Why it is easy to write bad questions

Fowler, Floyd Jackson Jr.

WHY IT IS EASY TO WRITE BAD QUESTIONS[1]

FLOYD JACKSON FOWLER JR.

There are at least 8 standards that survey questions should meet. One reason it is easy to write poor questions is that researchers focus on some standards but ignore others. A more complex problem is posed by the fact that designing a question that is good according to one standard can make it a poor question when judged against another standard. Third, typical pretest and question evaluation procedures do not provide information about all 8 standards. It is easy to ignore a standard if there is no information about it. Finally, researchers often are committed to questions that have been used previously, even when there is evidence that they are flawed. This paper presents the 8 standards, the ways in which they can be in conflict, the challenges of evaluating questions, and the implications for standard survey practice.

1. Introduction

"Total survey design" refers to the idea that in designing or evaluating a survey project, it is important to consider all aspects of the data collection protocol, sampling, data collection procedures, interviewing, non-response, and question design, not just a few. The quality of survey data will be no better than the worst aspect of the data collection protocol, considering all aspects of it (Groves, 1989; Fowler, 1993).

In the area of survey question design and evaluation, a comparable construct, "total survey question design", may be appropriate. One of the important reasons that designing good questions is difficult is that there are numerous standards that survey questions should meet, and they are not necessarily related to one another. In a manner parallel to the concept of total survey design, if a question proves to be in-

1 This manuscript was prepared while the author was a guest professor at the Zentrum für Umfragen, Methoden und Analysen (ZUMA) in Mannheim, Germany. ZUMA's hospitality and support is warmly acknowledged.

adequate with respect to any one of these design standards, it is likely to be a poor survey question.

The purpose of this paper is to lay out 8 standards that survey questions should meet, to briefly describe why each is important, and then to discuss the implications of applying these 8 standards for the design of good survey questions.

2. The 8 Standards

The following are 8 standards that survey questions need to meet.

Content standards

Researchers have to decide what questions to ask, what content to ask about. Although that might seem an obvious standard, one that would be hard to fail to meet, in fact methodologists often find that researchers have not thought through what they really should ask about.

Decisions about question content emerge from an integration of two kinds of information. First, researchers need a set of analytic objectives, analytic questions that they want to answer based on the survey data they collect. Second, they need to find out what respondents have to say, what they are able and willing to tell them, that can provide the information they need. Often, a result of careful pre-survey evaluation of questions is that researchers become more informed about what questions they should ask in order to gather the information they need.

The content standard has two components:

a) Whether or not the answers to the questions asked will meet the analytic objectives;

b) Whether or not the questions are the right ones to get respondents to provide the information they have that will enable the researchers to address the analytic questions.

Cognitive standards

The official beginning of the cognitive aspects of survey methodology (CASM) movement in the Unitet States is usually identified as 1983. In that year, an advanced seminar on CASM was convened, bringing together statisticians, survey researchers, and researchers from the cognitive sciences to explore their mutual interests. A similar event was held at ZUMA in 1984. The results of these two conferences

were published by Jabine/Tanur/Tourangeau (1984) and by Hippler/Schwarz/Sudman (1987).

Since those conferences, and the related publications, there has been growing agreement that questions should meet cognitive standards before they are used in surveys (see Sirken, et al. 1999). Cognitive standards include:

a) That questions should be consistently understood by all respondents.

b) That respondents have access to, that is they know or can remember, the information required to answer the questions.

c) That the answers given to the questions accurately reflect the reality that respondents are being asked to describe.

