CONSTRUCTION STRATEGIES FOR COMPLEX SURVEY QUESTIONS

Paul Beatty, National Center for Health Statistics; Floyd J. Fowler, Jr., and Gregory Fitzgerald, Center for Survey Research, University of Massachusetts--Boston
Paul Beatty, 6525 Belcrest Rd. Room 915, Hyattsville, MD 20782

Keywords: Questionnaire design, behavior coding, cognitive aspects of survey methodology

Standardized survey questions must often convey a large amount of information to respondents. Consider the following sample question:

Q1.A What kind of place do you go to most often when you need routine medical care, such as a physical examination? Is it a doctor's office, a clinic or health center, hospital emergency room, hospital outpatient department or some other place?

The question contains many details, all of which are important: the emphasis on *routine* care; the example of a *physical examination* to illustrate the intent of the question; and the *list of possible places* we want to be considered as eligible responses. Survey researchers expect that an interviewer reads this question accurately to a respondent, who comprehends and retains all of these components, ultimately providing a codeable response. Ideally, we want this to happen with minimal intervention from the interviewer (Fowler and Mangione, 1990).

Yet questionnaire designers have little guidance on how to actually assemble these pieces into a whole that is clear and comprehensible. We could just as easily imagine an alternative version of the question above:

Q1.B People can get medical care in different places, including at a doctor's office, a clinic or health center, a hospital emergency room, a hospital outpatient department, or some other places. What kind of place do you go to most often when you need routine medical care?"

This revision uses almost identical words, but they are organized differently. Which organizational strategy is better for respondents? That is, which one results in fewer errors, less respondent confusion, and fewer time-consuming interactions with interviewers?

The preceding example represents only one of the sorts of complexities that are commonly introduced in survey questions. In this study, we attempt to identify some of these common cognitive complexities and develop construction guidelines to help questionnaire designers minimize their cognitive burden.

Question structures and cognitive tradeoffs

The first version of the medical care item (Q1.A, above) asks a complete, self-contained question before explaining the initial response categories. A potential problem with this is that respondents might start to formulate a response before these categories are read, providing answers that do not correspond with what we want. Responses such as "Well, I usually go to my HMO" will require additional probing, which is inefficient and could lead to interviewer-related error. The alternative version (Q1.B) gets around this problem by specifying the response categories at the beginning-- but respondents may have trouble making sense of these until they have heard the rest of the question.

In order to decide which structure is preferable, it is useful to break the question down into major components that serve specific cognitive functions. Borrowing from the language of mathematics, we refer to first component as the domain, which tells the respondent what the question is about. The second component, which we refer to as the range, tells the respondent what we expect them to give back to the interviewer. In the initial example, the domain is the type of place that a respondent receives most medical care; the range is the list of medical care facilities that are provided as response categories. The original version of our example begins with the domain and ends with the range; the alternative version specifies the range first, followed by the domain.

We are interested in learning whether one structure is more cognitively effective than the other. This is a particularly important issue when asking complex survey questions, because important details can easily be lost if the concepts within the question are not presented in the most logical manner. It may be that one structure is not clearly "better" than the other, but that tradeoffs are involved. For example, respondents may forget earlier parts of questions, or "tune out" later parts, regardless of how they are structured. In such cases, we need to prioritize the most important details of the question-- or be open to rewrites that split single questions into multiple ones (see Fowler, Beatty, and Fitzgerald, in press).

Methods

Beginning with a review of several major national surveys (primarily the National Health Interview Survey), we identified 19 questions that were particularly complex. After selecting these questions, we constructed alternates

for each of them (the alternates retained the same subject matter, terms and response categories, but were structured differently). Next, we randomly selected either the original or the alternate version of each question for inclusion in an instrument labeled "Questionnaire A." The questions that were not selected were assembled as "Questionnaire B." Thus Questionnaires A & B both contained the 19 questions in either original or alternate versions.

In the first phase of the study, the questionnaires were tested in cognitive interviews, with half of the subjects responding to Questionnaire A and the other half responding to Questionnaire B. All subjects were probed about their responses and debriefed following the interviews. Based on probe responses and debriefings, we evaluated the strengths and weaknesses of the alternate questions. In a few cases, it was clear that *both* versions ignored more serious cognitive problems than were not adequately addressed by our restructuring. When this occurred, we rewrote the alternate question versions.

