DIFFERENT RESPONDENTS INTERPRET ORDINARY QUESTIONS QUITE DIFFERENTLY[1]

Anna Suessbrick, New School University
Michael F. Schober, New School University
Frederick G. Conrad, Bureau of Labor Statistics
Michael F. Schober, Dept. of Psychology AL-340, New School University,
65 Fifth Ave., New York, NY 10003

Key Words: question clarification, data quality, conversational interviewing, measurement error, standardized interviewing

INTRODUCTION

Consider the first question in the Tobacco Use Supplement to the Current Population Survey (CPS): *Have you smoked at least 100 cigarettes in your entire life?* At first glance, it seems to consist of ordinary, non-technical words that should be easy for respondents to understand. However, as Belson (1981, 1986) observed, there can be substantial variability in people's interpretations of concepts in straightforward survey questions like this. For example, in one of Belson's studies 16% of respondents interpreted "you" in *How many hours of television do you watch each weekday?* to include other people, and 61% counted days other than the five weekdays. Our question is whether such conceptual variability is as widespread as Belson suggested, and more importantly whether it actually harms survey data quality.

The findings of our earlier laboratory experiments (Schober & Conrad, 1997, 1998; Schober, Conrad & Fricker, 1999) and field study (Conrad & Schober, 2000) suggest that conceptual variability can indeed affect data quality under certain circumstances, at least for a sample of questions about facts and behaviors excerpted from full-length government surveys. For example, respondents answering the Current Point of Purchase Survey question *Last year, did you purchase or have expenses for household furniture?* interpreted the question quite variably when they were answering about purchases of floor lamps, televisions, or appliances; some treated them as household furniture purchases, and others did not (Conrad & Schober, 2000; Schober & Conrad, 1997). Interpretation was far more uniform for straightforward purchases like end tables or sofas. In these studies, uniformity of interpretation—and thus data quality—could be increased dramatically when respondents were provided with clarification about the meaning of the words in the questions.

Here we examine the effects of conceptual variability on data quality in a full-length established survey with complex skip patterns, the Tobacco Use Supplement to the CPS. The Tobacco Use Supplement is sponsored by the National Cancer Institute and administered by Census Bureau interviewers using Computer Assisted Telephone Interviewing (CATI) once a year, in most years since 1992, to all CPS households. It assesses respondents' current and previous smoking and tobacco use, as well as opinions about related topics. Respondents answer from twelve to thirty-six questions, depending on skip patterns. All respondents answer the initial behavioral filter question and a similar question later in the survey about *pipes, cigars, chewing tobacco, and snuff.* Only those respondents who have smoked answer additional behavioral questions, for example *Have you EVER stopped smoking for one day or longer because you were TRYING to quit smoking?* All respondents then answer all the opinion questions, which include questions like *In restaurants, do you THINK that smoking SHOULD be allowed in all areas, allowed in some areas, or not allowed at all?*

Although the questions in this survey all seem quite straightforward, they all allow multiple interpretations. Has one "stopped smoking" if one cuts down during an illness? Should "restaurants" include outdoor seating areas and restrooms? Even *Have you smoked at least 100 cigarettes in your entire life?* might be difficult to answer for a respondent who isn't sure whether to include clove or marijuana cigarettes, cigarettes that have never been inhaled, or cigarettes from which only a puff or two were taken. A misinterpretation of this filter question could lead a respondent to answer the wrong questions later on.

Unlike in our earlier laboratory studies (Schober & Conrad, 1997, 1998; Schober, Conrad, & Fricker, 1999), respondents in the experiments reported here

[1] We thank Tom Nardone, Cathy Dippo, and Bill Mockovak at the Bureau of Labor Statistics; Kimberly Clark, Ron Tucker, Mike Haas, and Fran Faull at the Bureau of the Census; and members of the Psycholinguistics Laboratory at the New School University for their assistance. This material is based upon work supported by the National Science Foundation under grants No. SBR-9730140 and ITR-0081550. The opinions expressed are those of the authors and not of the Bureau of Labor Statistics.

answer about their own lives rather than fictional scenarios, and so we do not control the frequency of problematic circumstances like the purchases of floor lamps. Unlike in our earlier field study (Conrad & Schober, 2000), respondents here answer questions in the laboratory, rather than at home, so that we can measure their conceptualizations in greater detail. Unlike in any of the earlier studies, respondents answer opinion questions as well as factual ones.