Interpersonal Standards

A survey instrument is not only a set of questions for respondents to answer. When it is used by an interviewer, it becomes a script for (at least one side of) an interaction. Thus, the fact that the questions will be used in an interpersonal context, in a particular way to collect data, produces another set of issues that need to be addressed in designing a survey instrument.

a) The questions need to be designed so that they can be asked exactly as written. When the questions are read orally, they should be clear enough that the respondent is ready to answer the question. Questions that require extensive probing, or that lead to respondents to frequently ask for clarification, have been shown to have adverse effects on data quality (Fowler/Mangione 1990; Fowler/Cannell 1996).

b) The interview schedule also must produce an interaction that seems reasonable and appropriate to respondents. Instruments that seem repetitive, overly detailed, or irrelevant to the purposes of the survey from the respondents perspective may have an adverse effective on respondent motivation.

c) The content of the survey instrument also must appear appropriate to the respondent. If questions call for answers that are seen by respondents to be highly personal or potentially embarrassing, special efforts are likely to be needed in order to obtain accurate reporting. Researchers need to assess the appropriateness of the data collection context, with particular attention to the kind of relationship the interviewer and respondent may have developed, when thinking about what questions to ask and how to ask them.

Psychometric standards

Survey questions must meet standards with respect to the information that they provide. The three main standards include:

a) How answers are distributed is one key to how much information is provided by a question. If there is little variation in the answers, there may be little value to the question. Also, if questions apply to only a subset of a sample, their value is related to the number of people to whom they apply and who will actually answer them.

b) Measures of validity, how well the answers to a question measure what the researcher intends to measure are, of course, central to any assessment of the quality of a survey question (Cronbach/Meehl 1955).

c) Reliability, the extent to which answers are consistent over time (in the absence of change) is another aspect of psychometric analysis. Reliability does not ensure validity, of course, but unreliability places a limit on how valid measures can be.

Usability

Survey instruments usually are in some sort of paper form or on a computer. They will be used either by interviewers or by respondents themselves. Survey instruments need to be as easy as possible to use so that whoever is using them, interviewer or respondent, can focus on the question-and-answer process and not face challenges in figuring out which questions are to be answered or how to answer them.

Multi-mode capability

Increasingly, it is desirable to design survey questions that can be asked in more than one mode of administration. Survey instruments can be interviewer administered or self administered; interviews are done in person and by telephone. It is highly desirable to be able to compare answers that are collected by different modes. A particular incentive for comparability across modes is that multi-mode designs, whereby more than one mode of data collection is used to maximize the rate of response, are becoming increasingly used (see Dillman 2000). However, to use such designs, it is essential to be able to assume that the answers to questions are not affected by the mode of data collection itself.

To meet this standard, researchers need to think about how questions would be administered both with and without and interviewer. Questions should be designed so that they form a script for an interviewer, as well as being appropriate for self-administration. The response alternatives need to be simple enough that they can be administered on the telephone, as well as being appropriate in self administration.

Multi-language capability

It may be obvious that considering how well questions can be translated into other languages should be a concern in the design of survey questions. In the past, at least in the United States, researchers often designed questions in English, then worried about ease of translation later. We know that if researchers think about how easy or hard it is to translate a particular question at the time of its initial design, the likelihood that questions can be comparable across languages is greatly increased.

Cost effective use of survey time

Finally, there is always pressure on survey researchers to use survey time well. It is very common for researchers to want to ask more questions than budgets will permit or than respondents will put up with. Thus, in addition to all the previous standards, researchers also must attend to setting priorities and to using the survey time to ask the questions that will yield the most valuable information for the research purposes; in short, they must try to achieve their research objectives in the most parsimonious way, consistent with data quality.

Table 1 provides a summary of the 8 standards.

Table 1: Summary of Standards

1. Right content?
2. Cognitive Standards.
3. Interpersonal Standards.
4. Psychometric Standards.
5. Usability Standards.
6. Multi-mode Capability.
7. Multi-language Capability.
8. Cost-effective use of survey time.

3. Why it is easy to write bad questions

There are four main reasons that researchers write bad questions:

1. There are many standards; researchers often fail to attend to all of them, either through inattention or lack of appreciation of their importance.
2. The standards sometimes conflict with one another; making a question better in one respect can make it worse in some other respect.
3. Presurvey evaluation procedures often are inadequate to provide information about whether or not standards have be met.
4. Researchers do not like to change questions that have been used before.