In the second phase of the study, the revised questionnaires were administered through telephone interviews of a small national sample (n=156). The first 112 of these interviews were also tape recorded and behavior-coded. The behavior coding procedure provided us with counts of how often interviewers read questions exactly as worded, were asked for clarifications, were interrupted during reading, and used various probes in order to get an adequate answer (see Fowler and Cannell, 1996). Our analysis of the strengths and weaknesses of alternative question versions is based on differences in tabulated responses, and differences in distributions of behavior codes.

Results

Presenting results for all 19 questions is clearly beyond the scope of this paper. Rather than providing a cursory overview of our findings for each question, we present a few key examples that illustrate some of the major design decisions that we want to highlight.

Example 1: Domain/Range Presentation

Returning to our initial example about sources of medical care: cognitive interviews suggested that both versions failed to address a serious problem. Regardless of how the material within the question was structured, both versions make an "embedded assumption" that respondents actually have a place of "regular care." This often proved to be incorrect, and created a variety of cognitive problems. We therefore revised the alternative version (Q1.B) by splitting it into two questions (below):

Q1.B1 People can get routine medical care in different places, including in a doctor's office, a clinic or health center, a hospital emergency room, or a hospital outpatient clinic. Do you have a place you go to most often for routine medical care such as a physical examination?

Q1.B2 *(If yes:)* What kind of place do you go to most often for routine care, such as a physical examination?

Question Q1.B1 contains essentially the same information as the original question, Q1.A (the emphasis on routine care, and mentions about the sorts of places we have in mind). The difference is that Q1.B1 explicitly accounts for the possibility that someone does not have such a place. Without asking this lead-in question, respondents are likely to "force" their answers to conform to one of the categories provided due to assumptions of communicative relevance (Clark and Schober, 1992, Schwarz and Hippler, 1991).

This is in fact the case, as tabulated responses from telephone interviews reveal (Table 1, below). Almost 20% of respondents took the "out" provided in Q1.B1, whereas only 1% of respondents volunteered such responses in Q1.A. This suggests that our "embedded assumption" about having a place for routine care is incorrect for many respondents. Furthermore, the original version of the question creates misleading results by not explicitly providing for this outcome.

Table 1: Places of routine medical care

	Q1.A	Q1.B
Doctor's office	72%	64%
Clinic	19%	14%
Emergency room	1%	0%
Outpatient clinic	4%	3%
Other	3%	0%
No Routine Care	1%	19%

Although Q1.B represents a significant improvement, it is still clear that both versions of the question have problems. While Version B eliminates the "embedded assumption," behavior coding results (Table 2, below) reveal that its added verbiage creates new difficulties. The first part of Version B (Q1.B1) required more probing and interviewer clarifications than Version A, and interviewers misread Version B more often as well.

Fortunately, it may be possible to combine elements of both versions into a promising solution. By retaining the 2-question format of Version B, but eliminating the response categories in the first question, we may be able to minimize verbiage while eliminating unwarranted assumptions within the question:

Q1.C1 Do you have a place you go to most often for routine medical care such as a physical examination?

Q1.C2 (If yes) People can get routine medical care in different places, including in a doctor's office, a clinic or health center, a hospital emergency room, or a hospital outpatient clinic. What kind of place do you go to most often for routine care?

The most important lesson here is that the structural issues ("domain first" vs. "range first") are relatively minor compared to conceptual ones-- however, structural decisions *can* exacerbate problems created by unnecessary conceptual complexities. In this example, the conceptual problems called for a solution that restructuring alone could not solve.

However, after breaking the question into two parts, it seems most logical to put the *range first* in Q1.C2. We have already established the general domain of interest through the preceding question, and want to make sure that respondents listen carefully to the full range of responses without interruption. That is not to say that "range first" is always the best strategy. It probably makes sense to do so if the response categories need to be clearly specified, and if the conceptual domain of the question is relatively simple (or established in the preceding question). In other situations, it may be preferable to read the domain first-- which is, after all, the way most people pose questions in everyday language.

Example 2: Dangling Qualifiers ("exclusive" type)

Many survey questions employ "qualifiers" that alter the original meaning of the domain. In the following question, the qualifier is designed to *exclude* certain types of information from the response:

Q2.A During the past 12 months, how many times have you seen a doctor or other health are professional about your own health at a doctor's office, a clinic, or some other place? Do not include times you were hospitalized overnight, visits to a hospital emergency room, home visits, or telephone calls.

