STANDARDIZED SURVEY INTERVIEWING

Minimizing Interviewer-Related Error

Floyd J. Fowler, Jr.
Thomas W. Mangione

Applied Social Research Methods Series
Volume 18

SAGE PUBLICATIONS
The International Professional Publishers
Newbury Park  London  New Delhi
an average of 50 interviews each, such an intraclass correlation would inflate standard errors by 32 percent; obviously the effect on standard errors is greater for questions with higher intraclass correlations.

Looked at from another perspective, Sanders (1962) reported that more than half the variance in the number of health conditions reported in a survey could be attributed to the interviewers.

As an example of results from a study using record checks, Cannell et al. (1977a) found a correlation of .72 between the size of an interviewer's assignment and the percentage of known hospitalizations reported to interviewers in a health survey.

CONCLUSION

Using any of the several approaches to detecting interviewer error described above, there is a great deal of evidence that interviewers are an important source of error in surveys. Although it is easy for research organizations and researchers to ignore interviewer-related error, because it is not apparent without special efforts to detect it, there is no doubt that a lack of standardization among interviewers is responsible for a great deal of survey data being less good than it should be. Moreover, in contrast to some strategies for reducing error in surveys, such as increasing sample sizes, the cost of strategies for reducing interviewer-related error are often surprisingly low. On that note, let us turn to what we know about how to minimize interviewer-related error in surveys.

3

Standardized Interviewing Techniques

Although it is not at all easy to carry out a good, standardized survey in which all interviewers behave consistently, the procedures for interviewers to follow in handling the question-and-answer process in a standardized way are simply stated:

1. Read the questions exactly as worded.
2. If the respondent's answer to the initial question is not a complete and adequate answer, probe for clarification and elaboration in a nondirective way; that is, in a way that does not influence the content of the answers that result.
3. Answers should be recorded without interviewer discretion; the answers recorded should reflect what the respondent says, and they should only reflect what the respondent says.
4. The interviewer communicates a neutral, nonjudgemental stance with respect to the substance of answers. The interviewer should not provide any personal information that might imply any particular values or preferences with respect to topics to be covered in the interview, nor should the interviewer provide any feedback to respondents, positive or negative, with respect to the specific content of the answers they provide.

There are, however, two main obstacles to actually carrying out an interview in a standardized way:

1. An inadequate survey instrument. If the questionnaire is not designed so that it can be administered easily in a standardized way, then it is unlikely that standardized procedures will be followed. This is a subject we will say more about in Chapter 5.
2. Respondents do not understand what is expected of them. The measurement process in a survey interview is a team effort, requiring both participants to play their roles as prescribed. A major reason interviewers have difficulty in performing their job properly is that they are unable or do not know how to train the respondent to make the interview process work. That topic is addressed later in this chapter.

There are two other factors that increase the likelihood that interviewers will not do a good job of being standardized interviewers, particularly if they are having trouble with the questionnaire or the respondent:
Interviewers generally want answers to be accurate, so they have trouble being standardized when the goals of standardization and accuracy seem to be in conflict.

Interviewers like to be personable and responsive to respondents, and they sometimes have trouble being standardized when they feel a conflict between behaving as they are trained and maintaining the kind of relationship they think the respondent wants.

The goal of standardization does not have to conflict with obtaining accurate data or being responsive to the respondent. The last part of this chapter deals with strategies for achieving standardized interviewing without neglecting those other important goals. Table 3.1 provides a summary of the techniques for standardized interviewing and its impediments.

**READING QUESTIONS AS WORDED**

Virtually every interviewer's manual that we have examined has "reading the questions the way they are written" as a basic first principle of good interviewing technique. On the surface, it would appear to be a rule that is both easy to understand and easy to follow. Hence, it may be somewhat surprising to learn that interviewers often do not read questions the way they are written.

In four studies in which interviewer-respondent interactions were coded, the rates at which interviewers changed question wording ranged from 20 to 40 percent. (Bradburn & Sudman, 1979; Cannell, Fowler & Marquis, 1968; Fowler & Mangione, 1986; Cannell and Oksenberg, 1988.) Moreover, it is important to know that these studies were all done in organizations that put more than the average emphasis on methodological rigor. These numbers are probably conservative with respect to the rates at which interviewers actually change question wording.

Why do interviewers change wording? Certainly in some cases the person who wrote the questions bears major responsibility. Interviewers are likely to change wording if a question is hard to read. They also will change a question to provide an emphasis that they think will make it easier for the respondent to grasp the question or what is wanted. Such explanations, however, can only account for a portion of interviewer lapses. In our opinion, the major force pushing interviewers to change question wording is an effort to make the interaction somewhat more conversational and casual. One way they do that is to add their personal touches to the questions. This practice is likely to continue and grow over an interviewer's career unless

**Table 3.1**

<table>
<thead>
<tr>
<th>Techniques for Standardized Interviewing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Read questions as written.</td>
</tr>
<tr>
<td>2. Probe inadequate answers nondirectively.</td>
</tr>
<tr>
<td>3. Record answers without discretion.</td>
</tr>
<tr>
<td>4. Be Interpersonally nonjudgemental regarding substance of answers.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obstacles to Standardized Interviewing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Inadequate survey instrument.</td>
</tr>
<tr>
<td>2. Respondents who do not know how to play their role.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reasons Interviewers Fail to be Standardized</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Goal of accuracy seems to conflict with goal of standardization.</td>
</tr>
<tr>
<td>2. Goal of maintaining rapport appears to conflict with goal of standardization.</td>
</tr>
</tbody>
</table>

Supervisory practices involve monitoring and feedback when interviewers do not read questions exactly as worded. Bradburn and Sudman (1979) found that more experienced interviewers were more casual about the way they read questions than were comparatively new interviewers.

Of course, most of the changes in wording appear to be minor; certainly most interviewers would say they are basically reading the questions the way they are written. The critical issue from the point of view of measurement is whether or not the question wording changes that occur make any difference to the quality of measurement. The answer is, we do not know for sure, although we do know that small differences in the way questions are worded can have a major impact on answers.

One way we looked at this was to see whether questions that were most likely to be misread by interviewers were distinctively likely to have large interviewer effects, as measured by the intraclass correlation. Our finding was negative in this respect. When interviewers are given a hard question to read, the changes that they make do not generally produce significant interviewer-related effects on the data.

A second approach was to see whether interviewers who were distinctively casual about question reading seemed to produce distinctively biased data. Our assessment is restricted to 20 interviewers who tape recorded their interviews, which were then coded. Since each interviewer's sample was a probability subsample of the total, it was meaningful to look at whether the estimates derived from an interviewer's subsample differed from the sample as a whole. We simply counted the number of estimates from each interviewer's sample which we thought were potentially biased; that is, that fell in the biased direction compared to the total sample mean. That count...
STANDARDIZED SURVEY INTERVIEWING

was correlated with various ratings of an interviewer's performance with respect to standardized interviewing skills.

It was found that the ratings of various skills, including reading questions, probing, and recording answers, were intercorrelated, so we cannot pull out the distinctive contribution of good question reading from good probing. With only 20 interviewers involved, the raw correlations did not reach statistical significance, but the direction of association for all of these standardized behaviors, including reading questions, is in the expected direction; that is, interviewers who were rated as showing better interviewing skills, including reading questions, appear to obtain less biased data.

The third approach to answering that question is to look at studies of question wording. Schuman and Presser (1981) report a number of experiments in which small changes in question wording were made on purpose, with the results being compared. Basically, they find that sometimes small changes in wording make a big difference in the distribution of answers; in other cases, apparently large changes in question wording have minimal effect on the answers people give.

Question A: Should Communists be forbidden to speak in public places in the United States?

Question B: Should Communists be allowed to speak in public places in the United States?

It could be argued "forbidden" and "not allowed" are equivalent concepts, but to respondents they are not. For example, when comparable samples were asked the two forms of the question, nearly 50 percent said the U.S. should "not allow" Communists to speak in public, while only about 20 percent said the U.S. should "forbid them to speak." This finding suggests that if an interviewer chose to substitute "forbid" for "not allow" in a question, it would have a marked impact on the data, and would produce clear interviewer-related error. Such a change might seem innocuous and conversational to an interviewer, making the question better, if you will, but it would be an excellent example of why we tell interviewers to read questions as worded.

Interviewers also are instructed carefully to read all the alternatives. It might be tempting to add or delete an alternative that says "or do you have no opinion on that topic." Again Schuman and Presser give us a good example that the alternatives matter. Comparable samples were asked whether they favored or opposed a fictitious Agricultural Trade Act of 1978. One sample was explicitly offered the option, "or do you have no opinion on that?"; the other sample was offered no such option; it was asked only if it favored or opposed the act. The result: 69% of the sample volunteered that they did not know the answer, but 90% chose the "no opinion" option when it was offered.

Whether or not a "no opinion" category is included has a major effect on the distribution of answers. Not reading one category might seem to be a small change that would make a question easier to read, but it produces interviewer-related error.