EXPERIMENT 1

In Experiment 1 we examined (1) the variability in respondent interpretations of survey concepts and (2) the degree to which this variability affects responses. To do this we first administered the Tobacco Use Supplement by CATI. Interviewers in Hagerstown, MD, called respondents in our laboratory and carried out a strictly standardized interview in which the interpretation of survey terms was left entirely up to respondents. Then respondents filled out two paper-and-pencil questionnaires. The first questionnaire assessed conceptual variability by determining the extent to which respondents' interpretations matched official survey definitions. The second questionnaire assessed how much the initial responses would change if respondents were provided with uniform concept definitions, relative to how much responses would change without subsequent definitions.

Survey concept definitions. We used the sponsors' definitions when they existed. For those concepts for which the sponsors had not provided definitions, we defined the concepts to conform with the survey designers' intent, to the extent that we had evidence of it. When there was no evidence, we defined the concepts in ways that seemed reasonable to us. An example of a sponsor-provided definition is: "*Past 12 months* means 12 months from today, NOT from the first of the month and not just the last calendar year." An example of a definition we created is: "By *smoked* we mean any puffs on any cigarettes, whether or not you inhaled AND whether or not you finished them."

Participants. Fifty-three paid respondents (27 Female, 26 Male) were recruited, using newspaper advertising and word-of-mouth, from the New York City area and the New School University community. Their mean age was 33.4 years and they ranged in ethnicities and educational backgrounds. Ten interviewers (8 Female, 2 Male) were recruited from the Hagerstown, MD, Bureau of the Census telephone facility. Interviewers averaged 59.9 months of interviewing experience. Each conducted five or six interviews.

Interviewer training. Before the experiment was conducted, interviewers were trained on the survey concepts for about two hours. Interviewers studied the key survey concepts and then took a quiz, followed by a group discussion. Although these interviewers were not to provide definitions to respondents during the survey, concept training allowed interviewers to know when to probe and ensured comparability with future experimental conditions.

Following concept training, we provided additional training in the strictly standardized interviewing techniques from the CPS training manual, conforming to procedures advocated by Fowler & Mangione (1990), among others. In a standardized interview, interviewers are instructed to read each question exactly as worded and to probe non-directively, either by re-reading the entire question; requiring respondents to provide a codable response (e.g., *I need a number*); re-presenting the complete list of response alternatives; or encouraging respondents to interpret questions for themselves (e.g., *Whatever "fairly regularly" means to you* or *We need your interpretation*).

Conceptualization questionnaire. In the first paper-and-pencil questionnaire after the CATI interview, respondents were asked their interpretations of the survey concepts. For example:

Have you smoked at least 100 cigarettes in your entire life?

When you answered this question, did you interpret "smoking" to include: *(Pick one)*
( ) Only puffs that you inhaled
( ) Any puffs, whether or not you inhaled

How did you interpret "cigarettes"? *(Pick all that apply)*
( ) Cigarettes that you finished
( ) Cigarettes that you partially smoked
( ) Cigarettes that you only took a puff or two from

Did you interpret "cigarettes" to include: *(Pick all that apply)*
( ) Manufactured cigarettes
( ) Hand-rolled cigarettes
( ) Marijuana cigarettes
( ) Cigars
( ) Clove cigarettes
( ) Something else. Specify: ______________

