained in most final surveys.

A second strategy is to ask a wide variety of questions about an issue, including a spectrum of information and belief questions, as well as attitudes, and then give a number of univariate results if any are to be given at all, thus emphasizing the complexity of the issue. And, of course, the exact wording of each question should be given also. This will obviously be difficult to do in brief newspaper story form, but it is essential to try, and many good survey reports do just this. Furthermore, this strategy requires a good deal of forethought in question construction if it is not to conflict with the equally desirable need to create a small set of permanent indicators that can be replicated over time to monitor social change.

I would like to suggest still a third strategy that is especially useful when dealing with controversial issues like abortion, where the different sides may not even frame key questions in the same way. The idea is to enlist representatives of opposing sides to help in developing questions that take account of their different frames of reference. Such a strategy can provide both the reality and the appearance of fairness. Both reality and appearance are important, the former for obvious reasons and the latter because otherwise any set of results on a controversial issue is apt to be attacked as biased by the side that was not consulted and that feels disadvantaged by the outcome.

Here is one example of this strategy, which I became involved in somewhat by accident. Last summer a University of Michigan Medical School professor asked my research center to carry out a survey about the use of animals in medical experiments, and more specifically about the transfer of unclaimed dogs and cats from dog pounds to laboratories for use in such experiments. The Michigan legislature was soon to consider a bill prohibiting such transfers from pounds, and the Michigan Society for Medical Research wanted to show the relevant legislative committee that public sentiment opposed such a bill.

Views concerning experimentation on animals are like debates about abortion and other moral issues: there are highly committed groups at

both extremes—and I do not use the word *extremes* in an invidious sense—and then a large part of the population that can easily be made aware of the issue but has not really thought it through. As director of the Survey Research Center I was reluctant to have us carry out a survey for political purposes where only one side of an emotionally charged issue would be involved in the design of the questionnaire and the reporting of the results. Moreover, this was a classic case where the sponsor was interested in the survey solely as a referendum—that is, in the marginals—and not in its analytic use for understanding public attitudes and beliefs.

I proposed to the sponsor that we would carry out the survey, provided that someone on the other side of the issue was allowed to review critically drafts of the questionnaire and indicate questions they thought unfair. Somewhat to my surprise, the sponsor agreed, and I obtained the cooperation for this purpose of the president of the Michigan Federation of Humane Societies, who had recently written an article in support of the bill prohibiting pound transfers and whom I will refer to as the critic.

I also agreed to help personally in developing the questionnaire, because of the opportunity it provided to try out this approach. I did so by first talking at length with the sponsor, constructing a draft questionnaire, reviewing the draft with the critic, and then repeating these steps in the course of pretesting and revision. My goal, of course, was to arrive at a set of questions that seemed desirable to the sponsor, reasonably fair to the critic, and adequate from my own standpoint as a relatively neutral observer and from the standpoints of the pretest interviewers. I did not try to bring the two parties to the conflict together in one room because, although each was quite sophisticated, each felt the near-immorality of the other's position too strongly to allow direct collaboration.

This kind of shuttle diplomacy worked well for a number of the questions. In some cases the critic had no serious objection to a question, while in other cases her objection led me to revise the question in a way that the sponsor accepted. In two instances, however, the wishes of the sponsor and the reactions of the critic could not be reconciled even after the most strenuous efforts, and I decided to make use of a split-ballot approach to allow both versions of each question to be asked. We would simply see whether it made any difference if we used the original version or the one preferred by the critic.

In the first case, the sponsor wanted to precede a question with a sentence that the critic felt would distract people from the main moral issue. I didn't personally regard either the question or the disagreement

as very important, and was glad to find that the experiment showed no difference in distribution (or in any other way) between the two versions.