On the other hand, Schuman and Presser report other experiments where major changes in wording seem to have little effect on answers. For example, substituting the term "abortion" for the term "end a pregnancy" had no effect on answers that people gave.

In conclusion then, the theoretical argument for having interviewers ask questions exactly the way they are written is easy to understand. If the interviewer does not ask questions the way they are written, the researcher does not know for sure what question was posed. However, general instructions to interviewers not to change wording do not suffice. Interviewers do reword questions unless significant efforts are made to keep them from doing it. Moreover, they are likely to increase the practice over time unless that tendency is checked. The motivations for interviewers rewording questions are generally innocuous, or even constructive, trying to make questions clearer, trying to make the interaction with the respondent go more smoothly, in short trying to improve on the work of the researcher. Sometimes the effects of those efforts are innocuous, but sometimes they create substantial differences. Basically, it takes effort and work on the part of the researcher to write questions that can be asked as worded, but it is an effort that must be made if serious measurement is to be achieved.

PROBING INADEQUATE ANSWERS

In an ideal situation, the researcher writes a great question, the interviewer reads it as written, and the respondent provides a complete answer which meets the question objectives. Of course, that does not always happen. If the initial reading of the question does not produce a satisfactory answer, then the interviewer must engage in some kind of behavior to move the process along and reach the desired end point. The interviewer's behavior cannot be completely preprogrammed, because the problem to be solved will vary from situation to situation. However, the goal is to have interviewers handle the problem in a way that is consistent across interviewers and respondents and that does not influence the content of the answers that
In the case above, where the respondent said, "I don't understand," the interviewer did not respond appropriately. The interviewer should have asked, "Can you explain what you mean by "I don't understand"?" This would have allowed the interviewer to understand the respondent's confusion and provide a more appropriate response.

Respondent: The schools around here are very good.

Interviewer: How would you rate your schools—very good, good, fair, or poor?

Respondent: Very good.

Interviewer: I see. Thank you for your time.

The key to understanding the response is to interpret the answer accurately. In this case, the interviewer correctly understood the respondent's answer as "very good." However, had the interviewer misinterpreted the statement, the conversation would have been less effective.

In some cases, the interviewer may need to ask follow-up questions to clarify the respondent's meaning. For example, if a respondent said, "I don't understand," the interviewer should ask, "Can you explain what you mean by "I don't understand"?" This would have allowed the interviewer to understand the respondent's confusion and provide a more appropriate response.

When asking questions, it is important to ensure that the respondent understands the question. Asking a question that is not clear or is difficult for the respondent to understand may lead to inaccurate or misleading answers.
it would be understandable if the interviewer probed something like: "Well, would you say fair or poor?" That is bad practice that will affect the answers. It can be shown that the distribution of answers to a scale that has "good, fair, poor" versus one that has "very good, good, fair, poor" is quite different. Respondents respond to the number of categories and the position of a category on a scale, as well as to the words, when classifying themselves. A truncated version of a set of responses is not the same stimulus, and it will affect the answers.

Probing Numerical Answers

When an answer calls for a numerical response, not in categories, the most common problem faced by the interviewer is one of precision. A respondent may answer with a range or a rounded number, and the interviewer may want to attempt to get the respondent to answer more precisely.

One inappropriate behavior is a directive probe. A directive probe is one that increases the likelihood of one answer over others. There are many different ways that interviewers can create directive probes, but the easy way to recognize one is that it can be answered with a "yes" or "no" response. The reason such probes are called directive is that in essence they suggest a particular answer as a possibility. Respondents are more likely to say "yes" than "no" when asked a question like that. So, any probe that can be answered with a "yes" or "no" is directive. In addition, any probe that lists or mentions some possible answers, but excludes others, is also directive because it increases the likelihood that the mentioned answers will be chosen.

Questions that call for numerical answers often require interviewers to probe for more specific details, and they are amenable to directive probing.

QUESTION: In the last seven nights, how many times have you gotten fewer than eight hours of sleep?
RESPONSE: I usually get eight hours of sleep.
DIRECTIVE PROBE 1: Well, for the last seven nights, would the answer be 0?
DIRECTIVE PROBE 2: Well, for the last seven nights, would the best answer be 0, 1, or 2?
NON DIRECTIVE PROBE 1: In the last seven nights, how many times have you gotten fewer than eight hours of sleep?
NON DIRECTIVE PROBE 2: Well, for the last week, would the best answer be more than 2 times or 2 or fewer times?
FOLLOW-UP NON DIRECTIVE PROBE 2: (If answer is "fewer") Well, for the past seven days, would the best answer be 0, 1 or 2?

The problem with the first two probes, obviously, is that they suggest an answer. The first is the worst, as it is clearly a probe that can be answered with a "yes" or "no." The second probe is less blatant, but the interviewer has already pretty well narrowed the field for the respondent. It would take an act of initiative, which respondents often do not do, to give an answer that is different from 0, 1, or 2.

In this case, the very best probe is to repeat the question, since the real problem with the respondent's answer was that the question was not answered. Repeating the question has the distinct advantage of being the most standardized approach also, since it involves no innovation or question creation on the part of the interviewer.

Nondirective Probe #2 is an acceptable response from the point of view of being nondirective, although it is less standardized. In this case, the interviewer is using a technique called "zeroing in." A reasonable guess is made by the interviewer of the general area in which the answer is likely to be found, and then a question is asked which does not suggest an interviewer expectation that the answer will fall on one or the other side of the cutpoint. Once an answer is obtained to the initial question, follow-up questions can be asked to further narrow the range.

Sometimes "zeroing in" is the only way, or the most efficient way, to get respondents to be more precise in their answers. Our own preference, however, is to explain the value of having the respondent, rather than the interviewer or the researcher, make an estimate about where in a range the accurate answer falls, and then let the respondent, armed with a clarification of his or her role, answer the question as originally posed. In short, as usual, our strategy is to train the respondent and stick with the question as written.

Probing Open-Ended Questions

The hardest probing tasks for interviewers involve those connected with open-ended questions. The interviewer has to make three judgments of any answer that is obtained: Does it answer the question? Is the answer clear? Is the answer complete?

In reviewing manuals for interviewers, we found that there is much greater variation among organizations in their instructions to interviewers about probing than there is about reading questions as worded. Some organizations seem to accept, or even encourage, interviewers to find a variety of conversational ways to get respondents to clarify or elaborate their answers. Our preference is to have interviewers stick with a very small list of probes.
In fact, we train interviewers that in addition to repeating the question, they only need to use three probes:

1. How do you mean that?
2. Tell me more about that.
3. Anything else?

These three probes are easy to remember. They are nondirective. They do not give interviewers any opportunity to innovate in ways that would make their interviews different across respondents or interviewers. Our feeling is that to the extent that organizations encourage innovative probing, any gains that may be realized in variety or conversational interest will be lost in lack of standardization. Moreover, if the creative probes are truly nondirective, they probably amount to one of the three mentioned above.

The interviewer's task is to decide which of those probes is appropriate, and that involves analyzing the respondent's answer. The four probes, including repeating the question, correspond to the four ways in which a respondent's answer can be inadequate:

1. The response can fail to answer the question; it answers some other question. The interviewer should repeat the question.
2. The answer contains unclear concepts or terms that make its meaning ambiguous. The interviewer should probe saying, “How do you mean (that)?”
3. The answer is not detailed enough or specific enough. The interviewer should probe saying, “Could you tell me more about (that)?”
4. A perfectly appropriate answer has been given, but there is a possibility that there are additional points that the respondent could make in answer to the question. The interviewer should ask, “Is there anything else?”

Below are some examples of situations in which these probes would be used.

**QUESTION:** From your point of view, what are the best things about living in this neighborhood?

**COMMENT:** This is one of the hardest kinds of questions for interviewers and respondents to deal with, because the kinds of “things” that count are not specified at all. It is up to the respondent and interviewer to decide what sorts of neighborhood features can appropriately be mentioned, and in how much detail.

**ANSWER 1:** In the last neighborhood in which we lived, it was very transient. People didn’t care about keeping up the neighborhood.

**COMMENT:** The problem with this answer is that it does not answer the question. Although by inference the description of the old neighborhood may be implying something about the characteristics of the present neighborhood, the question asks for a description of the current neighborhood.

**PROBE:** Repeat the question.

**ANSWER 2:** The people.

**COMMENT:** This plausibly could be an answer to this question, but no one could figure out what it means. We need some elaboration.

**PROBE:** Tell me more about that.

**ANSWER 3:** The people are good neighbors.

**COMMENT:** Someone might think that was an adequate answer. It is very hard to tell. One of the problems with this question, as noted, is that it does not give a clue to the interviewer or respondent what kind of answer would satisfy the researcher. Whether or not this is a specific enough answer depends on the question objectives and coding procedures. Nonetheless, what constitutes a “good neighbor” could clearly differ from respondent to respondent, and there is not any information in the answer so far on that topic. A good interviewer would probably probe to find out more about what the respondent meant.