Many Standards

It is a common experience for those testing questions to learn that designers of questions have not even thought through what they want to measure and why (standard 1), much less begun to think through carefully the next and more complex step of which questions to ask to achieve their poorly specified objectives. Many questions are given birth without any real thought about standards at all.

Fundamental to the problem of question design is that many of these standards have not been widely appreciated as being important. For many years, "usability", whether or not interviewers could administer the survey instrument, was the main focus of presurvey testing. In some disciplines, studies of validity and reliability are routinely done. In other disciplines, researchers are satisfied with "face validity", the appearance of validity, assuming questions measure what they appear to measure, without any empirical evidence.

New research has documented the importance of standards that were not previously appreciated. It may seem strange to say that it is new to demand that questions be consistently understood and ask for answers that respondents are able to provide. However, it is only in the past decade that cognitive evaluation of questions began to be done anywhere, and even today it is routine only in a small, though growing, number of centers (see Willis/DeMaio/Harris-Kojetin 1999).

It was over 30 years ago that Charles Cannell began documenting the effects of the way interviewers relate to respondents on data quality. He observed interviews and demonstrated that interviewers often asked questions in their own words, rather than the words that were provided (Cannell/Marquis/Laurent 1977). He showed he could manipulate interviewer behavior and improve reporting (Cannell/Oksen-

berg/Converse 1977). More recent studies show how standardized interviewing breaks down when questions are poorly designed and demonstrate that questions that consistently require interviewer probing and clarification are associated with more interviewer-related error (Fowler/Mangione 1990; Fowler 1991; Fowler/Cannell 1996), and how socially sensitive material is reported more favorably when an interviewer collects data than when questions are self administered (Dillman/Tarnai 1991; Aquilino 1994; Tourangeau/Smith 1998; Turner/Forsyth/O'Reilly, et al. 1998). Despite all the evidence, attention to how question design will affect the interviewer-respondent relationship is among the most neglected of the standards for questions.

Three other standards have emerged as important as the world has changed. Until recently, designers of survey questions thought their questions would be used in a single language. In the United States that is no longer the case. Hardly any surveys of importance in the U.S. are done only in English, even for studies restricted to the U.S. Researchers in other parts of the world have been concerned about multi-lingual use far longer than those in the U.S. Moreover, the desire for cross-national studies is growing daily.

There are fundamental difficulties in achieving comparable measurement across languages and cultures. However, the chances of success are better if question designers think about multi-language use before, not after, their questions are designed. For too long, researchers designed their questions in the primary language, then turned to the problem of translation. Some words and concepts translate more easily than others. To the extent that question designers take that issue into account when designing their instruments, the comparability of questions across languages will be better.

A similar issue arises with respect to mode of data collection. Increasingly, researchers want to collect data using more than one mode of data collection: with and without interviewers, with and without computer assistance. There are some features of question design that facilitate the comparable collection of data across modes. Interviewers need scripted questions that can be read as worded. Response alternatives that are long and wordy may be all right in a self-administered form, but they are hard to use on the telephone. Complex skipping instructions are easy for interviewers, especially with computer assistance. However, they pose a challenge for self administration. If a researcher designs an instrument with only one mode of data collection in mind, the chances are good that it will not adapt easily to other modes. Designing for multi-mode data collection from the start will greatly increase the chances that an instrument can produce comparable data across modes.