The second question under dispute was much more crucial to the questionnaire. It is also more important to us theoretically because it shows how two sides can phrase questions differently not because of minor variations in wording, but because of fundamental disagreements about the central issue under debate. The question was the first in the questionnaire to deal directly with the release of pound animals to laboratories, and the sponsor wanted the question to indicate that the animals to be released would otherwise be killed by the pound because they had not been claimed as pets. The critic, however, rejected this assumption entirely. Her position was that such arrangements between laboratories and public pounds tend to reduce the motivation by pounds to curb the unwanted animal population in nonlethal ways. In other words, she saw cause and effect to run in the direction opposite to that assumed by the sponsor. I can't go into the details of the argument here, but I was persuaded that both sides were quite sincere, that each argument was at least somewhat plausible, and that it was worth asking the question with and without the disputed assumption to see what difference it made. After all, if the distinction made no difference to the general population, it lost relevance as far as the survey results went.

As it turned out, omitting the statement did lower significantly the support for release of pound animals to laboratories, but the actual percentage drop was small—from 61 percent to 54 percent—and did not change the fact that a majority supported such release with or without the statement. In addition, the experimental variations produced unexpected order effects later in the questionnaire, which also reduced noticeably, but not massively, the evidence for public support of laboratory use of pound animals.

Our final report for the sponsor, which was also given in draft form to the critic for comment, drew two conclusions from the split-ballot experiments. The first was that the changes in wording that we actually performed had some effect, but only a small one, on the results. Readers were alerted to the fact that percentages can change for reasons other than sampling error, but also that in the instances studied experimentally, the changes did not alter the broad conclusions. As you can imagine, the sponsor was reasonably happy with this summary.

At the same time, we noted that the question manipulations we used were not necessarily the most powerful that might have been tried. Other nonexperimental results from the survey strongly suggested that the main way to appeal to the majority of the population on this issue

would be to emphasize that dogs and cats in medical experiments not only die, but do so often with a great deal of suffering. Whether in fact that would be a valid claim was not for the survey investigator as survey investigator to determine, any more than the issues involved in the two split-ballots we actually carried out, but had the key questions included such an assumption, my guess is that the marginals would have been altered more substantially. Thus the critic may have learned something relevant to her position in the longer run.

Looking at this unusual experience from my own standpoint, I draw some additional conclusions. For one thing, I know that both the critic and the sponsor felt that our survey organization had gone to considerable lengths to avoid bias in the questions. It was also clear from their comments that both parties came away with much greater respect for the need for careful questionnaire design, and we in turn realized that by hearing two sides of the issue we had gained substantive understanding that was useful in the question construction process.

I don't claim that this unusual exercise would be easy to carry out in other settings, or with other sponsors or critics. But it does seem worth considering as a possibility when issues are highly controversial, for it can help protect survey research from two of the false beliefs that most bedevil it. One belief, which might be labeled survey fundamentalism, is the naive acceptance of the numbers in a survey report as a literal picture of public opinion, with or without the now conventional footnote about sampling error. The other and equally naive belief—let us label it survey cynicism—is that survey results are worthless because investigators can readily produce whatever numbers they wish by means of clever question wording. We need in every report, academic or commercial, to counter these two widespread beliefs—sometimes put forth, I should note, by the same person at different times depending upon how the results come out! The most effective way to do so is by demonstrations from which both the public and the survey investigator can learn.

Let me end by reiterating that I do not wish to deny the need at times to present univariate distributions, as indeed was the case for the animal experimentation report. Where trends are not possible because an issue is new, or where one is attempting to obtain a fuller picture of public thinking about a complex policy question, it may be essential to present and attempt to interpret response marginals in a direct way, as well as to deal with the data more analytically. A good example was the crisis in 1981 over the Falkland or Malvena Islands—the choice of name itself, of course, reflected a basic, not a trivial, disagreement over wording. It was of paramount interest in the early days of that conflict to learn what degree of public backing the British and Argen-

tine governments had as they moved toward war—yet neither trend nor other relevant comparison data were available to aid in understanding.

In this sense, what I described at the beginning of my talk as the fundamental paradox of survey research—the referendum point of view *versus* the analytic point of view—lies deep within each of us, as well as between the public and the survey investigator. We all wish at times to know how the public as a whole feels about an important issue—whether it is the bombing of Libya or the spread of AIDS in America. Therefore, we need both to remind ourselves and the public of the limitations of single variable results, and at the same time take whatever steps we can to reduce those limitations.