Response change questionnaire. This questionnaire assessed the extent to which respondents' variable interpretations would actually affect responses. In this self-administered "re-interview" respondents answered exactly the same questions they had answered in the original interview. Half the respondents (27) were asked to use the official definitions in answering the questions; the other half (26) were presented with the identical questions without definitions. By comparing response change when definitions were provided with response change when definitions were not provided, we could determine if the conceptual variability assessed by the conceptualization questionnaire actually changed responses. That is, if response change is greater when definitions are provided, this suggests that there was sufficiently variable—and unintended—interpretation in the original interview to affect responses. Here is a sample item from the self-administered reinterview with definitions:

> *Have you smoked at least 100 cigarettes in your entire life?*
>
> Definition:
> - We want you to include any puffs on any cigarettes, whether or not you inhaled AND whether or not you finished them.
> - We want you to include hand-rolled cigarettes as well as manufactured ones, and tobacco cigarettes with additives like cloves.
> - We DON'T want you to include cigars or non-tobacco cigarettes, like marijuana cigarettes.
>
> Keeping this definition in mind, how would you answer this question?
> - Yes
> - No

The comparable item from the self-administered reinterview without definitions simply presented the question and the response alternatives.

RESULTS

Conceptual variability. We calculated for each respondent the percentage of their concept interpretations (for questions they had answered) that matched the survey definitions. Conceptual fit between respondents' interpretations and survey definitions was poor, averaging 39% overall. Fit for concepts in opinion questions, at 46%, was also poor, although it was reliably better than for behavioral questions, at 33%, $F(1,51) = 36.84, p < .0001$.

These findings might indicate simply that the survey definitions were unsound—that is, counterintuitive to respondents. This characterization of the data would be supported if most respondents agreed on one interpretation, regardless of whether it matched the survey definition. But this was not the case. Respondents' interpretations were not uniform but, rather, tended to be distributed among multiple interpretations, suggesting that the integrity of the definition is not at issue. For the 37 concepts in questions answered by all (95% or more) of the respondents, on average only 51.3% of respondents endorsed the majority interpretation (61.7% for behavioral concepts; 48.9% for opinion concepts). A striking example is the very first question of the survey, which contains the seemingly ordinary concepts "smoking" and "cigarettes" (see Tables 1 and 2).

Concept Interpretation of "Smoking"	Percentage of Respondents
Only puffs inhaled	46%
*All puffs, whether or not inhaled	54%

Table 1: Conceptual fit: Percentage of respondents who interpreted the concept of "smoking" as either only puffs inhaled or all puffs. An asterisk indicates the interpretation that corresponds to the survey definition.

Concept Interpretation of "Cigarette"	Percentage of Respondents
Only cigarettes you finished	23%
Cigarettes you finished or partly smoked	23%
*Even just one puff	54%

Table 2: Conceptual fit: Percentage of respondents who interpreted the concept of "cigarette" as cigarettes they finished, partly smoked, or took even just one puff of.

Respondents' understanding of the same concepts may vary considerably, which could undermine the comparability of their responses.

Response change. It appears that conceptual variability affected responses substantially. When respondents were given definitions in the "re-interview," the responses changed about twice as often as when they were not given definitions, $F(1,51) = 13.02$, $p = .001$. Response change was at least as great for opinion questions as for behavioral questions, with more response change for opinion questions, $F(1,51) = 4.35$, $p < .05$ (see Table 3).

Survey Item	No Definitions	Definitions
Behavioral Qs	5.0%	10.3%
Opinion Qs	6.5%	16.3%

Table 3: Response change with and without definitions for both question types.

These data demonstrate that respondents can interpret even the most seemingly straightforward questions quite differently. This seems to be as true of opinion questions as of factual questions. Experiment 1 is consistent with our previous findings that respondents provided with standard concept definitions interpret survey questions more uniformly, and thus provide more comparable answers.

But what is the best way to give definitions to respondents? In Experiment 1, respondents were given definitions in a self-administered laboratory questionnaire. In Experiment 2, we explore two different ways of providing respondents with definitions during the interview itself.

