Experiments on the Structure and Specificity of Complex Survey Questions

Paul C. Beatty [1], Carol Cosenza[2], Floyd J. Fowler, Jr.[2]
National Center for Health Statistics, Centers for Disease Control and Prevention[1]
Center for Survey Research, University of Massachusetts-Boston[2]

Abstract

Survey questions must often present specific and complex information to respondents. Some questions include terms with meanings that are not completely self-evident, which need to either be defined or illuminated through examples. Questions may also include detailed instructions about factors that should be considered or not considered in responding, as well as very specific answer categories. In considering these issues, questionnaire designers must decide how much information to present within a single question and how to structure that information. This paper presents results from several split-ballot experiments that were designed to provide guidance on these issues. We identified questions that included complex concepts, definitions, or response categories and constructed alternative versions that varied either question structure, or the manner in which complex concepts were defined or explained. These alternative versions were fielded via split ballot in an RDD telephone survey (n=454). Most interviews were also tape recorded and behavior-coded. We evaluate whether these design decisions had a significant impact upon response distributions or respondent behaviors.

Keywords: questionnaire design and evaluation; split-ballot experiments; behavior coding

1. Introduction

Even a casual examination of the questionnaires used on major surveys reveals that many questions are quite complex. This complexity is primarily driven by designers' needs to obtain very specific pieces of data (e.g., the number of times a respondent has spoken with a primary care doctor in the last year, excluding specialists, but including contacts made over the phone or otherwise not in person). Survey costs often prompt researchers to obtain these data points using as few questions as possible, making it necessary to include definitions, explanations, qualifying information (e.g., include or exclude certain things while answering), and specific response options within a single question.

Assuming that all of this information is in fact important, and needs to be conveyed within a single question, we may have several options regarding how to structure this information within a question. In some other cases, general explanations may be functionally equivalent to detailed definitions or examples that are meant to give respondents an exhaustive and airtight frame of reference.

While some guidance from experimental research is available regarding these decisions (see Sudman, Bradburn and Schwarz, 1996, for a review), recommendations on the structure and specificity of complex questions are more commonly driven by general principles, past experience, or common sense. At other times, guidance is unavailable or even contradictory. This paper presents results of some research that was designed to provide additional evidence and guidance regarding the design of complex survey questions.

2. Methods

We selected a number of examples of complex questions from major federal health survey questionnaires. Some were actually administered on surveys such as the National Health Interview Survey, and others had only been proposed in draft form for inclusion. For each of these questions, we constructed an alternative version that was designed to obtain the same information. Some alternatives used almost identical words, but varied the question structure. For example, some questions provide "qualifiers" after the formal question has been asked; an alternative version incorporated the qualifier into an introductory statement *prior* to the question. Other alternatives included simplified definitions or alternative means of illustrating concepts (e.g., replacing a list of examples with a general definition). In this paper we will discuss five separate experimental manipulations.

Study procedures were similar to those reported by Beatty, Fowler and Fitzgerald (1999) in another split-

The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention, nor those of the University of Massachusetts-Boston.

ballot study. Question alternatives were embedded in one of two questionnaire instruments. These questionnaires were administered via an RDD telephone survey (n=454) conducted by the University of Massachusetts—Boston. Because the purpose of the study was split-ballot experimentation (including a number of experiments not covered in this paper) and not the creation of population-based estimates, we accepted any adult from contacted households to serve as the respondent. There was no conversion attempts on initial refusals and callbacks were minimal.

In each experiment, we were interested in whether the survey responses differed across question versions. Of course, it is not always possible to determine which of two distributions is more accurate, but we generally had a priori hypotheses about how we expected the experimental manipulations to affect responses.

In addition, we tape recorded the interviews whenever respondents gave permission to do so. These recordings were behavior-coded using procedures summarized in Fowler and Cannell (1996). The purpose of behavior coding is to help us understand how easy the questions are to administer. A number of different codes were used, but this paper will focus on a select few: the number of times that the initial response to the question was inadequate; the number of times the respondent interrupted the question before it was fully read; the number of times some sort of probing was required to get a response; and the number of times the respondent asked for clarification, repeat of question, or similar assistance.