**PROBE:** How do you mean good neighbor?

**ANSWER 4:** They keep to themselves. They leave you alone. You don’t have to worry about being sociable and you don’t have to worry about what they think.

**COMMENT:** This surprise answer shows the value of probing. Based on the initial answer, some might have expected the “good neighbor” answer was going to lead to a description of how warm, friendly, and helpful everyone in the neighborhood was. We now understand what the respondent means by people and good neighbors. The question, however, calls potentially for the respondent to mention more than one “thing” about the neighborhood.

**PROBE:** Okay, I have that down. Anything else?

**COMMENT:** With a question that allows the respondent to make an unlimited number of points, an interviewer should continue to ask “Anything else?,” until the respondent says “no.”

**Probing Don’t Know Answers**

When the respondent answers a question by saying “I don’t know,” it poses a special probing problem to interviewers. “I don’t know” can be a legitimate answer to a knowledge question. It also can mean:

A. It is a respondent response style, a kind of preface to the answer while he or she is thinking about it.
B. The respondent has not thought about the question before, but if he or she thinks about it, an answer may be forthcoming.
C. The respondent knows an answer, but is not sure it is specific or accurate enough for the standards of the researcher.

When a respondent says “don’t know,” the interviewer’s first task is to attempt to diagnose the origin of the problem.

A. If “I don’t know” is considered to be an accurate, thoughtful answer to an information question, the interviewer writes down the answer and goes to the next question.

B. If it is a delaying response style, the interviewer gives the respondent time to think about the answer. The interviewer may want to repeat the question to help the respondent think it through.

C. If the respondent has not thought about the question, the interviewer would encourage the respondent to think about the question, emphasizing that the respondent is uniquely qualified to provide information on the topic. Then repeat the question.

D. If the respondent is not sure about the quality or precision of the answer, the interviewer should be reassuring. There are no right or wrong answers; the questions are designed to get people’s own perceptions and opinions. The respondent’s own best estimate will be better than not having any information at all. Then the interviewer would repeat the question.

Types of Probing Errors

Probing is certainly the hardest of the interviewer skills to learn. Interviewers make two main types of probing errors: probing directly and failing to probe an answer that requires probing.

Initially, researchers were concerned that interviewers would probe directly in order to make results come out the way they wanted them to be. So, for example, they thought that Republican interviewers would probe in ways that would increase the number of answers supporting Republican views. When interviewers are reasonably well trained, that sort of thing does not seem to happen (e.g., Hyman et al., 1954). Mainly, interviewers seem to probe directly when they think they know the answer the respondent wants to give and are having trouble getting the respondent to be explicit.

Directive probing is a strategy for easing the interviewer-respondent interaction. Interviewers find it stressful when they probe an answer several times and still cannot get the respondent to give an answer that meets the question objectives. When the respondent has said enough that both the interviewer and the respondent are fairly sure that the interviewer knows the answer, the easiest thing for the interviewer to do is to say, “I think you want to answer X; is that right?”

The other kind of error interviewers make is failure to probe answers that need to be probed or being inconsistent in choosing which answers they do and do not probe. Three kinds of situations have been found to be particularly prone to interviewer variation in probing.

First, Hyman et al. (1954) found that interviewer expectations affected their probing behavior. Specifically, when they obtained an answer that was consistent with what they expected, based on what they knew about the respondent and other answers given, interviewers tended to accept it without further probing. However, when respondents gave an answer that appeared to interviewers to be inconsistent, they were likely to probe it to make sure they had it right. This is an example of conscientious interviewer behavior that results in handling answers inconsistently.

Interviewers also are likely to be different in the number of answers that they get to questions for which multiple answers are possible. There is interviewer discretion in how often they ask for “anything else?” Some interviewers obtain more answers than others on a consistent basis because they consistently probe for more answers, and that affects the data.

Third, interviewers have been found consistently to differ in the way they handle the “don’t know” response or its equivalent. Some interviewers either work harder or are more successful in getting opinions or answers out of respondents when they say initially that they do not have an answer to the question.

Probing is one part of the question-and-answer process that cannot be completely standardized; if a respondent does not give an adequate answer when the question is first asked, the interviewer has some decisions to make. The response cannot be perfectly programmed. Anytime there is an opportunity for interviewer discretion, there is an opportunity for interviewers to be inconsistent across respondents and across interviewers, and that is when interviewer-related error occurs.

Is probing an important source of error in survey measurement? Absolutely. In Hyman’s early studies, differences in probing were the main error-producing aspects of interviewer behavior. The two problems cited above, the way “don’t know” answers are handled and the number of mentions obtained from respondents, are both characteristics of questions that have been consistently associated with interviewer effects (Groves and Magilavy, 1980). In our own studies, the quality of open-ended probing was related to the bias in answers interviewers obtained at a nearly significant level (p = .07). Moreover, as will be discussed in more detail in Chapter 5, the most important correlate of questions which are prone to interviewer effects is the likelihood that they will require interviewer probing.

As we will argue in Chapter 5, we think the most effective way to try
to minimize probing as a factor in interviewer-related error is to improve the quality of questions. The less interviewers have to probe, the less opportunity they will have to make errors. In addition, we are strong believers in minimizing the variety of probes that interviewers use. The more interviewers use innovation in creating the stimuli to which respondents respond, the more likely they are to be inconsistent and create error. Moreover, interviewers have plenty to do during an interview besides thinking up innovative probes. A good question will not only minimize the need for probing but will also reduce the inconsistency of probes when they are needed.

**RECORDING ANSWERS**

The job of the interviewer is to write down the answer the respondent gives. The key to standardized recording is to have no interviewer judgment, no interviewer summaries, no interviewer effects on what is written down. The rules for standardized recording differ for closed and open-ended questions, and by whether the question asks for the report of factual information or information about opinions and feelings. (See Table 3.2.)

For closed-response questions, the key interviewer task is to get the respondent to choose one answer and then to check or record the answer chosen. The only possible recording error, other than a clerical error, would be for an interviewer to indicate that a response was chosen by the respondent when in fact it was not.

The rule for recording open-ended responses to opinion or attitude questions is equally clear and simple: interviewers should write down the answer verbatim; that is, the interviewer should write down the exact words given by the respondent, without summary or omissions. It has been documented that summaries and paraphrases will vary from interviewer to interviewer (e.g., Hyman et al., 1954). One way to keep interviewers from affecting answers is to reduce interviewer discretion about what to record.

When questions are of a factual nature, whether answers are of the closed-response or open-ended variety, the rules are a little bit different. While the coding of subjective answers is highly dependent on the way things are phrased and the particular points that are or are not made, factual questions generally ask for some specific kind of information. The words that a respondent uses are not deemed critical. The interviewer is expected to write down the information provided by the respondent that was called for in the question, without trying to record the exact words. Moreover, if the question includes a set of response alternatives, and the respondent does not choose one, the best step probably is for the interviewer to write down the relevant information provided by the respondent without necessarily getting the respondent to choose a single category.

One of the important realities about asking factual questions is that sometimes it is hard to anticipate all the possible special circumstances that will be encountered. Moreover, respondents may come up with questions or problems with definitions that are not given in the question but nonetheless are essential to giving an accurate answer. It is not good practice for an interviewer to provide a definition or to give help to one respondent that is not going to be given consistently to others. It is intolerable for an interviewer to make up an idiosyncratic rule about how to handle a special situation that was not anticipated. For a factual question, good practice is to gather all the information needed to answer the question given alternative interpretations of key concepts that affect the answer. Then the researcher can make a consistent coding rule across all respondents.

In an earlier example, we discussed a question about visits to medical doctors. One respondent wanted to know whether a visit to a doctor's office counted when only a nurse was seen for inoculations. If the respondent or interviewer cannot tell from the question what the rule must be, the ideal solution would be to write down all the information needed given either interpretation. Hence, the interviewer might write down: "Four con-
tacts with medical doctor plus a series of ten visits when inoculations were given by a nurse only."

Another common recording challenge is exemplified by questions about how much formal education has been completed. Many educative post-high school experiences do not fall neatly on the typical educational ladder. Suppose the following categories are offered: less than high school graduate, high school graduate, some college, and college graduate. Respondents will report post-high school education in art, music, nursing, auto mechanics, and many other things. It is not feasible to give interviewers (and thereby respondents) rules for which of these do and do not count as college experience. The thing to do is have interviewers write down the exact pattern of education that the respondent has had, then let the coding department handle these problems consistently across all interviewers and respondents.

So for factual questions, the threat to standardized measurement is that interviewers will make arbitrary decisions. Verbatim recording is not that important, but getting all the information down so that the decision rules can be carefully evaluated and applied consistently is the critical step in standardized measurement for such questions.