EXPERIMENT 2

Our earlier studies (Conrad & Schober, 2000; Schober, Conrad, & Bloom, 2000; Schober, Conrad, & Fricker, 2000) suggest that interviewers empowered to provide clarification can improve uniformity of interpretation. Here we contrast two techniques: Respondent-Initiated Clarification, in which the interviewer provides clarification during the interview if the respondent explicitly requests it, and Mixed-Initiative Clarification, in which interviewers can also offer clarification during the interview whenever they think the respondent needs it, even if the respondent hasn't explicitly requested it. (In our earlier papers, this latter technique has also been called "conversational" or "flexible" interviewing) Our earlier studies suggest that response accuracy will be greater for Mixed-Initiative Clarification, which increases the likelihood that the respondent will be given clarification.

In this experiment, interviewers trained to use either the Respondent-Initiated Clarification or Mixed-Initiative Clarification technique administered the Tobacco Use Supplement. As in Experiment 1, they questioned laboratory respondents by telephone. Unlike in Experiment 1, interviewers instructed the respondents to ask for clarification if they were at all unsure how to interpret the questions. Next, respondents, with paper and pencil, filled out the conceptualization questionnaire and the self-administered re-interview questionnaire from Experiment 1. The re-interview questionnaire always included the survey definitions.

Participants. The respondents were 51 paid participants recruited from the New York City area and the New School University community, with demographics comparable to the groups in Experiment 1 (21 Female, 30 Male, with a mean age of 32.4 years). Respondents were randomly assigned either to the Respondent-Initiated Clarification group (n=25) or the Mixed-Initiative Clarification group (n=26). Nine interviewers (7 Female, 2 Male) who had not participated in Experiment 1 were recruited from the Hagerstown, MD, telephone facility of the Bureau of the Census, averaging 59.1 months interviewing experience. They were randomly assigned to one of the two interviewing techniques and were roughly matched for interviewing experience. Each interviewer conducted five to six interviews, except for one interviewer, who conducted ten.

Interviewer training. As in Experiment 1, interviewers were trained on key survey concepts using a quiz and group discussion. Afterwards, the Respondent-Initiated Clarification interviewers were trained to clarify survey concepts only upon the respondent's explicit request. The Mixed-Initiative Clarification interviewers were trained to clarify concepts whenever they thought the respondent needed clarification and whenever the respondent requested it. The amount and type of training was the same as used in earlier studies (Conrad & Schober, 2000; Schober & Conrad, 1997).

RESULTS

Despite the interviewer's opening instructions to respondents encouraging them to request clarification whenever necessary, respondents almost never asked for clarification. Furthermore, despite the concept training, and unlike interviewers in our previous studies, interviewers here almost never offered clarification. We propose that it simply didn't occur to respondents or interviewers that their interpretations did not match, and therefore they did not recognize a need for clarification (Schober, 1999; Schober & Conrad, in press).

Conceptual variability. Given how little clarification was delivered in the CATI phase of the experiment, it is not surprising that conceptual fit for both groups in Experiment 2 was no better than for the two groups in Experiment 1. Respondents almost never received definitions during the telephone survey, so when they completed the conceptualization questionnaire their understanding of the survey concepts could not have been affected. And, again, although fit was better for opinion than for behavioral questions, $F(1,49) = 43.31$, $p < .0001$, it was still poor, averaging less than 50%: In the Respondent-Initiated Clarification group, conceptual fit averaged 32% for behavioral questions and 51% for opinion questions. In the Mixed-Initiative Clarification group, conceptual fit averaged 32% for behavioral questions and 46% for opinion questions. However, there was no reliable difference in conceptual fit between the two groups.

Additionally, the variability of concept interpretation continued to be substantial: for the 37 concepts in questions answered by all (95% or more) of the respondents from both experiments, on average only 51.8% of respondents endorsed the majority interpretation (63.4% for behavioral concepts; 49.1% for opinion concepts). Furthermore, overall, none of the four groups differed in the rate of conceptual variability before the self-administered re-interview, $F(3,144) = 0.02$, *n.s.*

Again, it appears that respondents' concept interpretations do not simply differ from the survey definitions but vary substantially from one another's.

