WHEN AND HOW SHOULD SURVEY INTERVIEWERS CLARIFY QUESTION MEANING? [1]

Michael F. Schober, New School for Social Research
Frederick G. Conrad, Bureau of Labor Statistics
Scott S. Fricker, Bureau of Labor Statistics
Michael Schober, Dept. of Psychology AL-340, New School for Social Research, 65 Fifth Ave., New York, NY 10003

Key Words: Interviewing techniques, question clarification, data quality, conversational interviewing, standardized interviewing

INTRODUCTION

Most survey researchers advocate standardizing interviewing methods, but they differ in how best to do it. On one extreme, survey researchers argue that interviewers should present exactly the same words to all respondents, and they should leave the interpretation of those words entirely up to the respondents. Under this view (see Fowler & Magnione, 1990), interviewers, in the interest of collecting objective data, should avoid influencing responses—that is, they should refrain from any behaviors that might influence one respondent differently than another. If respondents ask what a question means, interviewers should not answer directly; rather, they should repeat the question, repeat the response alternatives, or tell respondents "whatever it means to you."

On the other extreme, researchers have argued that interviewers should standardize the *meaning* of survey questions, making sure that all respondents interpret the questions as the survey designers intended (Suchman & Jordan, 1990, 1991). To do this, interviewers may need to behave in a non-standardized "conversational" way, clarifying question meaning when respondents ask for help or seem to need it, asking respondents to describe their circumstances and then helping them choose the appropriate response, etc.

These alternative notions of standardization have been hotly debated throughout the history of large-scale surveys (Beatty, 1995). And the controversy continues: different organizations that consider themselves to implement standardization go about it in different ways. Some train their interviewers to adhere to the strictest standardization of wording, while others train interviewers to provide scripted clarification when respondents explicitly request it. Even within a single organization, interviewer behavior can vary substantially. For example, in one study (Schober & Conrad, 1999), some interviewers in the same organization never deviated from the strictest standardization, while others deviated for as many as 10 of 12 questions (83% of the time).

The fact that there can be such variability is problematic for organizations that maintain that their practice is standardized. Our primary question is: How does such variability affect the accuracy of responses? And how does it affect survey costs like interview length? To answer these questions, we try to disentangle different kinds of "standardized" interviewer behaviors and examine their effects on accuracy and costs.

In the laboratory study reported here, we systematically varied when and how telephone interviewers provided clarification to respondents. Interviewers either provided clarification only when respondents requested it or also when they believed respondents needed it; and they either read scripted definitions verbatim or used their own words to explain the definitions. The respondents answered fact-based questions from ongoing government surveys; they answered on the basis of fictional scenarios, so that we could directly assess response accuracy—the extent to which responses matched what the official government definitions for key survey concepts required. We compared response accuracy and survey costs across three groups of these respondents and two groups from an earlier study (Schober & Conrad, 1997), who came from the same population and answered the same questions in the same setting. In that study interviewers either provided no clarification at all (strict standardization) or provided unscripted definitions whenever they believed respondents needed them (what we call "conversational interviewing").

In the earlier study, we found that response

[1]We thank Cathy Dippo, Kimberly Clark, Laura Hahn Lind, Eduardo Vega, and Susan Brennan for assistance and advice. This material is based upon work supported by the National Science Foundation under grant No. SBR-97-30140 and by the Bureau of Labor Statistics. This paper reflects the opinions of the authors and not the Bureau of Labor Statistics.

accuracy was virtually perfect for both strictly standardized and conversational interviewing when respondents' fictional circumstances mapped onto the questions in a straightforward way. For example, if respondents were asked whether they had purchased household furniture, they were highly accurate—their answers matched the official definitions—when an end table had been purchased, irrespective of the interviewing technique. In contrast, respondents were quite inaccurate in standardized interviews when their fictional circumstances mapped onto the questions in a complicated way, for example, if the fictional purchase was a floor lamp, which might or might not be considered furniture. For such complicated mappings, accuracy in conversational interviews was almost 60% better. This improvement in accuracy came at a substantial cost: conversational interviews took over three times as long as strictly standardized interviews.

In the current study, we examine the effects of three additional sorts of standardized interviewing which are "intermediate" between the extremes of strictly standardized and conversational techniques examined in the Schober and Conrad (1997) study. These intermediate forms are worth examining not only because they correspond with what interviewers are trained to do at some organizations, but because they might produce substantial gains in response accuracy without the concomitant threefold increase in interview duration we saw in our earlier study (Schober & Conrad, 1997).