Our studies of interviewers suggest that they make relatively few errors in recording closed-response answers. The quality of verbatim recording of open-ended opinion questions varies more. It takes interviewer effort to do that well, and interviewers will not do it unless their supervisors insist on it. It is important, though, because interview summaries and paraphrases are not standardized. As noted, Hyman et al. (1954) found that interviewer expectations affected what they wrote down. Interviewers tended to make their recorded answers consistent with their perceptions of the respondent; their paraphrases and summaries left out the contradictions and subtleties in answers.

**BEING INTERPERSONALLY NEUTRAL**

All interviewer manuals encourage interviewers to be interpersonally neutral as part of the standardization process. Sometimes it is difficult to figure out exactly what is meant by that, but we think the proposed behavior includes, if it is not limited to, the following:

1. The interviewer would not volunteer personal information to the respondent about life situations, views, or values. That would be particularly true for any characteristics that might be related to the subject matter of the interview, but most survey organizations want to minimize such conversation altogether for at least three reasons. First, volunteering personal information may undermine the goal of establishing a professional, rather than a personal, relationship in which data gathering is the priority. Second, although interviewers cannot be identical in their observable demographic characteristics, talking about personal situations and views only exacerbates the differences across interviewers. Third, information about personal views and background may directly affect answers. The most serious way would be that a respondent would try to guess which answers would be most valued or preferred by interviewers.

2. During the interview interaction, the interviewer should be careful that the feedback provided to respondents does not imply any evaluation or judgment about the content of the respondent's answers. The goal of the interviewer is to get accurate and complete answers. It is natural for respondents to be concerned about how answers will look to the interviewer. The interviewer should be careful not to feed into that process by casual interpersonal behavior.

This aspect of standardized interviewing is something at which interviewers are pretty good if they have at least a little bit of training. Untrained interviewers are not good at it. In our studies, over a third of interviewers with minimal training were rated in need of improvement in the interpersonal area, but over 85 percent of interviews done by interviewers with more than minimal training were judged to be "satisfactory" or better in managing the interpersonal side of things. (See Chapter 7).

These data are consistent with Hyman's studies. He was concerned that interviewers' personal views would be communicated to respondents. He found that respondents usually reported they had no idea about interviewers' opinions. When they did think they knew where interviewers stood, those opinions often were not accurate. Significantly, when respondents thought they knew the interviewer's opinion, they almost always thought interviewers agreed with their views regardless of the interviewer's actual opinions.

In our observations of interview interactions, blatant evaluative comments are quite rare. In training, we tell stories about the interviewer who asked the respondent how much he drinks. When he says he has six drinks a day, the interviewer responds, "Oh my gosh, that's awful." Such events really do not happen in a reasonably well run survey.

What is harder to deal with are the subtle kinds of feedback processes. For instance, a respondent says, "I haven't had to go to the doctor in over a year now. I guess I have been pretty lucky." To that an interviewer might respond, "Isn't that great."

Now what's so bad about that? Don't we all share a general humanitarian wish for universal good health? Sure. However, we do not want that respondent to get the idea that this interviewer will think less of him/her or be unhappy or somehow care if, when the next question comes along, some
deviation from perfect health has to be reported. In this instant relationship in which respondents are looking for clues about how to do it right, subtle, seemingly innocent expressions of pleasure that a respondent is healthy or has taken good care of his/her health can lead to effects on the data.

How important is it? It is hard to say for sure. However, Marquis, Cannell and Laurent (1972) found clear evidence that subtle reinforcement of respondent behavior had a significant effect on the number of health conditions and visits to doctors reported. Moreover, in our studies, among tape recorded interviewers, there was a significant relationship between the rating of inappropriate feedback to respondents and the likelihood that interviewers obtained answers that appeared to be biased.

Thus, the evidence is that interviewers can be successful at avoiding blatantly biasing evaluative behavior; most reasonably well run surveys will not have that problem. On the other hand, there is a considerable amount of more subtle interaction that goes on between interviewers and respondents that probably affects some answers.

**TRAINING THE RESPONDENT**

In our studies, we have come to believe that one of the most important things an interviewer can do to carry out a standardized interview is to train the respondent. Although a main source of problems for interviewers can be a poorly constructed survey instrument, the real problem comes when the interviewer begins to feel awkward because of the way the interview is proceeding and the rules under which the participants are operating. One reaction of interviewers is to bend the rules of standardization in order to appear to be responsive to the respondent and to make the respondent more comfortable. We say a better solution is to explain what is going on to the respondent. First, it introduces the fact that this is a specialized interaction with a special set of rules. It is not like most other interactions the respondent has been in, including other interviews. It also legitimizes the notion that the interviewer may elaborate or explain further rules as this game progresses.

Second, it tells the respondent in advance what the interviewer is going to do, which will make it easier to do it. Also, once an interviewer has told a respondent that questions will be read exactly as worded, it will make it harder for the interviewer not to do it; it may increase standardization for that reason as well.

Once the interview begins, we believe an interviewer should stop the question-and-answer process every time the respondent fails to perform his role appropriately and explain the rules and why the rules matter. The following are among the most common issues:

**PROBLEM:** The respondent has partially answered, or even fully answered, a question that has not yet been asked. The interviewer feels awkward about reading the next question, since it will appear that the respondent's earlier answer was not heard.

**INTERVIEWER:** "The next question is one you have already dealt with to some extent. However, the way the interview works is that I need to have you answer each question specifically, so that we can compare the answers you give with the answers everyone else gives. Also, sometimes we find the answer is different to a specific question, even though it seems that the question has been answered before. So, let me read the question as it is worded here, and I would like you to give me the answer to make sure we have it right."

**PROBLEM:** A question contains a term that the respondent finds ambiguous or not well defined, and the question wording does not provide what the respondent considers to be an adequate definition.

**INTERVIEWER:** "I see what your problem is with the question. Even though these questions are carefully tested, sometimes we have one that is not quite
clear to some people, or which doesn’t quite fit everybody’s situation. Again, though, the way a survey works, we need people’s best answers to the questions as they are written. That way we can compare your answers with other people’s. If we change the question for each respondent, we wouldn’t be able to analyze the answers. Let me read the question again, and you give me the best, most accurate answer you can, given the way it is written.”

PROBLEM: Respondent does not want to choose one of the response alternatives in a closed-ended question.

INTERVIEWER: “With this kind of question, answers are analyzed according to which of these alternatives people choose. I need to have you choose one of these specific answers so that we can compare your response with those that others give. We know that in some cases none of the answers will fit the way you feel exactly; but other people will have that problem, too. The important thing is that we keep the question-and-answer process consistent across everybody, so we can see similarities and differences in the answers people give.”

PROBLEM: Respondent will not give an answer that is specific enough, because it would only be an estimate or a guess.

INTERVIEWER: “Well we would like it if you would make your very best estimate. Even though it may not be exactly right, no one is in a better position than you are to make this estimate. Just do the best you can.”

PROBLEM: Respondent is speaking so fast, and the interviewer is having trouble recording verbatim.

INTERVIEWER: “I have to write down your answer exactly as you give it, so that it is accurate. If I summarize, I might not get it right. For questions like this, it would help if you would speak slowly, and I may ask you to repeat some parts so that I can get it all down without making a mistake or leaving anything out.

PROBLEM: A family member wants to help a respondent form the answer to an opinion question.

INTERVIEWER: “On factual questions, things like how many times you’ve seen a doctor or been in the hospital, it is fine for you to get help from anyone who can be helpful, because we want the most accurate information we can get. However, when we ask for somebody’s feelings or opinions, there really is no one except you who can give us that answer. Again, it is a matter of being consistent across everybody. When we are asking how somebody feels or what they think, that person alone has to give us the answer that seems to fit best. Although lots of us know others very well, we don’t think that anyone else can accurately tell us what someone thinks or feels. Therefore, to be consistent, we make sure that people answer those kinds of questions for themselves.

PROBLEM: The respondent asks the interviewer for an opinion during the interview.

INTERVIEWER: “I’ll talk about anything you want after the interview, but not before it is over. The reason is that we have found that in some cases when interviewers give their opinions and ideas during an interview, we influence the answers we get. This whole interview process is set up so that the only thing that influences the answers is your situation and what you have to say.”

We could extend this list. The details are not important. What is important is the process. If an interviewer is trying to conduct a standardized interview and the respondent does not know how to play the role, the result will be an awkward interaction in which the interviewer will frequently be forced to choose between standardization and being responsive to the respondent. It is true that in order to train the respondent effectively, the interviewer has to be well informed; the interviewer has to know the rules and have a reasonable idea of the reasons for them. However, that level of knowledge is not difficult to achieve. Most interviewers receive these explanations in the course of their training. The trick is to make sure they use them in the right way during the interview process.

CONCLUSION

Carrying out an interview in a standardized way is a difficult task. From the point of view of skills, there is no doubt that probing is the hardest thing for interviewers to do in a consistent, nondirective way. It is the area in which interviewers are most likely to fall short of reasonable standards, and also the area from which interviewer effects are most likely to emanate. Recording open-ended answers to opinion questions is also very hard to do well, and also is a source of error, though the decline of the use of open-ended questions in standardized surveys makes that somewhat less of a problem than in the past.

