DOES CONVERSATIONAL INTERVIEWING IMPROVE SURVEY DATA QUALITY BEYOND THE LABORATORY?[1]

Michael F. Schober, New School for Social Research
Frederick G. Conrad, 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: Standardized interviewing, conversational interviewing, data quality

INTRODUCTION

Imagine the following exchange between an interviewer (I) and a respondent (R) in a Current Population Survey (CPS) telephone interview:

I: Last week, did you have more than one job, including part-time, evening, or weekend work?

R: Well, it depends...I babysit for different people—is that one job or more than one?

What should the interviewer do now?

If the interviewer is administering a strictly standardized interview (see Fowler & Mangione, 1990), she will not answer the respondent's query. Instead, through the use of "neutral" probing techniques she will require the respondent to interpret the question by himself. The interviewer must maintain a neutral stance in order to minimize the possibility that she will influence or bias responses. If all interviewers follow this procedure, the variability in the data due to interviewers should be low. To proponents of standardization, low interviewer variance is necessary for high quality data.

Another possibility, advocated by Suchman and Jordan (1990, 1991) among others, is that the interviewer should answer the respondent's query, even if this means deviating from the pretested survey script and probing in ways that wouldn't be licensed in a strictly standardized interview. For example, our interviewer might explain that for this question the sponsoring organization counts babysitting for more than one employer as only one job (see US Department of Commerce, 1994, pp. B1-4 to B1-6).

Proponents of this alternative, which we will call *conversational interviewing*, believe that the interviewer and the respondent should work together to

assure that the respondent interprets the question as the survey designer intended. Unlike in standardized interviews, conversational interviewers may clarify the question as needed, and they may say whatever it takes to help the respondent interpret the question as intended. The rationale is that in ordinary communication, the desired understanding is guaranteed only through conversational collaboration (see, e.g., Clark & Brennan, 1991; Clark & Schober, 1991; Clark & Wilkes-Gibbs, 1986). On this view, survey interviews should be seen as specialized conversations, and participants should be able to collaborate to understand each other just as they do in ordinary conversation. As a result, what should be standardized is not *wording*, but rather *meaning* (Suchman & Jordan, 1991).

Both standardized and conversational interviewing techniques are designed to improve data quality, but they are based on different philosophies. Proponents of standardization attempt to reduce variance due to interviewers; proponents of conversational interviewing attempt to reduce variance due to respondents' misinterpretation of questions.

Note that in actual practice, interviewers sometimes use a combination of these techniques, for example, providing respondents with scripted help but only when respondents explicitly ask for it. Such mixed practices make it all the more urgent that we understand what kind of help is best (if any), especially given how complicated the official definitions for some survey concepts are. Take the concept of "more than one job" (US Department of Commerce, 1994, pp. B1-4 to B1-6):

A job exists when there is a definite arrangement for regular work every week, or every month, for pay or other compensation (e.g., profits, anticipated profits, or pay in kind, such as room and board). A formal, definite arrangement with one or more employers to work on a continuing basis for a specified number of hours per week

or days per month, but on an irregular schedule during the week or month, is also a job...It is possible for individuals to have more than one employer, but only one job. If an individual does the same type of work for more than one employer in an occupation where it is common to have more than one employer, do not consider the individual a multiple jobholder. Examples include private household or domestic workers including babysitters, chauffeurs, gardeners, handypersons, cooks, and maids.

This is long and complicated enough that it might be impractical for interviewers to read the scripted definition in full.

In a recent laboratory experiment (Schober & Conrad, 1997), we compared the accuracy of responses collected in standardized and conversational telephone interviews using questions from ongoing government surveys. Respondents answered from experimenter-designed fictional scenarios—floor plans, work descriptions, or purchase receipts—and so response accuracy could be measured directly as the fit between responses and official definitions like the one for "more than one job." The scenarios were available to the respondents both before and during the interviews, but they were never available to interviewers, so interviewers never knew the correct answer.

For half of the scenarios seen by any one respondent, the circumstances mapped onto the official definitions of the concepts in the survey questions in a straightforward way. For example, a respondent asked "Last week, did you have more than one job, including part-time, evening, or weekend work?" would see a work schedule with the name of the one family for whom this fictional person babysat. For the other half of the scenarios, the circumstances mapped onto the official definitions in a complicated way. For example, a different respondent asked the same question would see a work schedule with the names of several different families for whom the fictional person babysat last week.