EXPERIMENT

We examined response accuracy and interview duration using exactly the same procedure as in the Schober and Conrad (1997) study, for comparability. The only difference was in interviewer training: interviewers were trained to implement three alternate forms of standardized interviewing.

Questions. All respondents were asked the same 12 questions as in the Schober and Conrad (1997) study. Four questions were about employment, adapted from the Current Population Survey (e.g., "Last week, did Chris do any work for pay?"); four questions were about housing, adapted from the Consumer Price Index Housing survey (e.g., "How many people live in this house?"); four questions were about purchases, adapted from the Current Point of Purchase Survey (e.g., "Has Alexander purchased or had expenses for college tuition or fixed fees?"). For each question, the sponsoring organization had developed official definitions for key concepts in the questions.

Scenarios. Respondents answered on the basis of the same fictional scenarios as in the Schober and Conrad (1997) study. These consisted of floor plans, work descriptions, and purchase receipts; the scenarios were never seen by the interviewers. For each respondent, half the scenarios described situations that mapped onto questions in a straightforward way, and the other half described situations that mapped onto questions in a complicated way; different respondents saw different scenarios for different questions. For example, a respondent asked "Last week, did Pat have more than one job, including part-time, evening, or weekend work?" would see either a scenario showing that Pat babysat for one family all week (straightforward mapping – this is clearly one job) or a scenario showing that Pat babysat for several different families (complicated mapping – does this count as one job or several?). The official concept definitions always clarified what the correct answer should be (in the second case, Pat has one job even if she has multiple employers).

As in the earlier study, the interviewers never knew, nor could they predict, what the correct answers were.

Participants. The 33 interviewers were professional Census Bureau interviewers (26 F, 7 M) calling from the Hagerstown, MD telephone facility. They averaged 62 months of interviewing experience, ranging from 2 to 165 months. There were no reliable differences in interviewing experience between the different interviewing groups. Each interviewer telephoned two respondents in the Bureau of Labor Statistics laboratory in Washington, DC.

The 66 paid respondents were recruited from the Bureau of Labor Statistics subject pool; they had responded to an ad in the *Washington Post.* They represented a range of demographic characteristics comparable to the range in the Schober and Conrad (1997) study; there were a total of 38 women and 28 men; 16 were black, 47 were white, 2 were Asian, and 1 was Hispanic, and there were comparable numbers of respondents from each category in each of the three groups. Respondents averaged 16.3 years of education, which is comparable to the educational level of respondents in the Schober and Conrad (1997) study; the three groups did not differ reliably in education.

Interviewer Training. All interviewers were first trained together on the key survey concepts for about an hour. This training included a quiz and group discussion; as the definitions could be quite long and complicated, we wanted to make sure that all interviewers understood them thoroughly, and as thoroughly as the interviewers in the Schober and Conrad (1997) study had.

Then interviewers received additional training in one of three interviewing techniques. Two of the

groups were trained to clarify the meaning of questions only if respondents explicitly requested clarification. We defined explicit requests fairly rigorously; interviewers were only to provide clarification when respondents asked explicit questions like "Does babysitting for two families count as one job or two?" or expressed their uncertainty directly, as in "I'm not sure what you mean by that question." They were not to provide clarification if respondents described their circumstances rather than answering the question, as in "Well, I babysit for two families," nor were they to provide clarification if respondents' answers merely sounded uncertain.

Of these two groups, one group was trained to read scripted definitions; this consisted in reading at least one full sentence of the definition at a time, up to reading the entire definition. The second group was trained to explain the concepts in their own words (although they were allowed to rely on reading parts of definitions if they preferred).

The third group was trained to provide clarification whenever they felt respondents needed it, whether or not respondents had explicitly asked for clarification. This meant that they were licensed to provide unsolicited clarification when they deemed it necessary. But in providing clarification they were to read scripted definitions verbatim (at least one full sentence at a time).