Interviewers also do not read questions as worded; they like to make changes in questions. As a source of error in surveys, reading errors probably are less important than probing errors. However, the rates at which interviewers have been found to be writing their own questions in published studies does not make one sanguine about that aspect of standardization. Moreover, it is likely that in more casually operated surveys, the rates at which interviewers are changing question wording are even more severe,
with more severe consequences, than those reported in the literature. Although it is hard to document the exact extent of that, it may be worth repeating that the bedrock of standardized measurement is that we know what questions people are being asked.

Being neutral is something that interviewers seem to understand and take to readily. Again, the perspective in this book may be slanted by the fact that people who study interviewers tend to work with well organized interviewing staffs where clear standards are emphasized. The ease with which interviewers can slip into interpersonal behavior that implies judgment and evaluation leads one to suspect that this, too, may be more of a problem than is reflected in published data.

Finally, one of the most important contributions we have to make to improving the standardization of interviewing is to emphasize the importance of training the respondent. Based on our observations, a main reason for a breakdown in standardization is that respondents fail to cooperate or fail to understand the process. Consider the task of trying to play chess with someone who thought he or she was playing a slight variation on checkers. There would be a general idea about what was supposed to happen, but confusion about the detailed rules would make it an extremely frustrating experience. In a way, survey respondents are in a similar situation. They have a general idea about interviews, but they do not have a clear understanding about how a specific standardized interview should work. Telling them that the rules are different, and then briefing them on the details of the rules as they become relevant in the course of the game, only makes sense, and it makes for a much better game.

4

Establishing the Context for Standardized Interviews

A precondition to successfully carry out a standardized interview is to establish relationships in which the respondents are willing, and the interviewers able, to carry out their respective roles in the measurement process. Most thinking and writing about interviewing has focused on the quality of the question-and-answer process and how that is managed. Additionally, the important contribution of Charles Cannell and his associates has been to focus attention on the importance of the context in which that question-and-answer process occurs and to point to the critical role that interviewers play in setting up the interview as a measurement experience. There is no doubt that the way the question-and-answer process is handled plays an important role in the quality of measurement, as discussed in Chapter 3. However, there also is no doubt that the way the interviewer sets up the interview interaction, the relationship with the respondent within which the question-and-answer process occurs, also plays a critical role in the quality of the data that result. Moreover, that responsibility rests heavily with the interviewer. It is a very demanding part of the interviewer’s job, and it is one that has been found to be done inconsistently by interviewers.

In this chapter we discuss three important aspects of setting up the interviewer-respondent interaction.

1. Explaining the purpose or reason for the interview. Respondents must have some reason for contributing their time for an interview, and there must be some sense of what will be accomplished by so doing. In the first section of this chapter, we discuss what is known about why respondents give interviews and what difference their motivations make for the quality of data that result.

2. Establishing the tone of the relationship between the interviewer and respondent. On the one hand, the relationship needs to be positive enough so that the respondent wants to give the interview and be cooperative. On the other hand, it also has to be the kind of relationship in which answering the questions accurately and completely is seen as appropriate and desirable. In the second part of the chapter we will discuss what we know about how to formulate that relationship and how that affects the quality of measurement.
On the surface, it seems quite plausible that the concern of a question would affect the exam to which interviewers influence the answers. Two.

QUESTION CONTENT

mean in order to obtain an adequate answer.

Those questions for which interviewers must exert discretion and judge.

Four and up to this point: Interviewer effects are likely to be higher for

the research literature. Our answer will not be surprising. Given what we

we examined a number of issues that have been explored in

interviewers. We examined a number of issues that have been explored in

problems that are at the high and of the distribution, that are most affected by

The focus of this chapter is to describe the characteristics of these answers.

By interviewers than others.

Now very clear we are in the area that some questions are more affected

our distribution. At the moment, we are less interested in this distinction

and ability of the survey instrument but all were basically similar to

stress to study design. Design in the quality of the interview and the com-

studies of Groves and Morgenstern (1986). These distributions varied

interviewer effects. Those whose position similar distribution from these

interviewers affect the answer. Perhaps 25% drastically interviewers,

In any survey, the questions will exhibit a range in the exam to which

with regard to whether the interviewers obtain. In this chapter, we discuss the options available to researchers to increase

We focus on improving the questionnaire itself because the better the qu-

yield data if they were part of their standard operating procedures.

Feedback is that most survey respondents would be able to produce better

Going forward, ways to help the interviewers set standards. On going

standardized instructions, asking for verbal commitments to part-

The Role of Question Design

One of the most important ways for a researcher to ensure standardized

in Standardized Interviewing
Table 5.1. Cumulative Distribution of Rho for 130 Survey Items

Aspects of question content seem particularly plausible: the sensitivity of a question and whether a question asks about attitudes and opinions or about facts.

It is well documented that material that is likely to be sensitive or embarrassing tends to be underreported in surveys. Locander et al. (1976), for example, showed that bankruptcy and arrest for drunken driving were reported at very low rates, despite the fact that they were unlikely to be forgotten. Cannell et al. (1965a, 1977) showed that hospitalizations associated with diagnoses that were likely to be threatening or embarrassing were reported at lower rates than others. Our issue, however, is somewhat different. It is not how well such events are reported in the absolute but, rather, the extent to which the measurement is consistent across interviewers. It does seem reasonable to think that some interviewers would be more successful than others in establishing a kind of relationship with a respondent in which sensitive material would be reported. However, there is very little, if any, empirical evidence to support that hypothesis.

The most support we have seen is reported by Sudman, Bradburn, Blair, and Stocking (1977). They found that interviewer expectations about whether a question would be troublesome or sensitive to respondents was related to the likelihood that they would get an adequate answer to the question. Interviewers who thought a question would be sensitive to respondents were less likely to get an answer at all. With respect to the substantive answers that were obtained, however, Sudman did not find significant interviewer effects distinctively associated with the questions he (or the interviewers) thought were likely to be more sensitive or embarrassing.

In our own research, we coded a series of questions by how likely it was that the answers would be sensitive and embarrassing and looked at the relationship to intraclass correlations. We found no significant association with the coded sensitivity of the questions; in fact, if anything, the intraclass correlations were lower for items rated as sensitive. So, while the theory seems compelling, there is no research evidence to support the notion that sensitive material is harder to measure in a standardized way than nonsensitive items.

Kish (1962) explored the hypothesis that attitudinal or opinion questions were more likely to be influenced by interviewers than questions about factual material. The argument is that questions about facts (past events or current situations that can be independently verified) are rooted in some kind of objective reality. Questions about attitudes and opinions, in contrast, cannot be verified or observed by anyone other than the respondent. To the extent that someone thinks that the source of interviewer error is either that interviewers influence respondents to change answers or that respondents alter their answers to fit interviewer expectations, it is a sensible hypothesis that answers about attitudes and opinions might be more susceptible to such changes than answers about objective events.

As plausible as that hypothesis seems, again, the data are lacking to support that theory. Kish’s data did not indicate any differences between attitudinal and factual data from the point of view of interviewer effects. Our own research also failed to find a significant difference in the extent to which interviewers affected attitudinal and factual answers. Overall, we found factual questions were slightly more affected by interviewers than opinion questions.

That generalization was not true for all classes of opinion items, however. Opinion items that were rated especially difficult to answer, generally because respondents were unlikely to have well developed opinions on the topic, did turn out to be significantly affected by interviewers. However, we are confident that the explanation lies less with the content than with other characteristics of those questions which will be discussed below.
Overall, there is nothing in the research literature to support the generalization that the subject of a question has any consistent bearing on whether or not the researcher can design a standardized measurement procedure.

**QUESTION FORM**

One of the most obvious differences in question form is whether the respondent is given a set of answers from which to choose (a closed or fixed response question) or is asked to answer in his or her own words (an open-ended question). The latter type of question offers more potential for interviewer effect both because of possible ambiguity of what kind of answer will suffice and because interviewers have a more difficult recording task.

There are several kinds of interviewer-related error reported in the research literature that pertain primarily to open-ended questions. For example, Hyman et al. (1954) reported that interviewers tended to selectively probe ambiguous answers based on their perceptions of the respondent's "real position"; they also found errors in recording answers reflecting a tendency for interviewers to leave out points that were inconsistent with their sense of the main position of the respondent. Neither of these errors is pertinent to fixed-response questions.

When looking at intraclass correlations, Groves and Magilavy (1980) report that the number of codable answers given, including the likelihood of giving no codable answer at all, was highly correlated with the interviewer. Again, that only is an issue for open-ended questions.

In our own research, however, when we compared intraclass correlations for large samples of open versus fixed-response questions, we found no statistically significant difference. Although there was some evidence that open-ended questions were more likely to be problematic, it was not a statistically stable trend.