It is worth noting that behavior coding has been used for many years to identify problematic questions, i.e., those that are difficult to administer in a standardized manner. However, the use of behavior coding to compare administration of alternative versions of the same question is relatively new, and poses some challenges: generally the differences we are considering are subtle, and the problems captured by behavior codes may only affect a small proportion of the overall sample. We generally did not expect difference to be statistically significant and were looking for trends—however, we did perform significance tests and include those results below.

Note that the functional sample sizes reported in the results below are generally less than the full sample of 454 respondents, for several reasons: sometimes screener questions preceded the experimental items, diverting some respondents; sometimes we actually administered three variants of the questions, but for simplicity of presentation focused on the single most informative comparison here; and, behavior coding was only performed on a subset of the entire sample.

3. Experiment 1: Question Structure

Consider the following survey question:

> "What kind of place do you usually go for routine medical care? Is it a doctor's office, clinic or health center, hospital emergency room, hospital outpatient clinic, or some other place?"

This question includes an exact set of response categories for respondents. Ideally, we want respondents to hear the entire list of response categories, identify the one that best suits their situation, and report that single response to the interviewer. We do not want respondents to answer before hearing the entire list, because a more appropriate choice may be provided for them after they respond. We also want respondents to respond using one of these exact response categories—this minimizes the amount of interpretation or probing required by interviewers to obtain a quantitative response, and maximizes standardization. However, there is ample anecdotal evidence that respondents often interrupt before reaching the end of the list. Other times, respondents provide responses that do not perfectly conform to one of the response choices. Presumably this is because when we reach the question mark, respondents begin the process of formulating their answer, and some respondents are not paying attention to the information we provide about the specific response choices we would like for them to use. Given that, is there an alternative way to structure the question that might improve its performance?

The structure of the question above is very traditional: question followed by responses. One alternative would be to present the eligible responses before actually administering the question. Such a wording might look like this:

> "People can get routine care in different places, including a doctor's office, clinic or health center, hospital emergency room, hospital outpatient clinic, or some other place. Which of those places do you usually go to for routine medical care?"

In this version, the question mark is actually at the end of the question. At that point, respondents have heard everything that we want to present, and there is very little potential for them to interrupt before then. On the other hand, there are some reasons to dislike this structure. It is potentially more awkward. Are respondents actually able to keep response categories in mind and make use of them before they know how these categories will be applied? Is it possible that this structure actually causes more administration problems than it fixes?

As we suspected, Table 1 shows that the first version of the question was interrupted often (28% of the time). Generally, when interviewers were interrupted, they failed to completely read the response categories.

Table 1: Behavior codes for Experiment 1

	V1	V2	signif
Interruptions	28.2%	2.9%	p<.01
First response inadequate/help	14.7%	19.4%	n.s.
Sample	(n=156)	(n=175)	

However, slightly more respondents initially gave an answer that was inadequate (or was a request for help or clarification) for the revised question than the original one. The difference is not statistically significant, but is in a direction that makes sense if respondents had trouble using response categories before they heard the core of the question. (Other behavior codes we considered did not vary across questions to any notable degree.)

Table 2 presents actual responses to the questions. If respondents fail to fully consider the response categories in the original version (V1), then we might see a migration to latter responses in the alternative (V2), since respondents will have heard them all. However, the response distributions to the two versions are identical. If there is a problem with over-reporting in the early categories with V1, then V2 does not correct this tendency.

Table 2: Responses for Experiment 1

	V1	V2
Doctor's office	78.9%	78.0%
Clinic/health center	14.9%	12.4%
Hospital outpatient clinic	5.2%	5.3%

(differences not significant at .05 level)

We should note that one answer category ("doctor's office") is the overwhelming favorite in both versions. So for this question, it may be that interruptions are not particularly important—respondents only do so when they very clearly hear a category that applies to them. For this question, there may in fact be little chance that a better response lies ahead; however, it is possible that the results could be different for a different set of response categories. As of this writing, we are in the field with another experiment using alternatives of a different question—one for which we judged that it would be more important for respondents to hear all response categories before answering.