So, combined with the two conditions (strictly standardized and conversational) from the Schober and Conrad (1997) study, the three conditions in the current study lead to an experiment design in which *when* and *how* interviewers provide clarification is parametrically varied:

	WHEN		
Never	Only when explicitly requested	Also unsolicited	HOW
POQ 97			Scripted
		POQ 97	Paraphrase

Table 1. Experimental design

RESULTS

Before turning to the findings on response accuracy and interview length, we first needed to verify that interviewers had implemented the different interviewing techniques correctly. (We already knew that the interviewers in the two groups in the Schober & Conrad [1997] study had implemented their techniques appropriately). One way to do this is to examine transcripts of the interviews to see how often interviewers' clarification resulted from respondents' explicit requests for help. Interviewers trained to provide clarification only when it was explicitly requested did so more often (93.0% and 97.3% of the time) than interviewers who were also allowed to present unsolicited clarification (73.4% of the time). And interviewers trained to read scripted definitions presented exactly verbatim information reliably more often (88.6% and 89.8% of the time) than interviewers who were allowed to use their own words (72.2% of the time).

Response accuracy. As Figure 1 shows, across all five interviewing conditions respondents' answers were almost perfectly accurate (they matched what the official definitions required) for scenarios with straightforward mappings. For complicated mappings, our intermediate interviewing techniques produced intermediate response accuracy: reliably greater response accuracy than strictly standardized interviewing (contrast between strictly standardized and scripted only when explicitly requested, $F(1,104) = 4.62$, $p < .001$), but reliably less accurate responses than for our fully conversational interviewing (contrast between scripted unsolicited and scripted paraphrased , $F(1,104) = 3.18$, $p = .002$). The three intermediate groups did not differ reliably from each other. Table 2 shows response accuracy for complicated mappings as a function of when and how interviewers provided clarification:

Focusing on the data from the four groups where interviewers provided clarification, responses were reliably more accurate when interviewers provided unsolicited clarification than when they only responded to explicit requests for clarification, $F(1,84) = 16.15$, $p < .001$. Responses were marginally better when interviewers used their own words to clarify question meaning rather than reading scripted definitions, $F(1,84) = 3.85$, $p = .053$, but this was really because of the substantial increase in accuracy in the "fully conversational" case (*POQ* 97), interaction $F(1,84) = 5.94$, $p < .02$; we interpret this to mean that *how* interviewers provided help really didn't matter.

Of course, all cases where interviewers provided clarification produced reliably greater response accuracy than when they didn't give any clarification at

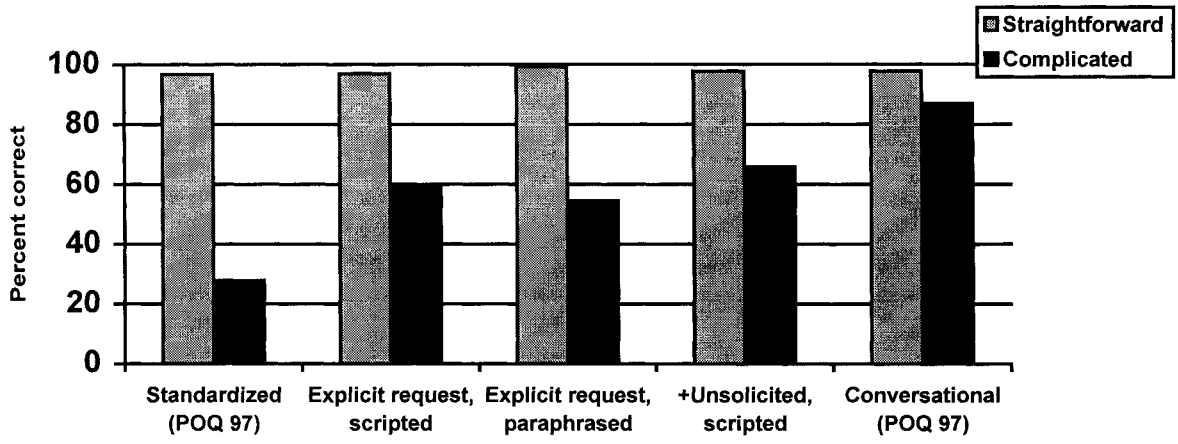

Figure 1. Response accuracy

	WHEN		
Never	Only when explicitly requested	Also unsolicited	HOW
28% (POQ 97)	59%	66%	Scripted
	55%	87% (POQ 97)	Paraphrased

Table 2. Response accuracy for complicated mappings

all. As Figure 2 shows, in the three "intermediate" conditions response accuracy was substantially greater whenever any clarification was provided.

Quality of clarification. The clarification that interviewers in the three "intermediate" groups provided was highly accurate. In the 272 cases where they provided clarification, they presented completely accurate information 98.9% of the time. And inaccurate information didn't always lead to inaccuracy; in the 9 cases where interviewers presented any inaccurate information, respondents still produced accurate answers for 8 cases.