The difficulty with this question is that the qualifier "dangles" after the question mark. It seems plausible that respondents begin to formulate an answer before the qualifier has been completely read. An alternative plan would be to convey the "qualified" information before actually asking the question, as follows:

Q2.B This question is about times you have seen a doctor or other health care professional in a doctor's office or clinic, but not counting overnight hospital stays, emergency room visits, home visits, or telephone calls. During the past 12 months, how many times have you seen a doctor or other health care professional about your own health?

Cognitive interviews provided little illumination regarding which version was better. In addition, tabulations from telephone interviews revealed that the two versions generated very similar response distributions. Behavior

Table 2: Behavior codes for Q1, places of routine medical care

Question	Minor read errors	Major read errors	% Read errors	% interrupt	Repeat question	Repeat cate-gories	Other probe	% probed	% R asks for clarif.
Q1.A	4		7.8%	15.7%			3	5.9%	
Q1.B1	13		18.8%	1.6%	2		5	11.5%	4.9%
Q1.B2	5	4	18.8%	29.2%			1	2.1%	4.2%

codes (Table 3, below) reveal slightly more information. For example, interruptions are more common in Version A. This makes sense, because respondents tended to interrupt at the question mark, not waiting for the interviewer to read the qualifier before answering. On the other hand, more respondents asked for clarification of the second version, suggesting that it might be less clear-- i.e., the qualifier may not be meaningful before the rest of the question.

We would argue that *in this case*, the confusion created by Q2.B is more serious than the interruptions caused by Q2.A. After all, those respondents who have not seen any doctors during the last 12 months at all will never find the "excluding qualifier" to be relevant-- that is, since the answer is already zero, there are no instances to exclude from the response. We suspect that these respondents account for most of the interruptions. In general, however, we believe that it is better to avoid dangling qualifiers, if it is possible to do so without adding to question complexity. In this example, the overarching problem for both versions is excessive verbiage, regardless of how the question is structured. (Note that both versions required probing about equally, more than one third of the time-- a likely indication that respondents were not grasping the full intent of the question).

Example 3: Dangling Qualifier ("inclusive" type)

Similar structural issues emerge with qualifiers that ask respondents to *include* certain information. Consider the following question:

Q3.A During the past 12 months, that is since (month) a year ago, have you had your vision checked by an eye professional? Include optometrists or eye doctors who can prescribe glasses.

"Inclusive" qualifiers, unlike the "exclusive" ones, *always* have the potential to expand the respondents frame of reference. An exclusive qualifier may contain more information than a respondent needs, but every inclusive qualifier has the chance to alter the meaning of the question. Thus it is even more important that they are always read completely. As an alternative, we therefore restructured the question as follows:

Q3.B This question is about eye professionals such as optometrists or eye doctors who can prescribe glasses. During the past 12 months, that is since (month) a year ago, have you had your vision checked by an eye professional?

The response distributions for the two versions were almost identical. However, behavior code results (Table 4, below) show that there were fewer interruptions in Version B. The question seems to contain a minimum of verbiage, and there were no obvious comprehension problems in either version. Given that these potential pitfalls are avoided, it seems useful to convert the dangling qualifier into an introductory definition (as in Version B), to improve the chances that it is completely heard. We believe that this is generally a sound practice. On the other hand, it is possible that a dangling qualifier could be justified in some cases, e.g., for relatively complicated questions. Nevertheless, a better solution in those circumstances would be to reduce the overall complexity of the question.

Table 3: Behavior codes for Q2, times seen a health professional

Question	Minor read errors	Major read errors	% Read errors	% interrupt	Repeat question	Repeat cate-gories	Other probe	% probed	% R asks for clarif.
Q2.A	8	1	17.6%	9.8%	6	5	7	35.3%	5.9%
Q2.B	11		18.0%	1.6%	8	3	11	36.1%	13.1%

Table 4: Behavior codes for Q3, visit to eye professionals

Question	Minor read errors	Major read errors	% Read errors	% interrupt	Repeat question	Repeat cate-gories	Other probe	% probed	% R asks for clarif.
Q3.A	9	3	19.7%	14.8%	3		1	6.6%	1.6%
Q3.B	7	1	18.4%	2.0%	2		2	8.2%	0.0%

Example 4: Dangling qualifier ("inclusive" type)

The following question also employs an inclusive dangling qualifier, this time in order to expand the definition of "dentist." Once again, the potential problem is that respondents will interrupt or "tune out" the qualifier once they reach the question mark.