Finally, the introduction of computers into all phases of data collection adds a new and challenging dimension to the concept of usability. A solid premise of survey instrument design is that the tools of data collection should be as easy to use as possible, so the participants can concentrate on the question-and-answer process. Initially, when questions were put on computers, designers more or less mirrored the designs they had used on paper. However, recent studies of the interaction between interviewers and computers has shown that there is much room for improvement in the way computer-assisted data collection tools are designed (Couper/Hansen/ Sadowsky 1997) . Interviewers are found to waste considerable time trying to navigate, have great difficulty making corrections, cannot take advantage of many of the tools and aids that computers have to offer, and spend more time looking at the computer than looking at respondents. There are similar issues with computer-based surveys designed for respondents to use themselves (see, for example, Dillman 2000). The point is that the concept of usability has risen to a new level of complexity with the introduction of computers, and how easily interviewers and respondents can use them is a critical part of the evaluation of how well a survey instrument is designed.

Conflicting Standards

A second set of challenges for designing good questions stems from the fact that sometimes improving a question from the perspective of one standard can make it worse with respect to another standard. Two places where standards are particularly likely to conflict include:

a) Providing detailed definitions and explanations to make questions clearer can also make questions complicated and harder to administer, particularly for interviewer administration (see Conrad/Schober 2000).

b) Having more response categories, and labeling the response alternatives with adjectives, have both been shown to improve the psychometric performance of questions. However, having numerous response alternatives makes a question harder for an interviewer to administer, particularly on the telephone. Also, using numerous labels on alternatives makes it less likely that a question can be easily translated into other languages. The following pairs of questions provide examples of the conflicts that are inherent in question design.

Example 1a: In the past 7 days, how many times did you exercise?

Example 1b: Exercise can sports, running, swimming include, and cycling, as well as working out at a club or gym. It can also include walking. In the past 7 days, not counting exercise you got while working or doing chores at home, how many times did you exercise—none, 1 or 2, 3 to 5, or 6 or more?

Comment: Example 1a leaves the definition of "exercise" up to the respondent. Also, by asking for a specific number of times, it may pose a difficult task for respondents to be as precise as the question demands.

Example 1b provides information about what is meant by exercise, and it allows respondents to answer in categories. However, it is a much wordier question, which may prove to be confusing to respondents and hard for an interviewer to administer. Also, the addition of labeled categories of times may prove to be harder for respondents to cope with than giving a precise number, particularly in a telephone interview, in which respondents must remember the categories.

Example 2a: How would you rate the way the Chancellor is doing his job - excellent, good, not so good, not good at all?

Example 2b: Using a rating scale from 0 to 10, where 0 is as bad as possible and 10 is as good as possible--- What number from 0 to 10 would you use to describe the way the Chancellor is doing his job?

Comment: Example 1a is a classic 4-category rating, with each category labeled. The numerical rating in 2b only has labels at the ends. Also, explaining the 0 to 10 task will take more time and be more complex than asking question 2a. However, using numbers has three advantages. First, psychometricians will like the fact that it provides 11 categories, which is likely to provide more information than 4 categories. Second, in a telephone interview, it is easier to remember all the possible answers, 0 to 10, for the numerical task than for the 4 labeled answers. Finally, the problem of translation across languages is easier for numbers than for words.

It is important to note that standards do not always conflict. Sometimes, applying several standards leads consistently to the same conclusion about which question is best.

Example 3a: Tell me what you think about the following statement: The news reported in newspapers and on television is not to be trusted—do you strongly agree, generally agree, neither agree nor disagree, generally disagree, or strongly disagree?

Example 3b: How much do you trust the news reported in the newspapers and on television—very much, some, a little, or not at all?

Comment: Example 3a is one of the most popular question forms in survey research. Question 3b is an alternative way to accomplish a similar research objective, to measure trust in the media. Putting aside for the moment whether or not trust in newspapers and magazines is a good concept about which to ask and whether or not newspapers and television should be combined in one question or separated, if one applies the standards outlined in this paper, question 3b is definitely a better question than question 3a.

1. Applying cognitive standards, the agree-disagree format has been consistently found to be confusing to respondents. It is particularly confusing when a respondent has to disagree with a negative statement in order to express a positive opinion. In this case, respondents have to disagree that media are not to be trusted to say that they can be trusted.