The results showed that response accuracy was extremely high in both standardized and conversational interviews when the mappings between scenarios and definitions were straightforward (97% and 98%, respectively). But when the mappings were complicated, accuracy in standardized interviews was quite poor (28%), while accuracy in conversational interviews was quite high (87%).

These results suggest that conversational interviewing can substantially improve response accuracy when respondents are uncertain about how their own circumstances correspond with official

definitions of concepts in the survey question. But this accuracy can come at a substantial cost: Conversational interviews took more than three times as long as standardized interviews.

For these results to be useful for survey practitioners, evidence is needed that they extend beyond the laboratory. It is entirely possible that they do not. First, mappings between question concepts and actual respondents' circumstances may be straightforward most of the time, and so inaccurate responding due to misunderstanding questions may be relatively rare. If this is the case, then the benefits of standardization probably outweigh the costs; if not, conversational interviewing is worth exploring further. Second, respondents may be unwilling to engage in lengthier conversational interviews, even if these interviews helped them understand questions better.

EXPERIMENT

Measuring understanding in real survey settings is more difficult than it is in the laboratory. Without direct access to respondents' circumstances, we can't easily tell if responses match the official definitions. And directly measuring accuracy has its own problems: Record checks and diary checks are expensive, and the records may not themselves be accurate.

But we *can* determine if conversational interviewing *changes* a respondent's understanding of a question concept. The logic of our study is this: The effects of conversational interviewing can be seen by observing whether respondents in an initial standardized interview change their responses in a subsequent conversational interview more than they would in a subsequent standardized interview.

Another way to measure comprehension in a real survey setting is to ask respondents to explain their answers. For example, if respondents answer that they did indeed make household furniture purchases in the last year, they can be asked to list what those purchases were. Each item in their list can then be coded as "legal" or "illegal" according to official definitions.

We carried out an experiment to compare data quality for conversational and standardized interviewing techniques, using these two measure. The 227 respondents were interviewed at home by 20 experienced telephone interviewers calling from Westat, a social science research firm in Rockville, MD. Each respondent was asked the same survey questions twice, each time by a different interviewer, with about one week between interviews. The first interview was always standardized. For half the respondents (118), selected at random, the second

interview was also standardized. For the other half (109), the second interview was conversational.

If conversational interviewing improves comprehension, then responses should change more from the first (standardized) interview to the second if the second interview is conversational rather than standardized. And the percentage of "legal" explanations should be greater in the conversational interviews than in the standardized interviews.

Note that, unlike in the laboratory experiment (Schober & Conrad, 1997), we do not control the frequency of complicated mappings. This suggests that if conversational interviewing really does produce more response change and more "legal" explanations, then complicated mappings are frequent enough to worry about in actual survey settings.

Questions. Respondents were asked the same 10 questions in both the first and second interviews. Five were about housing, adapted from the Consumer Price Index Housing (CPI-Housing) survey; these required numerical responses. For example, one question was "How many bedrooms are there in your home?". The other five were about purchases, adapted from the Current Point of Purchase Survey (CPOPS); these required yes/no responses. For example, one question was "During the past year, that is since July of 1995, have you purchased or had expenses for household furniture?". Respondents who answered "yes" to any purchasing questions were asked to list the purchases.

Participants. Respondents were chosen from a nationally representative sample of residential households in the continental U.S. with telephones, generated through a random digit dialing process. The 118 respondents whose second interview was standardized were 44 men and 74 women; 88 were White, 17 Black, 1 Asian/Pacific Islander, 1 Native American, and 10 Other (1 refused to report); 9 had no high school diploma, 38 had completed high school only, 36 had some college but no degree, 19 had a college degree, and 15 had postgraduate degrees (1 other). The 109 respondents whose second interview was flexible were 36 men and 73 women; 88 were White, 16 Black, 1 Asian/Pacific Islander, 1 Native American, and 2 Other (1 refused to report); 8 had no high school diploma, 40 had completed high school only; 35 had some college but no degree; 14 had a college degree, and 11 had postgraduate degrees

The 20 interviewers were all professionals with no reliable differences in experience: 29.6 months (standardized interviewers) and 34.4 months (conversational interviewers) at Westat. 15 interviewers used standardized technique, 10 in the first interview

and 5 in the second interview. Of these 15, 13 were women and 2 were men; 5 were Black, 7 were White, and 3 were Hispanic; 3 had additional professional interviewing experience elsewhere. The 5 flexible interviewers were 4 women and 1 man; 2 were Black and 3 were White; 3 had additional professional interviewing experience elsewhere.