Response change. Given how rarely anyone received a definition during the telephone survey portion of the experiment, we did not expect to see different rates of response change between the two groups. The rates were in fact comparable to response change for respondents who received definitions in the first experiment. And response change for all three groups receiving definitions in the self-administered re-interview was still greater than for the group that did not receive definitions in the re-interview (No-Clarification-No-Definitions), Helmert contrast (comparing the mean of the first group to the mean of the other three), $F(1,100)=18.07$, $p<.001$. Presumably this is the result of greater conceptual alignment between the respondents' interpretations and the survey definitions when definitions were provided (See Table 4).

	Response Change
No Clarification-No-Definitions (Exp. 1)	5.8%
No Clarification-with-Definitions (Exp. 1)	13.3%
Respondent-Initiated Clarification (Exp. 2)	13.4%
Mixed-Initiative Clarification (Exp. 2)	17.4%

Table 4: Response change: Rate of response change, by experimental condition.

Conceptual variability can lead to misinterpretation of questions, and this can be particularly costly when filter questions are involved; the wrong answer on a filter question can lead the respondent down the wrong survey path. Given this concern, we examined response change for the first question in the survey—*Have you smoked at least 100 cigarettes in your entire life?*—which largely determines the sequence of questions that follows. A full 10% of respondents (n=8) given a definition in the self-administered re-interview changed their answer from "yes" to "no" or from "no" to "yes." In contrast, no respondents in the group that didn't receive definitions in the re-interview changed their response. If we assume that the 10% rate of response change reflects incorrect interpretations of the first question, then a substantial number of respondents may have been asked inappropriate questions and not asked appropriate ones.

CONCLUSIONS

This study shows that respondents can interpret seemingly straightforward questions quite differently than intended, and quite differently from each other. The problem is not that the survey definitions are counterintuitive, but that respondents' interpretations of the question concepts are so variable. It appears that no single definition will conform to all (or even most) respondents' interpretations. The results also show that conceptual variability can affect the quality of the data not only for single questions but for the entire questionnaire by directing respondents along the wrong route through the instrument.

As in our previous studies, respondents, given definitions, can interpret survey questions more uniformly and, presumably, data quality can thus be improved. What's more, this appears to hold for questions about opinions as well as behaviors. Opinion questions, like those about behaviors, involve concepts whose interpretations can vary. While notions of response accuracy and measurement error are not directly applicable to opinion questions (Sudman, Bradburn, & Schwarz, 1996), researchers should be as sure as possible that their measurement of opinions is based on comparable interpretations by all respondents.

Clearly, some instances of mismatched conceptualizations have fewer consequences than others. For example, a smoker who never inhales, but interprets "smoking" as inhaling only, will answer *Have you smoked at least 100 cigarettes?* inaccurately and take the wrong path down the remainder of the survey—as 10% of the respondents in our study did. But a regular smoker of both clove and "conventional" cigarettes who did not include clove cigarettes in her interpretation of "cigarette" would still correctly answer "yes" to the question.

We propose that almost all survey concepts—even the seemingly ordinary ones—are open to multiple interpretations, just as concepts in everyday, unscripted language use are. This variability appears to be a consequence of differences in individuals' circumstances and perspectives. These differences are ordinarily negotiated in conversations through grounding (e.g., Clark & Brennan, 1991), an iterative process by which conversation does not progress until the participants agree that an utterance has been understood well enough for current purposes. Given the variability that appears to characterize survey concepts, we are skeptical that more thorough pretesting of survey concepts is the full solution to this puzzle. Rather, concepts could be discussed by interviewer and respondent until both agree that the respondent has interpreted them as the survey designers intend.