**THE QUESTION-AND-ANSWER PROCESS**

After concluding that we could not produce generalizations about the quality of questions as standardized measures simply from observing their characteristics, we decided we had to study the interaction between interviewer and respondent in more detail. We took 100 tape recorded interviews done by 57 interviewers and developed a detailed code about what happened between the interviewer and respondent in the process of trying to get an answer to the question. Among other things, we counted the frequency with which eight different behaviors occurred that might have some bearing on the measurement process:

1. whether or not the question was read exactly as worded;
2. whether or not the interviewer correctly used a probe to obtain an adequate answer;
3. whether or not an interviewer used a directive probe;
4. whether or not an interviewer failed to probe when a probe was needed to get an adequate answer;
5. (for open-ended questions only) whether or not the interviewer's recording of the answer was an accurate verbatim recording;
6. whether or not an interviewer gave the respondent inappropriate feedback to an answer;
7. whether or not the interviewer engaged in inappropriate interpersonal behavior; and
8. whether or not there was laughter during the question and answer process.

The incidence of each of these behaviors was counted, with careful check coding, for each interaction in the interview. We had previously calculated which questions were and were not subject to significant interviewer effects, as measured by the intraclass correlation coefficient. We then analyzed the interviewer behaviors we had coded to see whether or not certain kinds of behavior could differentiate between the questions that interviewers did influence and those that they did not. These findings are shown in Table 5.1.

Inappropriate feedback or interpersonal behavior did not occur at a high rate per question, and that may explain why it did not show up as a correlate of individual question results. This does not mean that interpersonal behavior does not affect the quality of the data, but among trained interviewers such behavior does not explain why certain individual questions are more or less likely to be affected by interviewers.

We thought that questions that were frequently reworded or misread by interviewers would be those most affected by interviewers. This was not the case. Across all questions, whether or not interviewers read the questions exactly as worded was not correlated with the size of interviewer effect.

We did find other aspects of the interaction that were correlated with interviewer effects. The major correlate of interviewer effects was the extent to which a question required probing. All three of the probing counts, the number of correct probes, the number of directive probes, and the number of occasions when the interviewer failed to probe, were associated
with the size of the intraclass correlation. Basically when a question routinely requires probing, it produces an opportunity for interviewers to make mistakes. Every time a question requires a probe, it gives the interviewer a chance to use a directive probe rather than a nondirective probe; it gives the interviewer an opportunity to fail to probe an answer that requires probing. Moreover, such decisions affect the data in ways that can be associated with the interviewer.

The other correlate of the interviewer-related intraclass correlation was the likelihood that interviewers would make a recording error. Although such errors were considerably less common across all survey items than probing errors, our findings were consistent with Hyman et al.’s (1954) conclusion that selective recording is a significant contributor to nonstandardized measurement across interviewers.

We have reviewed the research literature and looked in more detail at the interviewer-respondent interactions in our own work to see if we could understand the kinds of question features that required probing, and whether or not there were specific kinds of situations that were distinctively likely to produce interviewer effects. We consider this an important area for ongoing research and do not think our answers are definitive yet. However, we have identified some question characteristics that appear likely to produce interviewer effects on data.

**CORRELATES OF PROBING**

Although our studies of question characteristics did not lead us to identify any broad classifications of questions that routinely are more affected by interviewers than others, there are some generalizations about question types that are more likely to require probing, which in turn are more likely to be subject to interviewer effects (see Table 5.2). Specifically, open-ended questions are much more likely to require probing than are closed-ended questions. There also is a tendency for questions about opinions, rather than facts, and questions that were judged by raters to be more difficult than average for respondents to answer (either because they involved a potentially difficult recall task or required formulating an opinion about something respondents were unlikely to have thought about before) to be more likely to require probing. Also, difficult questions and questions about opinions were likely to elicit directive probes.

Hence, our generalization is not that any one of these types of questions (difficult, open-ended, or opinion) is in and of itself distinctively suscepti-

### Table 5.1

<table>
<thead>
<tr>
<th>Interviewer Behavior</th>
<th>Correlation with Rho</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laughing</td>
<td>-</td>
</tr>
<tr>
<td>Incorrect reading of question</td>
<td>-</td>
</tr>
<tr>
<td>Correct probe</td>
<td>.23</td>
</tr>
<tr>
<td>Directive probe</td>
<td>.20</td>
</tr>
<tr>
<td>Failed to probe</td>
<td>.49</td>
</tr>
<tr>
<td>Inappropriate feedback on answer</td>
<td>-</td>
</tr>
<tr>
<td>Inappropriate interpersonal behavior</td>
<td>-</td>
</tr>
<tr>
<td>Incomplete or inaccurate verbatim recording on open q's</td>
<td>.39</td>
</tr>
</tbody>
</table>

**SPECIFIC THREATS TO STANDARDIZED MEASUREMENT**

Identifying questions that routinely require probing can be done with reliability by tape recording and coding interviewer-respondent interactions.

### Table 5.2

<table>
<thead>
<tr>
<th>Interviewer Behavior</th>
<th>Question Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Difficult</td>
</tr>
<tr>
<td>Laughing</td>
<td>.34</td>
</tr>
<tr>
<td>Incorrect reading of question</td>
<td>.52</td>
</tr>
<tr>
<td>Correct probe</td>
<td>.34</td>
</tr>
<tr>
<td>Directive probe</td>
<td>.59</td>
</tr>
<tr>
<td>Failed to probe</td>
<td>.28</td>
</tr>
<tr>
<td>Inappropriate feedback on answer</td>
<td>.29</td>
</tr>
<tr>
<td>Inappropriate interpersonal behavior</td>
<td>.28</td>
</tr>
<tr>
<td>Incomplete or inaccurate verbatim recording on open q's</td>
<td>.28</td>
</tr>
</tbody>
</table>

**THE ROLE OF QUESTION DESIGN**

**Table 5.1**

<table>
<thead>
<tr>
<th>Significant Correlations Between the Incidence of Specific Interviewer Behaviors and Rho</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviewer Behavior</td>
</tr>
<tr>
<td>Laughing</td>
</tr>
<tr>
<td>Incorrect reading of question</td>
</tr>
<tr>
<td>Correct probe</td>
</tr>
<tr>
<td>Directive probe</td>
</tr>
<tr>
<td>Failed to probe</td>
</tr>
<tr>
<td>Inappropriate feedback on answer</td>
</tr>
<tr>
<td>Inappropriate interpersonal behavior</td>
</tr>
<tr>
<td>Incomplete or inaccurate verbatim recording on open q's</td>
</tr>
</tbody>
</table>

**Table 5.2**

<table>
<thead>
<tr>
<th>Significant Correlations Between the Incidence of Specific Interviewer Behaviors and Question Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviewer Behavior</td>
</tr>
<tr>
<td>Laughing</td>
</tr>
<tr>
<td>Incorrect reading of question</td>
</tr>
<tr>
<td>Correct probe</td>
</tr>
<tr>
<td>Directive probe</td>
</tr>
<tr>
<td>Failed to probe</td>
</tr>
<tr>
<td>Inappropriate feedback on answer</td>
</tr>
<tr>
<td>Inappropriate interpersonal behavior</td>
</tr>
<tr>
<td>Incomplete or inaccurate verbatim recording on open q's</td>
</tr>
</tbody>
</table>

**NOTE:** Cells with dashes showed nonsignificant correlations, p < .05. Correlations are based on coding of behaviors while asking 65 items for 100 different interviews. Degrees of freedom = 64.
However, we have also identified at least five kinds of question properties that may be identified by simple pretest procedures that have been found by us or others to increase the likelihood of significant interviewer effects. These are:

1. Number of points mentioned;
2. The hidden screening question;
3. Unclear terms in question;
4. Unclear response expectations; and
5. Field coded answers.

Number of Points

Most often, it is an open-ended opinion question in which the number of different responses given by a respondent affects the data.

Example: What are the things you like best about living in this neighborhood?

The question permits the respondent to give multiple answers. The fact that there is a plural word, "things," would indicate to most interviewers that they should probe at least once to try to get a second response. However, the question is unclear to respondents and interviewers about how many answers or points is enough. As noted previously, Groves and Magilavy (1980) reported that the number of different points made by respondents was one of the characteristics of answers that is most affected by interviewers. Anytime the researcher does not specify what is expected, there will be inconsistency across interviewers.

The Hidden Screening Question

The following are examples of hidden screening questions:

1. What is the best book you read last year?
2. For whom do you think you will be voting in the next election—Harry Truman or Thomas Dewey?
3. What is the best thing that happened to you yesterday, the event that pleased you most?

All three of the above questions, and many other questions that are asked by survey researchers, include at least two questions at once. One of the questions is a hidden question about whether or not the main question actually applies to the respondent.

Analysis: The first question assumes that the respondent has read at least one book (but actually to answer the comparative question, two or more books would be better) during the past year. The respondent might be able to say that the question does not make any sense, and does not apply, because no books were read. A respondent, further, might be able to reject answering the question because none of the books read were liked. Although in theory one can choose a favorite book among disliked books, there is reason to wonder if the question applies in those circumstances.