Q4.A About how many months has it been since you last saw or talked to a dentist? Include all types of dentists, such as orthodontists, oral surgeons, or all other dental specialists, as well as dental hygienists.

Originally, we constructed an alternative that re-worked the same words into a different structure (e.g., "This question is about dentists, including orthodonists, oral surgeons..."). However, cognitive interviewing led us to believe that the *wordiness* of the question may have been its most significant liability. Restructuring the same words did not seem to improve this problem. We therefore abbreviated the question as follows:

Q4.B About how many months has it been since you last went to a dentist office for any type of dental care?

The assumption behind Version B is that it is possible to convey the full meaning of Version A with fewer words. Version A focuses on the word "dentist," but the question is really about receiving dental care from any dental professional. Refocusing the question on "care" eliminates the need for detailed definitions.

Again, the response distributions for the two versions were almost identical. However, it is clear from behavior coding results (Table 5) that Version B is easier to administer, with no respondent interruptions, no reading errors, and fewer interviewer activities in general. (The equal-- and relatively high-- amount of probing associated with each question is probably attributable to the fact that neither question provides a specific range of responses.) Once again, the problem with Version A has less to do with the structure of the question than the sheer amount of words included within. There is no evidence that the alternative version conveys any less information than the first, and it is clearly easier to administer.

Discussion and Conclusions

Without extending these few examples too far, we do feel that our results offer a few ideas of general interest to questionnaire designers. We suggest that restructuring the words within complex questions has relatively minor payoff. Rather, the majority of cognitive burden comes from two sources: unwarranted assumptions embedded within the question, and excessive verbiage. Eliminating unwarranted assumptions can significantly change the response distributions for some questions (we believe making them more accurate); also, respondents can become confused when faced with too many concepts or too many details all at once. This point seems obvious, but even a casual look at major national surveys shows that it is often overlooked.

Many survey questions contain verbiage that pushes respondents' working memory to the limit (and sometimes beyond). Simplifying the questions seems to generate similar response distributions, while reducing the amount of interviewer effort required to obtain an adequate response. When these cognitive stumbling blocks are removed, other challenges posed by the questions become much more manageable.

Tackling these major problems-- unwarranted assumptions and excessive verbiage-- should be of the highest priority. This can often be accomplished through simplification, or the use of multiple questions rather than one very complex item. Restructuring survey questions without resolving such difficulties is unlikely to be of substantial benefit.

However, once these problems are resolved, modest improvements may be possible if questions are optimally structured. For example, we believe that "dangling qualifiers" should generally be avoided-- if this is accomplished while reducing the overall verbiage. It is also worth noting that changes in question structure can

Table 4: Behavior codes for Q4, recent dental care

Question	Minor read errors	Major read errors	% Read errors	% interrupt	Repeat question	Repeat categories	Other probe	% probed	% R asks for clarif.
Q4.A	4	2	9.8%	9.8%	1	7	8	26.2%	4.9%
Q4.B	0	0	0.0%	0.0%	1	3	9	25.5%	2.0%

influence the particular type of difficulties that are identified during question administration. For example, danglers may create more interruptions, but long definitions presented earlier in the question may require more interviewer probing and repeats of questions—which may be worse. Hopefully these observations can be of use to those who develop survey questions on complex topics.

References

Clark, H. and Schober, M. (1992). "Asking Questions and Influencing Answers." In Tanur, J. (ed.), Questions About Questions. New York: Russell Sage Foundation.

Fowler, F., Beatty, P., and Fitzgerald G. (in press). "When Two Questions Are Better Than One." Proceedings of the Section for Survey Research Methods, American Statistical Association, 1999.

Fowler, F. and Cannell, C. (1996). "Using Behavioral Coding to Identify Cognitive Problems with Survey Questions." In Schwarz, N. and Sudman, S. (eds.), Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research. San Francisco, CA; Jossey-Bass.

Fowler, F., and Mangione, T. (1990). Standardized Survey Interviewing: Minimizing Interviewer-Related Error. Newbury Park, CA: Sage.

Schwarz, N. and Hippler, H. (1991). "Response Alternatives: The Impact of Their Choice and Presentation Order." In Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S. (eds.) Measurement Error in Surveys. New York: John Wiley and Sons.