Interviewers provided the information that respondents needed to hear 90.4% of the time. But interviewers sometimes also presented unnecessary parts of definitions, telling respondents more than they needed to hear. And in 26 cases (out of the 272) interviewers provided only irrelevant parts of definitions—that is, they provided accurate but unhelpful clarification. In these cases, as one might expect, respondents were not much more accurate (10 out of 26 cases, 42%) than when no clarification had been (27%).

Interview duration. Interviews took longer when interviewers provided more clarification. Table 3 shows the median interview duration in minutes for all five types of interviewing The three "intermediate" types of interviewing didn't reliably differ in how long they took, but they all took reliably longer than strictly standardized interviewing (contrast of standardized and scripted on demand $F(1,103) = 2.30$, $p < .025$), and they all took reliably less time than fully conversational interviewing (contrast of scripted unsolicited with paraphrased unsolicited $F(1,103) = 5.71$, $p < .001$).

In the Schober and Conrad (1997) study, the threefold increase in interview duration for conversational interviews wasn't merely because interviewers spent more time clarifying complicated mappings; they also spent a great deal of additional time discussing straightforward mappings, which really didn't need to be discussed. Figure 3 plots how much time was spent in all five kinds of interviews on complicated vs. straightforward mappings, as measured by the number of words spoken by interviewer and respondent together per question. As the figure shows, when interviewers provided clarification only when explicitly requested, the amount of time spent on straightforward mappings wasn't much greater than in strictly standardized interviews, while the amount of time spent on complicated mappings increased. In the condition where interviewers could also provide unsolicited clarification, the amount of "unnecessary" time spent on straightforward mappings increased, although not nearly so much as in the fully conversational interviews.

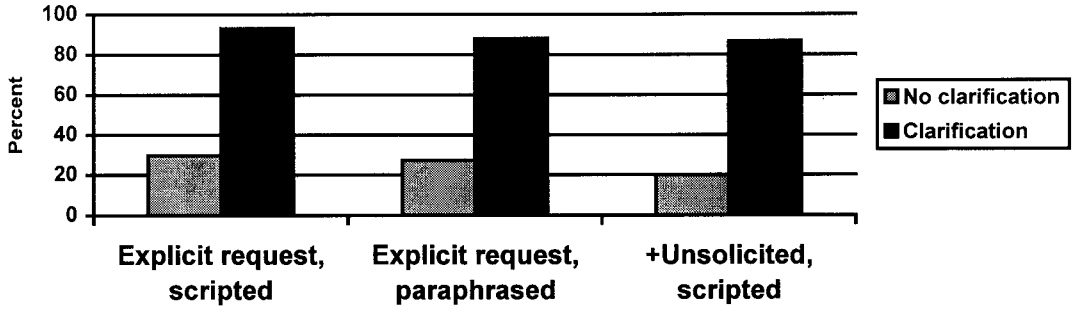

Figure 2: Response accuracy for complicated mappings, when clarification was provided

	WHEN		
Never	Only when explicitly requested	Also unsolicited	HOW
3.41 (POQ 97)	7.83	7.62	Scripted
	7.43	11.47 (POQ 97)	Paraphrased

Table 3. Median interview duration in minutes, for all five types of interviewing

Focusing on the four groups where interviewers clarified question meaning, we see that interviews took longer when interviewers provided unsolicited clarification than when they only responded to explicit requests, $F(1,83) = 36.32$, $p < .001$. Interviews took reliably longer when interviewers used their own words to clarify the questions rather than reading the script, $F(1,83) = 17.41$, $p < .001$, but as the table shows, the real increase in duration was for the fully conversational case, interaction $F(1,83) = 10.09$, $p = .002$; this leads us to conclude that how interviewers provide clarification isn't the real determinant of interview length.

Across all five interviewing groups, then, we see an emerging pattern: response accuracy for complicated mappings increases with interview duration. The duration of the interviews conducted with intermediate levels of clarification was more than twice that of the strictly standardized interviews, and the duration of the conversational interviews was more than three times that of the strictly standardized interviews. In fact, there is a strong linear relationship ($r = .98$) between interview duration and response accuracy. For each additional minute that interviewers and respondents spent on clarification, there was a 7% gain in accuracy. This suggests that more clarification to respondents improves response accuracy more, but at a linear increase in interview duration, and thus in survey costs.