In this case, the disadvantages of V2 may outweigh its potential benefit. While V2 forestalls interruptions, this seems to have no substantive effect on responses. The restructuring also comes at a cost of a (modest) increase in administration problems. Rather than adopting a potentially counter-intuitive question structure, more gains might be made in interviewer training to ensure that all response categories are read before accepting a final response.

4. Experiment 2: Presentation of Qualifiers

Response categories are not the only part of a question that can dangle after the question mark. Survey questions also include qualifiers, as in the example below:

> "How many months has it been since you last talked to a medical professional about your own health? Include in-person visits, telephone calls, or times you were a patient in a hospital."

This qualifier contains potentially important information, asking respondents to consider situations that may not be obvious from the core question alone. As in the example above, we might be concerned that respondents would either interrupt before considering this information, or pay minimal attention to it, as they are already engaged in the process of providing a response. If so, the respondent could fail to include certain visits. Fowler (1995) advises against such question constructions for that reason.

An alternative, of course, would be to move the qualifying material prior to the question mark. For example:

> "People talk to medical professionals in person, over the phone, or as patients in a hospital. Including any of those, how many months has it been since you last talked to a medical professional about your own health?"

This alternative version uses virtually identical words, but is structured differently. While this has some possible advantages, we again wonder whether the alternative structure is actually more confusing than the original. That is, can respondents actually make use of the qualifier when it is presented prior to the core question?

Responses to the question include slightly more recent reports of doctor contacts when the qualifier is read first

(V2). Although the difference is not statistically significant, it is in the direction that is consistent with respondents including more potential contacts in their answers. Perhaps more puzzling is the fact that more respondents failed to ever provide an answer to the original than the alternative. This difference is significant at the .05 level and suggests that for whatever reason, more respondents were unable to make sense of the original question with the dangling qualifier.

Table 3: Results for Experiment 2

Qualifier:	V1 (after q)	V2 (begin of q)	signif
Mean months since last dr. contact	5.2	4.7	n.s.
No answer given	3.6%	0.9%	p<.05
Interruption	4.4%	0.0%	p<.01
Initial resp inadeq	22.0%	17.1%	n.s.
Sample	(n=182)	(n=193)	

Several behavior coding results also favor the alternative version of the question. Although interruptions are not as frequent as in the previous experiment, they are present for V1, and V2 eliminates them. Also, the percentage of initially inadequate responses is lower for V2 (a five percent gain, although not statistically significant).

Thus, there may be advantages to incorporating qualifying information into questions, rather than allowing them to dangle afterwards. This advice may be limited to situations where the qualifier is simple enough to be incorporated in a reasonable manner, but the general principle may be worth considering. Whereas it may not make as much sense to present response categories first, qualifiers are substantially different and may be useful in "setting the stage" for respondents' thought processes.

5. Experiment 3: Detailed Definitions

Survey designers often want to load up questions with detailed definitions, or an exhaustive set of examples, such as this one:

> "A health provider could be a general doctor, a specialist doctor, a nurse practitioner, a physician assistant, a nurse, or anyone else you would see for health care. In the last 12 months, not counting the times you needed health care right away, did you make any appointments with a doctor or other health provider for health care?"

The reason for this level of specificity is concern that respondents might not consider all of these situations unless the question mentions them explicitly. However, the question is long and difficult to administer. It is also potentially challenging for respondents to keep all of the details straight. In an effort to present all possible variations to respondents, questionnaire designers may overload working memory to the point that many details are forgotten anyway, or that the central premise of the question is lost amid the many examples.

It also seemed to us that the details in this case were not particularly important. The examples of "health care provider" are pretty straightforward. It seems unlikely that anyone would exclude contact with a nurse on the grounds that one would not be considered a health care provider. It is possible that the extra words stimulate memories, but it seems more likely that respondents' memories center on medical events rather than the particular type of provider seen.

We suggested that a shorter statement could convey the same information with less burden and less chance to lose track of more significant details: "a health provider is anyone you would see for health care." Nevertheless, some researchers have resisted such changes on the grounds that the simplification will reduce reports.