For the initial standardized interviews, 62.3% of those contacted agreed to participate. This rate would undoubtedly have been higher had we not asked respondents, at the outset, (a) to participate in both interviews and (b) for permission to audio-record both interviews. Response rates for the second interviews (proportion of those participating in the first interview) were quite high: 89.7% for standardized second interviews and 82.9% for conversational second interviews.

Interviewer training. All interviewers were trained on the key survey concepts for about one hour; this included a presentation about the concepts, a quiz and group discussion about the concepts.

The 15 standardized interviewers were trained for an additional hour to conduct a pure version of standardized interviewing where they read questions exactly as worded and provided only nondirective probes, but never provided definitions for the survey concepts. Probing techniques included rereading the question, providing the response alternatives, and otherwise probing neutrally.

The other 5 interviewers were trained for an hour to conduct conversational interviews. Like standardized interviewers, these interviewers were instructed to read the questions exactly as worded, but then they could say whatever they wanted to assure that the respondent had understood the question as the survey designer had intended. This included reading or paraphrasing all or part of a question, reading or paraphrasing all or part of a definition, and asking questions of the respondent to elicit information so that the interviewer and respondent could jointly reach a correct response. Interviewers could intervene at the respondent's request or voluntarily; that is, interviewers were licensed to intervene whenever they thought the respondent might have misunderstood the question.[2]

Conversational interviewers were also trained to explain to respondents that this interview would be different from the first one, and probably different from

---

[2] Note that this technique doesn't license interviewers to do anything they please or alter the research agenda for each respondent, along the lines of more radical proposals (e.g., Mishler, 1986).

most interviews in which they had previously participated. Interviewers were to explain to respondents that official definitions might differ from respondents' own ideas about what counted as a room or a furniture purchase. Interviewers were to encourage respondents to ask for help interpreting or answering questions whenever they had the slightest doubt about how to do this.

RESULTS

Implementation of interview techniques. In order to interpret our results, we need to be sure that interviewers correctly implemented both interviewing techniques. Because this involved detailed and labor-intensive inspection of interviewer-respondent interaction, we sampled 35 pairs of interviews from the full set of 227 pairs of interviews, and transcribed and coded the audiotaped interviews. The selected pairs were sampled at random, with the constraint that at least two interviews by each interviewer would be included. This ensured that about half the sampled pairs would have standardized second interviews, and about half conversational.

In this subsample, we counted all *directive interventions*—all those utterances by interviewers that would be "illegal" in pure standardized interviewing. These included providing definitions or clarifications of concepts in the survey questions (either verbatim from the official documents or improvised), requesting information from the respondent relevant to the definition, offering help, proposing potential purchases overtly telling respondents what the answer should be, and rewording the survey question (after reading it initially),. At least one directive intervention occurred in 86.7% of the questions in conversational interviews, but in only 5.7% of the interactions in standardized interviews. (The percentage was the same for second-week standardized interviews, as in all first-week standardized interviews). In fact, by more lenient criteria most of the "illegal" interventions in our standardized interviews could be counted as legal, since 87% of them, 26 of 30, were partial but verbatim repetitions of the survey questions. So clearly the two types of interviews were implemented in qualitatively different ways and much as we had intended.

Measure 1: Response change. In the full sample of 227 pairs of interviews, more responses changed when the second interview was conversational (21.8% change from first interview) than when it was standardized (11.0% change), $F1(1,225) = 45.61, p < .001$. This was true for all ten survey questions, $F2(1,18) = 4.60$, p < .05.[3]

The 11% rate of response change across the two standardized interviews is within the normal range for reinterviews in large government surveys like those sponsored by the Bureau of Labor Statistics, and probably reflects ordinary memory or reporting errors. The additional 11% change for conversational interviews must result from something beyond this.

The rate of response change in the subsample of 35 transcribed pairs of interviews is similar to the overall rate of response change in the full sample. Thus we can be confident that the subsample represents the full sample of 227. Respondents in conversational second interviews changed 35 of 165 answers from the standardized first interviews, or 21.2%; respondents in standardized second interviews changed 10 of 174 answers from the standardized first interviews, or 5.7%.