The voting question includes hidden questions about whether or not the respondent plans to vote, and probably more importantly, whether or not the respondent has made a choice or has an idea about which candidate to support. That one question is really a series of three questions which include whether or not the respondent is planning to vote, whether or not he has a current preference between the two candidates, and then what that preference is.

The question about what was the best thing that happened yesterday is technically a correct question: No matter how terrible things were yesterday, there must have been a best thing. However, if a respondent does not think anything happened that pleased them, then he may want to give the answer "nothing." Again, interviewers and respondents may actually treat the one question as two questions: "Was there anything that happened yesterday that pleased you or gave you pleasure?" Then: "What was it?"

These issues are variations on the "don't know" problem. For almost any question, there is the possibility that a respondent will say he does not have the information or has not given the question enough thought to give an appropriate answer.

From a standardization point of view, of course, the problem is that there are no rules or procedures written into the questions to handle the hidden question. The hidden question is not asked explicitly. Interviewers have to decide how to handle the situation when the respondent volunteers an answer to the hidden question, the one that was not explicitly asked but was assumed, which means that the main question does not apply. As is often the case when the procedures are not spelled out, interviewers handle the situation inconsistently. Studies have shown that the rates at which interviewers accept answers such as "don't know" or "does not apply" vary markedly. In addition, Schuman and Presser (1981) show that the handling of these "don't know" answers has a substantive effect on the distribution of answers.
Poorly Defined Terms

In our research, we thought questions that included terms that were not well defined would be distinctively prone to interviewer effects. We had raters code questions as to whether or not they included “ambiguous” terms. Although we did not have as large a sample of such questions as we would have liked, we found no evidence of a relationship between a question being so rated and the size of the interviewer associated intraclass correlation.

Nonetheless, when we studied interactions between interviewers and respondents in detail, it was clear that some significant interviewer effects stemmed from nonstandardized conversations about the meanings of key terms. We have since come to conclude that only certain types of ambiguities cause problems for standardization.

A key observation was that ambiguous terms did not consistently lead to probing. Some terms and questions were ambiguous, their meaning unclear, but they were not problematic for respondents. For example, respondents were asked how many days they had eaten breakfast in the preceding seven days. We are absolutely certain that respondents differed in what they considered breakfast to be. However, “breakfast” is a term that respondents think they understand, whether or not they define it the same as other respondents or as the researcher. Hence, while it was not a clear term from a measurement perspective, this question did not lead to unstandardized interactions between interviewers and respondents.

We think that a question that includes ambiguities that can appropriately be dealt with by simply repeating the question will probably produce a standardized measurement process. Of course, all questions will be better if they are consistently understood across respondents and if the respondents and researcher share the same view of what the question means. Questions produce response error if they are not clear, but the error is not related to the interviewer unless the question stimulates interaction between the respondent and interviewer, which in turn provides the interviewer with an opportunity to be inconsistent. If interviewers know how to behave and what to do when asked a question about a question, there will not be inconsistency.

The problems occur, then, when an appropriate interviewer response is not simply to reread the question or not to say, in response to a question about the meaning of a term, “whatever it means to you.”

Below are two examples of questions from our research which show high interviewer effects that we think suffered from this problem. The ambiguous term is highlighted with italics.

1. How long ago was the last time you were actually seen by a doctor for your health—within the last month, 1 to 6 months ago, 6 months to a year ago, or more than a year ago?
2. In the past 12 months did you have eczema or psoriasis?

In question 1, the ambiguity is what counts as a visit to a doctor; in question 2, the meaning of eczema or psoriasis is not familiar to all respondents.

In each case, the question wording itself does not deal with the uncertainty, yet it clearly does not make sense to ignore a respondent plea for clarification. The result was a nonstandardized interaction between interviewers and respondents on some occasions, which led to high interviewer effects.

Unclear Response Expectations

A particularly important kind of ambiguity associated with interviewer effects is when the question wording does not specify what constitutes an adequate response. In one sense, a special case of this is the issue discussed above, the number of responses required or expected to an open-ended question. However, as we studied question interactions, we found other instances of this same kind of phenomenon that are perhaps more general.

“From what source would you say you get the most information about health and what you should do to keep healthy?”

Although on the surface that may seem to be a reasonable question, it turned out to pose two problems for respondents and interviewers. First, there was some concern about what constitutes a “source.” There are processes whereby people acquire information, such as talking with people or reading or watching television. Do those constitute sources, or are they only conduits of information from some other source?

Then there is the related question of how specific one needs to be. For example, is “from reading” an adequate answer. If not, is greater specificity required about the sort of thing one is reading (magazines, books, newspapers) or is the source the issue (articles in the New England Journal of Medicine, Ann Landers, my horoscope).

There is absolutely no help in the question wording to deal with these issues. Consequently, the opportunities for interviewer discretion, and hence inconsistency, are abundant. First, interviewers get to decide whether or not an answer is detailed and specific enough to be adequate. If they decide not, they then have no real tools to help the respondent; so they get to write their own questions to solve the problem.
Field Coded Questions

We have a real bias against field coded questions, that is, questions which are asked in an open-ended format, but which the interviewer is asked to record by checking boxes. Such questions require interviewers to be coders. Interviewers are not chosen nor trained to be coders; they have lots of other things to do in the interview besides think about coding; and, worst of all, their coding cannot be check coded to detect errors or find out how bad they are at it.

Our own studies included only a few questions that were field coded. On average, they were not significantly worse with respect to interviewer effects than others. However, two of the worst questions in our study were field coded. One was taken verbatim from the National Health Interview Survey. Why did you go to the doctor the last time you went? Interviewers were supposed to record answers in one of the following categories:

1. Diagnosis or treatment
2. Check up
3. Pre- or postnatal exam
4. Eye exam
5. Immunization
6. Other (specify)

Another field coded question we made up ourselves. In what area of your life would you most like to see some changes? Interviewers were supposed to record answers in one of the following categories:

1. Job
2. Children
3. Finances, money
4. Health
5. Leisure
6. Other (specify)

Not only were the intraclass correlations high for these items, our review of interviewer-respondent interactions show they produced some of the worst (that is, least standardized, most directive) interviewing behavior that we saw in our survey.

The reason, of course, is that we put the interviewer in an absolutely impossible situation. On the one hand, these two questions happen to be distinctively nonspecific about what kind of answers constitute an adequate answer. "Why" questions are always problematic. Consider these possible answers to the question about why did you go to the doctor. "My husband said I should go" or "Because I believe in doctors." Although obviously many respondents figure out in context that the idea is to answer about the medical reason or problem for going to the doctor, there is nothing in the specific wording of the question that helps them get there.

However, the real problem is that the respondent has no idea what distinctions are sitting in front of the interviewer, and the interviewer has been given no standardized strategy for giving clues to the respondent. Respondents give answers that are relatively close to a response category, but not unambiguously so. The interviewer is fairly sure where the answer should go, but interviewers differ in how strictly they adhere to nondirective probing techniques to get the information needed to classify the answer. Moreover, reviewers could hear on the tapes how frustrating it was to interviewers not to be able to get a definitive answer. In the end, some of them would simply say: "So would you say that was a check up?"

There is not clear empirical evidence that field coding is a terrible thing to do under all circumstances. Only under some circumstances. Specifically, field coding may be okay if the kind of answer that meets the question objectives, both in terms of its character and its specificity, is clear from the question that is asked. However, it is a bad way to design questions if the respondent cannot discern from the question what constitutes the range of acceptable answers.

It may be that an open-ended question that does not specify well what answers will meet the question objectives will generate interviewer effects whether it is field coded or not. It may be that other aspects of the question, rather than the field coding itself, lead to interviewer effects. The quality of the coding categories—how clear, how nonoverlapping, how well they fit the range of answers given, how they are formulated—may also affect the interviewers' performance. Nonetheless, we continue to be very conservative about the use of field coding when designing standardized interview schedules.

DESIGNING SURVEY INSTRUMENTS TO MINIMIZE INTERVIEWER EFFECTS

Designing questions to minimize interviewer effects boils down to building into the question itself what the respondent needs to know in order
to answer the question. The idea is to write a question that provides a complete script for the interviewer, everything the interviewer will have to say, so that no idiosyncratic choices about how to clarify the question or how to probe the answers will be needed.

That may not sound profound or insightful. Possibly readers of Stanley Payne's book (1951) would have thought those principles were reasonably stated then. However, significant numbers of survey questions do not meet those standards, and, hence, are handled inconsistently by interviewers. Moreover, in our experience in working with colleagues about question design, we think there is far too little appreciation of the importance of minimizing the need for probing and anticipating within the question wording what kind of information will be needed to achieve question objectives.

Our discussion of question problems suggests some guidelines for improved question design. First, building in good definitions of key terms, particularly when counts of certain classes of events are needed, is sound question design. Interviewers and respondents should not have to cope with questions that lack needed definitions.