Table 4: Results for Experiment 3

Definition:	V1 (long)	V2 (short)	signif
Results– Saw a provider	69.2%	73.5%	n.s.
Sample	(n=224)	(n=230)	
Initial response inadequate	12.6%	5.7%	p<.05
Sample	(n=182)	(n=192)	

Contrary to this concern, the shorter form of the question actually increased reports of provider visits. The difference is not statistically significant, but is in the opposite direction of what would be expected if respondents failed to consider some provider types. In addition, the proportion of respondents who gave an initially inadequate response was cut in half, which is statistically significant at the .05 level.

Thus, for this question, the simplified version seems to be a reasonable alternative. This is not to say that all detailed definitions are superficial—there may be times when a high level of detail is necessary to convey key concepts. This does not appear to be one of those times, and a simplified question may be simpler to administer to respondents as well.

6. Experiment 4: Examples vs. Definitions

Many survey questions use a series of examples to illustrate a concept, as in this question:

> "The next question is about strenuous tasks done around your home. By strenuous tasks, we mean things like shoveling soil in the garden, chopping wood, major carpentry projects, cleaning the garage, scrubbing floors, or moving furniture."

Whereas the question in the previous experiment was designed to provide a near-comprehensive list, this question focuses on several concrete examples that are designed to express a range of possibilities. However, we have several concerns with this approach. Although the examples are intended to express a range of options, we believe that they often do exactly the opposite and focus respondent attention on a few examples that may or may not be representative of the key concept. In this case, it is not clear that these are good examples of strenuous tasks around the home. Most of them could be strenuous or not strenuous depending on the energy and amount of time committed to performing them. The potential for confusion about meaning seems to be high. Indeed, cognitive interview results have shown that a number of respondents were perplexed by the examples provided.

Partially in response to such findings, we have advocated the use of general definitions as an alternative to selective examples whenever possible. These have the advantage of conveying useful information without limiting the frame of reference. In this case, we constructed the following definition: "… By strenuous, we mean any activity that made you feel tired if you did it for 15 minutes or more…"

Table 5: Results for Experiment 4

	V1 (example)	V2 (definition)	signif
Responses (times/month)	5.2	5.9	n.s.
Initial response inadequate/ req for help	13.1%	24.1%	p<.01
Sample	(n=191)	(n=183)	

In our experiment, we expected the revision to lead to both higher reports of strenuous behavior and better performance. The alternative did lead to slightly higher reports of frequency—again, not to a statistically significant degree, but suggesting the possibility that respondents might have considered a broader range of activities in answering.

What is especially surprising, however, is that the alternative performs quite poorly on several administrative measures. The original led to inadequate initial responses 13% of the time, which is fairly high. However, the alternative is far worse, leading to inadequate initial responses almost twice as often (24%). The alternative also required significantly more probing and generated significantly more requests for clarification at the .05 level. All in all, this is quite poor performance.

We still believe that the problems identified with the original question are legitimate. However, it is clear that the alternative is measurably inferior. Upon closer evaluation, this is most likely due to the fact that the alternative mixes time and level of activity. It might be difficult to answer if someone performed an activity that they considered strenuous but for less than 15 minutes, or an activity that only became strenuous long after 15 minutes had passed, and so on. Perhaps another definition would work better, such as "by strenuous, we mean tasks or chores around your home that made you very tired by the time you finished them." We hope to evaluate this in a subsequent experiment.

7. Experiment 5: Cognitive Interview Findings

Cognitive interviews are often the basis for our recommendations for changes. Here is one example. We tested this question:

> "In the past 12 months, how many times have you seen or talked on the telephone about your physical or mental health with a family doctor or general practitioner?"

A number of cognitive interview participants answered "zero" to this question, but these responses made little sense given their answers to previous questions that strongly suggested they had seen a physician within the last few months. Probing revealed the nature of the problem: when we read the question, participants thought the question asked specifically about times they *talked on the telephone* to their doctors. The structure of the question and relative weight of the phrase "talked on the telephone" dominated the attention of several participants.

For a revision, we proposed the following:

> "In the past 12 months , how many times have you seen or talked with a family doctor or a general practitioner about your physical or mental health?"