Measure 2: "Legal" reported purchases. Whenever respondents answered "yes" to a purchase question, they were asked to list the purchases on which they based their response. In the full sample of 227 pairs of interviews, 56.8% of respondents' purchases in the first interview were consistent with the official ("legal") definitions. However, the type of second interview had a substantial effect on the proportion of legal reported purchases: 56.9% of the purchases that respondents reported were legal when the second interview was standardized but 94.5% of their reported purchases were legal when the second interview was conversational, interaction $F(1,206) = 86.62, p < .0001$. This advantage for conversational interviewing was not because the respondents in the conversational interview happened to conceive of the purchases in the same way as the sponsoring organization: They were no more likely to have reported legal purchases in the first interview (57.6%) than were their standardized counterparts (56.0%).

Why did respondents change their reports of purchases? By making sure that respondents understood the official definitions, conversational interviewers helped respondents to classify their purchases more accurately. Respondents in conversational interviews included purchases they should have included the first time, and they omitted purchases they had mistakenly included the first time. That is, in those cases where respondents changed their answer to a purchase question in the second interview to "yes," the newly reported purchases were legal 90.1% of the time. In

---

[3] *F2* is for an analysis of variance with items (questions) as the random factor.

those cases where respondents changed their answer in the second interview to "no," the previously included (now omitted) purchases were rarely legal (20.7%).

In contrast, when respondents in standardized interviews changed their responses to a purchase question in the second interview, the second answer was no more likely to reflect a legal interpretation than an illegal one. When standardized respondents changed their answer to "yes," the newly reported purchases were legal 46.6% of the time. When respondents changed their answer in the second interview to "no," the previously included (now omitted) purchases were just as likely to be legal (52.1%).

Analysis of the subsample of transcribed interviews gives essentially the same picture, this time for all ten questions. Every time that respondents in conversational interviews provided different answers than in their first interviews (35 times), interviewers had provided at least one directive intervention. In contrast, when respondents in standardized interviews provided different answers than in their first interviews (10 times), in only one of those cases did the interviewer provide any directive intervention. (For more details, see Conrad & Schober, in preparation).

Accuracy of interviewers' interventions. One concern that proponents of standardized interviewing have raised about conversational interviewing is that interviewers can mislead respondents. That is, even if interviewers sometimes provide information which helps respondents to produce accurate answers, interviewers may just as often provide information which can lead respondents astray. This doesn't seem to have been the case here, as seen in the subsample of 35 interviews. Conversational interviewers provided accurate official definitions (verbatim or improvised) for 116 of the 121 questions where definitions were presented, a rate of 95.3% accuracy. (The one standardized second interviewer who provided a definition presented inaccurate information).

This suggests that, as in the Schober & Conrad (1997) laboratory study, conversational interviewers can present highly accurate information to respondents, contrary to Fowler and Mangione's (1990) concerns.

Interviewer variability. Another concern raised by proponents of standardized interviewing is that conversational interviewing will lead to undesirable levels of error due to interviewer variability. One way to test interviewer variability is to use *rho*-int, as described in Fowler (1991), which requires an experimental design in which each interviewer's respondents are representative of the larger sample. Unfortunately, this was not under our control in this experiment, and so we are unable to use *rho*-int. But we *can* calculate whether the variability in responses for different interviewers is greater in conversational than standardized interviewers.

Responses did not differ for different interviewers in standardized second-week interviews, $F(1,4) = 0.68$, *n.s.*, nor did they differ in conversational second-week interviews, $F(1,4) = 1.24$, *n.s.*[4] Levene's homogeneity of variance test suggested a marginal possibility of greater interviewer effects in conversational second interviews, $F(1,4) = 2.06$, $p < .10$, than in standardized second interviews, $F(1,4) = 0.28$, *n.s.* This would need to be replicated with a larger sample to be considered reliable.

But even if we did find greater interviewer effects in conversational interviews than in standardized interviews, what should we make of that? Interviewer effects are only undesirable if we assume that they necessarily lead to poorer data quality. We propose that it may be better for variability due to interviewers to increase because some flexible interviewers increase data validity more than others, rather than for interviewer variability to be low in standardized interviews because accuracy is uniformly low.

Interview duration. Conversational interviews took 80% longer than standardized interviews. The median time to complete all standardized interviews was 5 minutes (timed in minutes by interviewers themselves), compared to 9 minutes for conversational. interviews, interaction F(1,224) = 60.88, p < .0001.