2. The response task is complicated in the agree-disagree form, with difficult categories to remember, making it difficult to administer by interviewers, particularly on the phone.

3. Interpersonal forces have been shown to produce acquiescence bias in some respondents, especially those with less education. That means that they consistently will be more likely to "agree" when a question is in this form than to give a comparable answer to a question in some other form. A significant number of respondents will agree with this question and with its opposite phrased positively (e.g. news in newspapers and on television is to be trusted) (see Schuman/Presser 1981; Dillman/Tanai 1991; Converse/Presser 1986).

4. Psychometrically, agree-disagree questions have disadvantages as well:

   a) The agree-disagree questions are almost always analyzed as two-category variables. In contrast, question 3b distributes answers across four categories. Thus, there usually is more information provided by the four-category scale.

b) There is frequent debate about where to place those who neither agree nor disagree: are they in the middle or should they be treated as missing data? It is not at all clear that those in the middle category can be put in order with the rest of the respondents.

One advantage of the agree-disagree question form is that one can ask almost any opinion or attitude question in that form. Thus, respondents (and interviewers) can enjoy a consistency of response tasks that is harder to provide with more direct questions. However, in virtually all other respects, alternatives to the agree-disagree question form will better meet the standards for question design.

In conclusion, there is the potential for real conflict among standards, which accounts for some of the shortcomings in the way questions are designed. Nonetheless, if researchers are mindful about the various standards, they usually can find a way to design a question that will do a reasonable job of meeting them all.

Inadequate Presurvey Testing

The third reason that poor questions are used in surveys is that question testing protocols prior to surveys often are not adequate to evaluate all aspects of the question that are important. Historically, pretesting consisted mainly of a small field test, in which senior interviewers conducted some test interviews and reported back on their experiences. In the last decade or so, question testing has been improved.

The core steps for presurvey question evaluation include:

a) Systematic review of questions (Lessler/Forsyth 1996; Fowler 1995)

b) Cognitive testing (Lessler/Tourangeau 1989; Willis/DeMaio/Harris-Kojetin 1999)

c) Behavior coding of field pretests (Fowler/Cannell 1991; Oksenberg/Cannell/Kalton 1991).

These steps are becoming increasingly common, particularly at centers that place great emphasis on methodological excellence. Each of the steps has the potential to identify question problems before a survey is done. Other presurvey evaluation steps that are done less frequently include:

a) focus group discussions prior to drafting the survey instrument to learn what respondents have to say on the survey topic and to explore vocabulary issues

b) an evaluation of the literacy level of questions by a specialist in reading

c) tests of the usability of survey instruments, in paper and pencil or in computer form

d) debriefing of respondents after field pretest interviews.

Table 2: Value of Various Evaluation Options for Providing Information About How Well Questions Meet Various Standards

	Standards for Survey Instruments						
Evaluation Options	Content	Cognitive	Inter-personal	Psycho-metric	Usability	Multi-mode Capability	Multi-Language
Pre-Field Testing							
Expert Review	YES	H	-----	-----	H	H	H
Focus Groups	YES	H	-----	-----	-----	-----	-----
Cognitive Ints	H	YES	-----	H	-----	-----	H
Usability Tests	-----	-----	-----	-----	YES	-----	-----
Field Testing							
Interviewer debriefing	-----	H	H	-----	YES	-----	-----
Respondent debriefing	-----	H	H	H	H	-----	-----
Behavior coding	-----	H	YES	-----	-----	-----	-----
Analysis of pilot data	-----	-----	-----	YES	-----	-----	-----
Reinterviews	-----	H	-----	YES	-----	-----	-----
Split Form (Split Ballot) tests	-----	-----	-----	YES	-----	YES	YES

YES = provides information; H = helpful, but not sufficient.