The revision is almost identical, except that it drops the word "telephone" and restructures the question slightly (whereas the original reads "talked about your health with a doctor," the alternative reads "talked with a doctor about your health")—this seems to flow more naturally. We thought this was a clear improvement, although some researchers expressed reservations about dropping the word "telephone," thinking that it would lead to underreports. We disagreed, and suggest that "talked with" is sufficient to cover telephone consultations, while minimizing the potential distraction caused by including the additional words.

Table 6: Results for Experiment 5

	V1 (original)	V2 (alternate)	signif
Mean contacts	3.5	3.8	n.s.
"Zero" responses	27.3%	18.2%	p<.10
Probed	13.3%	10.7%	n.s.
Req clarif/help	21.7%	15.7%	n.s.
Sample	(n=120)	(n=121)	

In our experiment, the revision did not cause reports to go down—in fact, they went up slightly. More significantly, the percentage of respondents who said "zero" dropped on the alternative. This is consistent with our cognitive interview findings. If respondents become fixated on the word "telephone," they would be more likely to answer zero ("I haven't talked to a family doctor on the telephone"). Without the word telephone, they are more likely to report contact with a family doctor. Not only are the responses to the alternative more plausible, but the alternative is also slightly easier to administer according to a number of behavior codes.

8. Conclusions

Experimental comparisons of both responses and response behaviors can be informative—such experiments allow us to compare not only how respondents answered but also how straightforward it is to administer alternate question forms. Collectively, the results of such studies often reveal tradeoffs involved in these decisions rather than "slam dunks" clearly favoring one or the other on all accounts. Still, the preponderance of evidence may land in favor of one particular version of a question.

The results presented here constitute a subset of our findings from a broader array of experiments, several of which deal with structuring of fixed material within a question. That is, given that we need to convey a certain set of words to respondents, what is the best way to structure the material?

Based on results of several studies (one of which is presented here), we suggest that the common use of "dangling" qualifiers should be reduced when possible. These qualifiers in effect change the rules of the game after the starting gun has gone off, giving respondents additional information to consider after the response task has been posed. If it can be done in a reasonable manner, it makes sense to incorporate this information up front. In contrast, we cannot make a general case for providing response categories in advance of core survey questions. While we are troubled to see that categories are often interrupted before they are read, we have not yet found evidence that adopting a counterintuitive construction produces any gain in data quality.

More generally, most of the differences observed from manipulating question structure are modest. In evaluating some complex questions, we have concluded that many are plagued by unwarranted assumptions about respondent experiences, inappropriate response categories, or other problems. The effects from these problems are probably much greater than effects from question structure. That being said, we think our recommendations can still make a contribution toward overall question clarity and will minimize response burden.

Several of our experiments have suggested that certain superfluous details of questions can be dropped without adverse effects. We want to be cautious and not over-extend these results. Clearly, some details can be critically important. But when cognitive interviewing suggests that excessive words are causing problems, or when a reasonable case can be made for simplified substitutions, it may be worthwhile to consider changes. We found modest evidence that such changes led to easier questions and may have led to more accurate responses as well. We also provided at least one cautionary tale that the identification of a problem in cognitive interviewing does not necessarily mean that the proposed solution constitutes an improvement.

This study had several limitations. First, it was restricted to easily-accessible telephone respondents. That should not limit the validity of the experimental comparisons, but such respondents may not be completely typical. Also, some of our findings may not apply to other modes of administration. We were also limited by small sample

sizes that may have lacked the statistical power to identify true differences. Some findings were more suggestive than definitive. In any case, more work will follow. More analysis remains to be performed on these experiments, and additional experiments building from these are currently in the field.

References

Beatty, P., Fowler, F.J., and Fitzgerald, G. (1999), "Construction Strategies for Complex Survey Questions." Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 973-978. Alexandria, VA: American Statistical Association.

Fowler, F.J. (1995). Improving Survey Questions: Design and Evaluation. Thousand Oaks, CA: Sage.

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

Sudman, S., Bradburn, N. and Schwarz, N. (1996), Thinking About Answers: The Application of Cognitive Processes to Survey Methodology. San Francisco: Jossey-Bass.