Some of the increased duration for conversational interviews must be the result of time that interviewers spent explaining the conversational interviewing technique to respondents. Despite this, the 80% increase in duration in the current study is far less than the threefold increase in duration found in the Schober and Conrad (1997) laboratory experiment, where interviewers didn't spend any time explaining conversational interviewing techniques. Why might this be? Presumably this reflects a difference in the frequency of complicated mappings. In the Schober and Conrad (1997) experiment respondents had complicated mappings between the questions and the scenarios from which they were answering fully 50% of the time. In the current study, mappings were not controlled, and complicated mappings may well have been less frequent.

---

[4] *Fs* reported for analyses of variance with interviewer as the random factor, for the questions which have numerical answers (the housing questions).

914

SUMMARY AND CONCLUSIONS

Our results show that conversational interviewing can increase response accuracy beyond the laboratory, as indicated by indirect measures of response accuracy (response change and reported purchases). As in the Schober and Conrad (1997) laboratory study, conversational interviewing substantially improved respondents' comprehension of survey concepts, but it came at a cost: increased interview length.

Complicated mappings between respondents' circumstances and question concepts seem to be less frequent in the real world than they were in the Schober and Conrad laboratory study, where for experimental design purposes mappings were complicated 50% of the time. But complicated mappings *were* frequent enough to compromise comprehension, and thus response accuracy, at levels high enough to warrant concern. We have strong evidence for this in the fact that fewer than 60% of purchases reported by respondents in standardized interviews should actually have been reported, according to official definitions.

How should survey researchers deal with this source of error (complicated mappings)? Our results suggest that conversational interviewing is a promising approach, though it is largely untested. Such testing is particularly important because current practice often includes some features of conversational interviewing.

We propose that different interviewing techniques may be appropriate for different circumstances. For any particular survey, researchers should estimate the frequency of complicated mappings in their sample, and they should determine how certain they need to be (given cost constraints) about response accuracy. Our results suggest that standardized interviewing may be the right approach when complicated mappings are rare and certainty needs are low. In contrast, conversational interviewing may be the right approach when complicated mappings are frequent and certainty needs are high.

REFERENCES

Clark, Herbert H., and Susan E. Brennan. 1991. "Grounding in Communication." In *Perspectives on Socially Shared Cognition,* ed. Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, pp. 127-149. Washington, DC: American Psychological Association.

Clark, Herbert H., and Michael F. Schober. 1991. "Asking Questions and Influencing Answers." In *Questions About Questions: Inquiries into the Cognitive Bases of Surveys,* ed. Judith M. Tanur, pp. 15-48. New York: Russell Sage Foundation.

Clark, Herbert H., and Deanna Wilkes-Gibbs. 1986. "Referring as a Collaborative Process." *Cognition* 22: 1-39.

Conrad, Frederick G., and Michael F. Schober. Does conversational interviewing improve survey data quality beyond the laboratory? Manuscript in preparation.

Fowler, Floyd J. 1991. "Reducing Interviewer-Related Error Through Interviewer Training, Supervision, and Other Means." In *Measurement Error in Surveys,* ed. Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman, pp. 259-278. New York: John Wiley.

Fowler, Floyd J., and Thomas W. Mangione. 1990. *Standardized Survey Interviewing: Minimizing Interviewer-Related Error.* Newbury Park, CA: SAGE Publications, Inc.

Mishler, Elliot G. 1986. *Research Interviewing.* Cambridge, MA: Harvard University Press.

Schober, Michael F., and Herbert H. Clark. 1989. "Understanding by Addressees and Overhearers." *Cognitive Psychology* 21:211-232.

Schober, Michael F., and Frederick G. Conrad. 1997 "Does Conversational Interviewing Reduce Survey Measurement Error?" *Public Opinion Quarterly,* in press.

Suchman, Lucy, and Brigitte Jordan. 1990. "Interactional Troubles in Face-to-Face Survey Interviews." *Journal of the American Statistical Association* 85 (409):232-253.

Suchman, Lucy, and Brigitte Jordan. 1991. "Validity and the Collaborative Construction of Meaning in Face-to-Face Surveys." In *Questions About Questions: Inquiries into the Cognitive Bases of Surveys,* ed. Judith M. Tanur, pp. 241-267. New York: Russell Sage Foundation.

U.S. Department of Commerce. 1994. *Current Population Survey Interviewing Manual* (CPS-250). Washingto n, DC: Bureau of the Census.