In our earlier experiments, Mixed-Initiative Clarification interviews have most efficiently aligned respondent and interviewer conceptualizations. In the current study, however, neither interviewer nor respondent initiated clarification. Respondents also didn't seem to give the sorts of cues we have seen correlate with the need for clarification (Bloom & Schober, 1999): *ums* and *uhs*, long pauses, restarts and repairs, and utterances other than answers. This suggests that both interviewers and respondents failed to recognize that their interpretations did not match, perhaps because the survey concepts seemed so straightforward. Respondent-Initiated Clarification and Mixed-Initiative Clarification interviews will not help if respondents and interviewers don't recognize that clarification is needed and respondents don't ask for it.

The problem seems to be, then, how should respondents get the clarification they need in order to answer questions accurately? We recommend testing different ways of administering clarification to respondents. One strategy is to provide definitions along with all survey questions. While this approach strongly promotes uniform interpretation by all respondents, it leads to lengthy interviews in which unneeded definitions are sure to be provided and is likely to be annoying. A second strategy might be to decompose the survey questions into a series of questions about each concept in the original question, but this too may lead to long interviews if concepts have many components, because each leads to an additional question. A third strategy is encouraging respondents to ask for clarification more often. In one study using a computer-assisted self-administered survey interview (Conrad & Schober, 1999), we found that participants were significantly more likely to request clarification when they had been instructed that survey definitions were essential to answering the questions accurately. Perhaps this strategy could translate to a human-human interview.

A fourth strategy that needs to be further evaluated is conducting Mixed-Initiative Clarification interviews with interviewers who understand just how variable respondents' interpretations can be. In the current study, our interviewers reverted to the standardized approach that they were used to using. Implementing productive Mixed-Initiative Clarification interviews, then, may require either more extensive training of professional interviewers or enlisting novice interviewers who have never been trained in standardized (or other) survey interviewing techniques.

REFERENCES

Belson, W.A.(1981). *The design and understanding of survey questions.* Aldershot: Gower.

Belson, W.A. (1986). *Validity in survey research.* Aldershot: Gower.

Bloom, J.E., & Schober, M.F. (1999). Respondent cues that survey questions are in danger of being misunderstood. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 992-997. Alexandria, VA: ASA.

Clark, H.H. & Brennan, S.E. (1991). Grounding in communication. In L.B. Resnick, J.M. Levine & S.D. Teasley (Eds.) *Perspectives on socially shared cognition* (pp. 127-149). Washington, DC: American Psychological Association.

Conrad, F.G., & Schober, M.F. (1999). A conversational approach to text-based computer-administered questionnaires. In *Proceedings of the 3rd International Conference on Survey and Statistical Computing* (pp. 91-101). Chesham, UK: Association for Survey Computing.

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

Fowler, F.J., & Mangione, T.W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error.* Newbury Park, CA: SAGE Publications.

Schober, M.F. (1999). Making sense of questions: An interactional approach. In M.G. Sirken, D.J. Hermann, S. Schechter, N. Schwarz, J.M. Tanur, & R. Tourangeau (eds.), *Cognition and survey research* (pp. 77-93). New York: John Wiley &Sons.

Schober, M.F., & Conrad, F.G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly, 61*, 576-602.

Schober, M.F., & Conrad, F.G. (1998). Response accuracy when interviewers stray from standardization. *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp. 940-945). Alexandria, VA: ASA.

Schober, M.F., & Conrad, F.G. (in press). A collaborative view of standardized survey interviews. In D. Maynard, H. Houtkoop, N.C. Schaeffer, & J. Van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview.* New York: John Wiley & Sons.

Schober, M.F., Conrad, F.G., & Bloom, J.E. (2000). Clarifying word meanings in computer-administered survey interviews. In L.R. Gleitman & A.K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 447-452). Mahwah, NJ: Lawrence Erlbaum Associates.

Schober, M.F., Conrad, F.G., & Fricker, S.S. (1999). When and how should survey interviewers clarify question meaning? *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp. 986-991). Alexandria, VA: ASA.

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