Table 2 is an attempt to relate different testing steps to the 8 standards outlined at the beginning of this paper. The cells are labeled "yes" to indicate that a test can provide information about how well a question meets a certain standard, "H" if it can "help" to evaluate it, and blank if it makes no contribution. Although the coding admittedly is arbitrary and subject to debate, two points stand out from the table: First, the most common testing strategies—expert review, focus groups, cognitive testing, and behavior coding of field interviews -- address the first three standards plus usability. However, they do very little to help with the other four. Particularly lacking are measures of data quality, validity and reliability, and comparability of data across modes and across languages. Second, there are three underutilized kinds of evaluation steps that would add considerably to knowledge about key elements of survey instruments: analysis of pilot or pretest data, reinterviews, and split-sample testing.

Pretest samples are often a little too small for meaningful analysis. Samples of 25 to 35 are large enough for usability testing and to get meaningful behavior coding measures for questions that are asked of most respondents. Probably having pretest samples of size 50 would be better if some psychometric evaluation is planned. By simply tabulating marginal distributions and cross tabulating variables that may potentially be redundant, questions can be identified that are yielding little information because there is little variation in answers, they apply to too small a percentage of the sample to be statistically useful, or they overlap with other questions. Correlational analyses to assess construct validity can be conducted to identify items that can be dropped from multi-item indices.

Reinterviews after pretests offer two potential contributions to question evaluation that are hard to achieve in any other way. First, they provide the opportunity to assess the reliability of answers. By asking the same question twice, at different points in time, a measure of the stability of answers can be obtained. Items measuring something thought to be stable that show low reliability are obvious candidates for revision or dropping.

An even more interesting use for reinterviews has become possible with the advent of computer-assisted interviewing. A reinterviewer now can ask a set of questions that have been previously asked, blind to the original answers. Then, either after a given answer has been given or at the end of the reinterview, the computer can check reinterview answers against original answers, notify the interviewer of discrepancies, and enable interviewers to ask probe questions to understand the reason for the differences. Such a process can be a valuable adjunct to cognitive testing,

particularly because it can reveal confusion and response inconsistency under realistic data collection conditions.

Split-ballot tests are among the most useful question evaluation activities. Their disadvantage, of course, lies in the fact that they entail additional data collection at the presurvey stage. However, the great strength of such designs is that they permit addressing a key question that cannot be addressed in any other way: whether two versions of the same question produce the same or different estimates. This, of course, is crucial to evaluating the comparability of data collected by phone and self administration, as well as assessing the comparability of questions to be asked in different languages. In addition, when changes are made to improve questions, based on findings from cognitive testing or field pretests, there is always a need to know how the changes will affect the results. If a change is designed to clear up some ambiguity found in a question, one would like to know if the resulting estimate is affected in the expected way. If a question was found awkward to administer, we should find out if the interviewer-respondent interaction is improved. In some cases, an improved question should produce a different result. In other cases, if a question has been streamlined to improve the ease of administration, the hope may be that the resulting estimate will be essentially the same as from its predecessor. In all these cases, it is only by comparing results from comparable samples that one will know the answer. Small split-ballot field tests, with samples of 150 to 200 interviews randomized to alternative forms of a survey instrument can provide information on these and many more topics in a cost-effective manner.

In conclusion, one important way to get researchers to tend to various standards is routinely to have test procedures to evaluate how well questions are designed. Recent developments in testing, such as cognitive testing and behavior coding of pretests, have definitely improved the evaluation of some aspects of questions. However, current protocols are generally inadequate to evaluate all the aspects of questions that should be evaluated. Any important aspect of questions that is not routinely evaluated is likely to be a continuing source of problem questions.

Resistance to New Questions

There are some reasons that have merit for wanting to repeat previously used questions. Using previously established measures permits replication of previous research findings, as well as the possibility of measuring change over time. Reviewers, for grant proposals and for articles submitted to journals, often are reassured when researchers propose to use, or have used, questions that have been used previously.

The question researchers and methodologists must face is whether these reasons are compelling enough to continue to ask a question that has demonstrable flaws.