Second, there is no doubt that being attentive to hidden screening questions and making them explicit will produce better, more standardized measurement. Although Schuman and Presser (1981) do not explicitly advocate the screening question as better, their data seem compelling to us. Rather than have respondents or interviewers decide whether or not a question applies in their own inconsistent ways, it is better measurement to set up a consistent process for having respondents decide whether or not the question should apply to them. Such design reduces error across respondents.

Third, researchers can reduce interviewer-related error by writing questions that make the response expectations clear. One obvious way to do that is to use closed-response rather than open-ended questions; there is no clearer specification of what responses are appropriate than to say, "pick one of these answers."

When open-ended questions are used, researchers would do well to think about ways in which they could narrow the discretion for interviewers and respondents about the kinds of answers that would count. The adverbial words, how, when, where, why, how much, should almost always be avoided; the kind of answer that is wanted can always be better specified by avoiding one of those words, and interviewers are often forced to rewrite questions in order to get answers in the correct terms (see Table 5.3).

Fourth, interviewer and respondent variability in the number of points they make in response to open-ended questions should be controlled by specifying a number of points to be made. A preference of ours is to ask for the "main reason" rather than "all the reasons," so that each respondent only gives one answer. If multiple answers are really desired, one could ask for the second most important reason. The goal is to eliminate interviewer and respondent uncertainty about what the measurement task is so that the researcher, rather than the interviewer or respondent, is making the decisions about what the respondent does in response to the question.

More important than any of these recommendations about particular question design strategies, however, is the need for careful evaluation of questions during the pretest and pilot phases of studies. Our research has shown that we cannot a priori identify classes of questions that are subject to significant interviewer effects. Questions that require a good deal of probing, particularly those where interviewers must go beyond simply repeating the question, in order to meet the question objectives, are most susceptible to interviewer effects. We have found such questions among attitudinal and factual questions. However, it is possible to identify many such questions during the pretest phases of research studies. We think identifying questions that generate a high rate of probing and improving them before finalizing the questionnaire is one of the most fruitful ways to reduce interviewer-related error in surveys.

| Table 5.3 |
| Adverbial Equivalents |
| **Adverbial Formulation** | **Alternative Interpretations** |
| 1. HOW do you get to work? | a. What kind of transportation do you use to get to work? b. By what route do you proceed to work? |
| 2. WHEN did you move to this address? | a. In what year did you... b. How many years ago did you... c. At what age did you... |
| 3. WHY did you vote for candidate X? | a. What characteristics of candidate X attracted your vote? b. What characteristics of candidate Y reduced your willingness to vote for him? c. Which interests or concerns of yours were reflected in your vote for candidate X? |
| 4. WHERE did you live when you were a teenager? | a. In which city or town did you live? b. With whom were you living? c. In what kind of housing structure did you live? |
| 5. HOW MUCH money do you make? | a. How many dollars are you paid? b. How does your income compare with that of others? c. How does your income compare with your needs? |
STRATEGIES FOR IDENTIFYING QUESTION PROBLEMS

A typical pretest of a survey instrument consists of experienced interviewers taking 10 to 20 interviews, then meeting for an hour or two with the researchers to discuss any problems they encountered. Although such pretests are useful, we are convinced researchers can and should do much more than that to evaluate their questions. We and others are in the process of designing and evaluating better ways to test questions. DeMaio (1983) and Converse and Presser (1986) provide two of the more thorough reviews of the strategies available for evaluating questions. The following are four steps that we think have considerable potential to improve question design:

1. **Focused discussion groups** are one of the oldest and best ways to begin the design of survey instruments. Groups of six to eight people typical of those to be interviewed should be brought together. Two to four groups are usually enough. The groups should be taken through a discussion of their experiences and thoughts about the areas to be covered in the survey. The availability of low-cost video taping makes it possible for a whole research team to review what is said in the groups without having to be present.

   From the point of view of minimizing interviewer effects, the specific outcomes of such discussions can include identifying terms or concepts that are not consistently understood and need definition, identifying the range of answers people give to potential questions, and identifying false assumptions that might be embedded in questions about the way respondents behave or think about issues to be covered. Such focussed group discussions will improve not only the ability of researchers to design a survey instrument that can be administered in a standardized way but will probably improve the extent to which they ask the questions that will provide the information they need. We believe that most survey research projects will benefit from several focussed group discussions prior to designing a survey instrument.

2. **Cognitive research techniques** have been newly introduced into the survey research field (e.g., Jabine et al., 1984). For many years cognitive psychologists have been studying how people process information, go about recalling information, and organize their thoughts, but only recently have their techniques been brought to bear on designing better questions in surveys. Their procedures primarily involve relatively intensive, lengthy sessions with a few typical respondents. The techniques which seem most widely used at the moment is the so-called “think aloud” interview. In this interview, respondents are asked to talk out loud about their thought processes as they go through a pretest version of a survey instrument. By listening to people talk about how they are understanding and thinking about the questions, researchers can identify ambiguities in what the questions mean and difficulties in the response task.

   A variation on this approach is to go through the interview schedule twice. First, a respondent is asked to answer questions as a normal pretest respondent would do. The interviewer then takes the respondent through the same interview schedule again, asking the respondent to discuss and explain answers and any difficulties a respondent had with the questions.

   These techniques are still being developed and evaluated. They are quite labor intensive, but obviously they have considerable potential payoff for identifying the kinds of problems that produce error.

3. **Better use of pretest interviews** is another area in which we think there is potential. Although interviewers with experience definitely have the ability to identify some characteristics of questions that will pose problems, our observation is that interviewers vary greatly in what they consider to be a problem. In fact, experienced interviewers tend to be good at solving the problems that researchers pose for them through poor question design, and we often find they are not sensitive to some of the difficulties questions create for them and for their respondents.

   We have been exploring the potential of giving interviewers more training in what to look for during a pretest. The goal is to sensitize them to questions that are not easy to read as worded and, in particular, to questions that routinely require probing and clarification in order to obtain adequate answers. We also believe having interviewers complete a standardized, written rating of each question will make their input more systematic and useful. We are still evaluating alternative ways to use interviewers in the question evaluation process, but we are convinced that more attention both to interviewer training and to the debriefing process can improve the ability of interviewers to identify question problems.

4. **Tape recording and coding** pretest interviews is, to our mind, potentially the most important innovation in the evaluation of survey questions. To use this strategy effectively, it is best to take at least 25 pretest interviews in order to increase the stability of measurement. Interviewers tape record interviews with respondent permission; this can be done in person or on the telephone. On the telephone, when the respondent is asked for permission to tape record the interview, be sure the respondent's answer is recorded on the tape.

   The tapes can be coded in about twice the time it takes to conduct an interview. The coding scheme that we have been working with requires a coder to count the number of times four or five different behaviors occur for each question: Whether or not the question was read exactly as written, whether or not the respondent asked for clarification, whether the respondent provided an adequate answer after the question was read or whether probing was required. We have used two approaches to the latter:

   1. coders count the number of times the interviewer either probes or repeats the question; or
2. coders count the number of times the respondent gives an inadequate or incomplete answer.

We then calculate the rates at which these events occurred; that is, the rates at which questions were misread, the rates at which respondents asked for clarification, the rates at which questions required additional probing or produced inadequate answers.

We are still working on the best way to code these events, the stability of the results, and the most cost effective way to use them as indicators of question quality. However, as indicated in data presented earlier in this chapter, we have evidence that questions that require additional probing are distinctively likely to be subject to interviewer effects. We also are sure that this technique reliably identifies questions that are difficult to read as worded. Although we expect to do further work and document the value of these activities, our work to date has convinced us that coding the interaction between interviewers and respondents with a particular eye to indications of difficulty of administering questions in a standardized way, provides meaningful, reliable clues to question problems that can and should be improved before they are put into the field for a full-scale survey.

CONCLUSION

In the next three chapters, we will talk about ways to get interviewers to do what we want them to do during the interview. However, it will be seen that there are limits to researchers’ abilities to control interviewers. As we said early in this chapter, in the studies which are published of interviewer effects on data, between one quarter and one third of the questions were subject to significant interviewer-related error. Moreover, almost certainly, the studies reported in that literature represent the high end of the continuum in terms of interviewer training and attention to methodological detail.

Although having interviewers who are able and willing to do standardized, nondirective interviews is one part of the process of error reduction, the task that they are given also plays an important role on how well they will be able to meet the goals of standardized interviewing. In our studies, we have found that some questions pose virtually impossible tasks for respondents and interviewers; the interviewers simply cannot use the words provided to meet the objectives of the question. Under those circumstances, interviewers have no choice but to write their own scripts. Sometimes that simply means innovations about which probes to use, and how often to use them. Sometimes it means completely writing new questions, developing free-lance explanations or instructions to respondents about what the question means in order to make the task manageable. Whatever the effort, when interviewers start doing that, they do it idiosyncratically, and the result is interview-related error. As others before have concluded, most notably Bradburn and Sudman (1979), one of the best courses to reducing interviewer effects on data is to write better questions; the critical step in writing better questions is careful, thorough development and pretesting procedures before a survey begins.