CONCLUSIONS

This study demonstrates that "intermediate" forms of conversational interviewing lead to intermediate levels of response accuracy. That is, response accuracy was better when interviewers provided clarification than when they didn't, but it was not as good as when interviewers clarified both (a) in their own words and also (b) whenever they deemed it necessary.

Across all interviewing types, response accuracy was better when interviewers provided unsolicited clarification than when they provided clarification only at respondents' request. But this improved accuracy came at a cost. We found intermediate interview duration for all three "intermediate" interview types; the three types did not take reliably different amounts of time. In general, our findings show that better response accuracy comes at the cost of increased interview duration.

These data are consistent with our findings on question clarification in a national telephone sample (Conrad & Schober, 2000) and in computer-administered self-interview questionnaires (Conrad & Schober, 1999): response accuracy can be improved through additional question clarification, but this clarification comes at a cost. The current study suggests that some benefits of fully conversational interviewing can be gained at lower cost, although the benefits won't be as great.

We believe this is particularly important to consider given how substantially current interviewing practices can vary across and within survey organizations that consider themselves to promote standardization. We propose that there is always a tradeoff between the need for accurate data and the costs of getting them (see

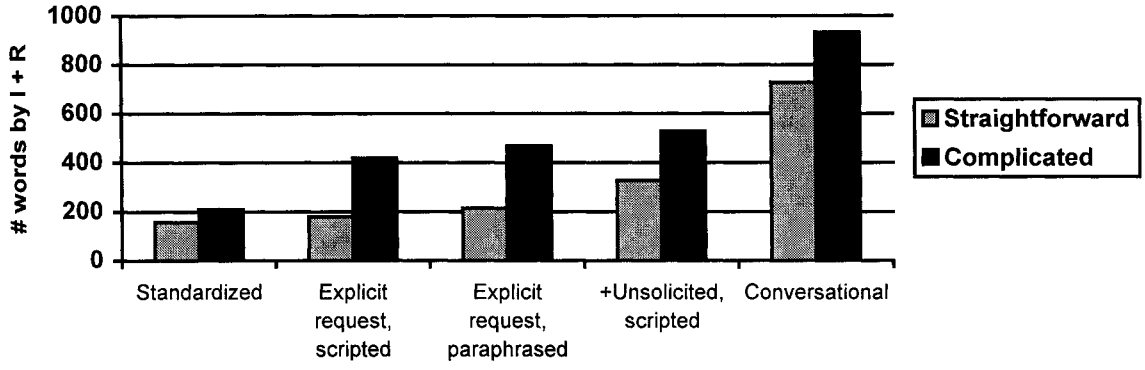

Figure 3. Interview duration (words per question)

Schober & Conrad, 2000); although current practice in some organizations may well reflect an optimum balance, we believe much remains unknown about the extent of variability in practice and just how this affects data quality.

The benefits of clarification, of course, depend on the frequency of complicated mappings. If complicated mappings are known to be rare, or if the need for precision is not pressing, then the extra costs of clarification may not be worth it. On the other hand, if the frequency of complicated mapping is unknown, or if they are known to be frequent, then encouraging interviewers to clarify may be a good idea. Just how much encouragement interviewers should have would then depend on how certain one needs to be that data are accurate.

Finally, our results show clearly that (1) relying on respondents to know when they need help may be insufficient. Respondents may not always ask for help when they give inaccurate responses (see also Conrad & Schober, 1999, 2000). Our results also show that (2) w*hen* interviewers offer definitions may matter more than how they word them.

REFERENCES

Beatty, P. (1995). Understanding the standardized/ non-standardized interviewing controversy." *Journal of Official Statistics*, 11,147-160.

Conrad, F.G., & Schober, M.F. (1999). A conversational approach to computer-administered questionnaires. *Proceedings of the American Statistical Association, Section on Survey Research Methods, 1998*, pp. 962-967. Alexandria, VA: American Statistical Association.

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

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

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. (1999). Response accuracy when interviewers stray from standardization. *Proceedings of the American Statistical Association, Section on Survey Research Methods, 1998*, pp. 940-945. Alexandria, VA: American Statistical Association.

Schober, M.F., & Conrad, F.G.. (2000). 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: Wiley, in press.

Suchman, L., & Jordan, B. (1990.) Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association* 85 (409):232-253.

Suchman, L., & Jordan, B. (1991). Validity and the collaborative construction of meaning in face-to-face surveys. In J.M. Tanur (ed.), *Questions about questions: Inquiries into the cognitive bases of surveys*, pp. 241-267. New York: Russell Sage Foundation.