Every discipline that uses survey research has a past filled with questions that were designed and used before we knew as much as we do today about question design. The repeated use of these questions constitutes an important source of poor quality survey research.

There may be compelling reasons in some circumstances for repeating questions that are less than perfect. However, there is no reason that all key questions in a survey instrument should not be evaluated. If a question to which a researcher is committed is found to be problematic in some way, the researcher then is in a position to make an informed choice among three reasonable options:

1. Revise the question to make it better
2. Use the flawed question, but be aware of its problems when analyzing the results and inform data users of the problems that were identified
3. Use both the original and a revised, improved question, so that results from the original question can be compared with previous studies, the impact of the revisions can be evaluated, and an even more informed choice, with improved question options, can be made by future researchers studying the same topic area.

4. Conclusion

Knowledge about the design of survey questions has progressed considerably over the past decade or so, as has the commitment to better question evaluation. However, better survey instrument design provides one of the most important opportunities to improve the quality of survey research practice. There are several needed steps.

1. We need continued research that links characteristics of questions to the quality of the resulting data. More empirically based generalizations about question design and survey error will accomplish two important goals: to teach us details about how to design better questions and to heighten general awareness of the significance of poor question design for the quality of survey estimates.
2. We need a broader appreciation of the complexity of designing good survey questions. While the 8 standards outlined in this paper might be subject to some debate, there is no doubt that good questions must meet numerous,

largely unrelated standards. To design good questions is not something that can be done by the amateur, by committee, or without serious evaluation activities and the input of experts.

3. More and better presurvey tests need to be used routinely. Even the testing that goes on at our best research institutions is not enough to provide the information researchers need to know about their questions. The next addition to routine testing protocols, beyond systematic expert review, cognitive testing, and behavior coding of field pretests, should be small split-ballot pilot tests of revised or alternative forms of questions. Such tests would provide information on whether the question revisions emerging from the early testing had the expected effect on results, and they would permit psychometric evaluation of questions prior to surveys. However, there are other kinds of evaluations mentioned above that are valuable. The important point is not to specify exactly which protocol should be used but to emphasize that attention should be given to more of the important aspects of questions than is the case with current protocols.

4. Finally, researchers, reviewers at funding agencies, and those who review journal articles need to be educated so they understand that new questions, properly evaluated, are likely to be better than old questions. A question that was used in the past, usually designed without the benefit of current testing wisdom, should not be an object of reverence, but rather a subject for evaluation and, usually, improvement. Researchers say that a certain question was used previously and it "worked", when there was no real evaluation of the response process or the quality the answers—in short, no real evidence that it "worked". The standard for whether or not a question is a good one for a particular purpose should rest not on its pedigree but on the evidence gleaned from an appropriate evaluation protocol

In conclusion, good researchers write bad questions: 1) because they do not appreciate all the standards that apply to survey questions; 2) because sometimes addressing multiple standards can be in conflict; 3) the evaluation protocols that would flag all the potential weaknesses of their questions are not routinely in place; and 4) because they overvalue previously used questions. The challenges for dealing with all these issues are obviously daunting. However, the real key to change is better question evaluation. When that is in place, better questions will follow.

Contact

Floyd Jackson Fowler Jr.
Center For Survey Research
University of Massachusetts Boston
100 Morrissey Boulevard
Boston, MA 02125-3393
U. S. A.
Phone (617) 287-7200; Fax (617) 287-7210
csr@umb.edu

References

Aquilino, W. S. 1994: Effects of interview mode on measuring depression in younger adults. Journal of Official Statistics, 14(1), 15-30.

Cannell, C. F./Marquis, K./Laurent, A. 1977: A summary of studies. Vital & Health Statistics, Series 2, 69. Washington, DC: Government Printing Office.

Cannell, C. F./Oksenberg, L./Converse, J. 1977: Experiments in interviewing techniques: Field experiments in health reporting, 1971-1977. Hyattsville, MD: National Center for Health Services Research.

Conrad, F. G./Schober, M. F. 2000: Clarifying question meaning in a household telephone survey. Public Opinion Quarterly, 64, (1), 1-28.

Converse, J./Presser, S. 1986: Survey questions. Beverly Hills, CA: Sage.

Couper, M. P./Hansen, S. E./Sadowsky, S. A. 1997: Evaluating interviewer use of CAPI technology. pp. 267-286 in: Lyberg, L. E./Beimer, P./Collins, M./et al. (Eds). Survey measurement and process quality. New York: John Wiley and Sons.

Cronbach, L./Meehl, P. 1955: Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Dillman, D. A. 2000: Mail and Internet surveys: The tailored design method. New York: John Wiley.

Dillman, D. A./Tarnai, J. 1991: Mode effects of cognitively designed recall questions: A comparison of answers to telephone and mail surveys. pp. 287-310 in: Beimer, P. N./Corres, R. M./Lyberg, L. E./Mathiewetz, N. A./Sudman, S. (Eds.), Measurement errors in surveys. New York: John Wiley.

Fowler, F. J./Mangione, T. W. 1990: Standardized survey interviewing. Thousand Oaks, CA: Sage.

Fowler, F. J. 1995: Improving survey questions. Thousand Oaks, CA: Sage.

Fowler, F. J. 1991: Reducing interviewer related error through interviewer training, supervision, and other means. pp. 259-278 in: Beimer, P. N./Groves, R. M./Lyberg, L. E./Mathiewetz, N. A./Sudman, S. (Eds.), Measurement errors in surveys. New York: John Wiley.

Fowler, F. J./Cannell, C. F. 1996: Using behavioral coding to identify cognitive problems with survey questions. pp. 15-36 in: Schwarz, N. A./Sudman, S. (Eds.), Answering questions. San Francisco: Jossey-Bass.

Groves, R. M. 1989: Survey errors and survey costs. New York: John Wiley.

Hippler, H. J./Schwarz, N./Sudman, S. (Eds.), 1987: Social information processing and survey methodology. New York: Springer-Verlag.

Jabine, T. B./Straf, M. L./Tanur, J. M./Tourangeau, R. (Eds.). 1984: Cognitive aspects of survey methodology: Building a bridge between disciplines. Washington, DC: National Academy Press.

Lessler, J. T./Forsyth, B. H. 1996: A coding system for appraising questionnaires. pp. 259-292 in: Schwartz, N. A./Sudman, S. (Eds.). Answering questions. San Francisco: Jossey-Bass.

Lessler, J. T./Tourangeau, R. 1989, May: Questionnaire design in the cognitive research laboratory. Vital Health & Statistics, Series 6, 1. Washington, DC: Government Printing Office.

Oksenberg, L./Cannell, C. F./Kalton, G. 1991: New strategies of pretesting survey questions. Journal of Official Statistics, 7(3), 349-366.

Schuman, H./Presser, S. 1981: Questions and answers in attitude surveys. New York: Academic Press.

Sirken, M. G./et al. (Eds.). 1999: Cognition and survey research. New York: John Wiley.

Tourangeau, R./Smith, T. W. 1998: Collecting sensitive data with different modes of data collection. pp. 431-453 in: Couper, M. P., et al. (Eds.), Computer assisted survey information collection. New York: John Wiley.

Turner, C. F./Forsyth, B. H./O'Reilly, J. M./et al. 1998: Automated self-interviewing and the survey measurement of sensitive behaviors. pp. 455-473 in: Couper, M. P., et al. (Eds.), Computer assisted survey information collection. New York: John Wiley.

Willis, F. B./DeMaio, T./Harris-Kojetin, B. 1999: Is the bandwagon headed to the methodicalogical promised land? Evaluating the validity of cognitive interviewing techniques. pp. 133-154 in: Sirken, M. G., et al. (Eds.), Cognition in survey research. New York: John Wiley and Sons.