Survey Research Methods (2020)
Vol. 14, No. 5, pp. 515-531
doi:10.18148/srm/2020.v14i5.7617

© *European Survey Research Association*
ISSN 1864-3361
http://www.surveymethods.org

The Benefits of Conversational Interviewing Are Independent of Who Asks the Questions or the Types of Questions They Ask

Frost A. Hubbard
Westat
Rockville, MD, USA

Frederick G. Conrad
Survey Research Center
University of Michigan, USA

Christopher Antoun
Joint Program in Survey Methodology (JPSM)
University of Maryland, USA

By clarifying the meaning of survey questions, interviewers help assure that respondents and researchers interpret questions the same way. This practice is at the heart of conversational interviewing and has been shown to improve response accuracy relative to standardized interviewing. This research investigates two issues: (1) Does conversational interviewing lead to improved response quality for opinion questions as it does for factual questions? and (2) Are some interviewers better suited to conduct conversational interviews than others? 490 respondents in the University of Michigan Surveys of Consumers participated in standardized telephone interviews after which they were re-asked five factual and five opinion questions. These questions were re-administered in conversational interviews for half the respondents; for the remaining half they were re-administered in standardized interviews. Interviewers also completed a nonverbal sensitivity questionnaire. Using response change between the two administrations of each question to measure response quality, the conversational technique improved quality, while increasing interview duration. The comprehension benefits of conversational interviewing were no greater for opinion than factual questions. Moreover, interviewers low in nonverbal sensitivity more often gave definitions before respondents were able to speak, but this did not affect data quality (response change). Taken together these results suggest that conversational interviewing can be effectively administered by a range of professional interviewers, although those who are more attuned to respondents' comprehension will be more efficient, and the technique will equally benefit the quality of responses to questions about objective and subjective phenomena.

Keywords: conversational interviewing; nonverbal sensitivity; measurement error

1 Introduction

For decades, survey researchers have administered "standardized interviews" in order to reduce interviewer-related measurement error and increase the comparability of responses. The crux of the method is to hold constant the words interviewers use when presenting questions to respondents. When the interaction requires interviewers to depart from the question script, the standardized interviewing method instructs interviewers to provide relatively content-free, "neutral", or "non-directive" probes (e.g., Fowler & Mangione, 1990). The thinking behind the approach is that if all interviewers in a survey provide the same information to respondents and behave in essentially the same way, there is little opportunity for them to influence responses, or at least to do so differently from one another.[1] However, by strictly adhering to standardized wording, the standardized interviewing protocol may increase misunderstanding by respondents and reduce the accuracy of their answers. People can interpret a word—even an ordinary word—quite differently than intended by the author of a question (e.g., Schober, Suessbrick, & Conrad, 2018). In addition, they can interpret the words as intended in a general sense but be uncertain what to include and what to exclude, e.g., if the question asks about the pur-

Contact information: Frost A. Hubbard, Westat, 1650 Research Boulevard, Rockville, MD 20850, USA (Email: FrostHubbard@Westat.com)

[1] The implication is that standardized interviewing should minimize interviewer variance and thus it should be lower than in conversational interviewing. However, West, Conrad, Kreuter, and Mittereder (2018) demonstrate that interviewer variance is not generally greater in conversational than standardized interviews and when it is, the reduced bias (increased accuracy) more than outweighs the increased variance.

chase of liquor, should they include fortified wines such as sherry (see Conrad & Schober, 2000)? If respondents recognize their possible confusion about the question's intended meaning and ask for clarification, their efforts will likely be thwarted as the logic of standardized interviewing prohibits interviewers from providing information about the question to some respondents if they do not give it to all respondents. As a result, respondents' misconceptions may be undetected and subsequently left uncorrected by the interviewer. Suchman and Jordan (1990, 1991) critiqued standardized interviewing, suggesting that the approach might promote reliable data that is not necessarily valid. As a result, they argued that interviewing should be built around every day conversational practices used to assure that both parties understand each other. One implementation of this idea, tested for factual survey questions, has been called "conversational interviewing" (e.g., Conrad & Schober, 2000; Schober & Conrad, 1997), the name reflecting the central role of conversational grounding (e.g., Clark, 1996), i.e., the back-and-forth used in most conversations to assure that participants understand each other well enough to carry out their communicative task successfully. In all of their studies comparing conversational to standardized interviewing, (e.g., Conrad & Schober, 2000; Schober & Conrad, 1997; Schober, Conrad, Dijkstra, & Ongena, 2012; Schober, Conrad, & Fricker, 2004; West et al., 2018), Schober and Conrad have instructed interviewers to read the questions exactly as worded and to then ground the concepts, that is to explain what the question is intended to mean based on official definitions, using whatever words they judge necessary with a particular respondent on a particular question. Interviewers are instructed to do this when their intuitions indicate that respondents are confused or have misunderstood what they were being asked.

As most of these studies involve telephone interviews, the evidence that respondents might benefit from clarification is spoken (as opposed to visual). This might include an explicit request for clarification (e.g., "What do you mean by 'usually'?") or a less direct request. For example, a respondent may describe one's circumstances rather than providing a response option, (e.g., when asked if she works for pay, a respondent might answer "I help on our family's farm.") rather than "yes" or "no" (see Schaeffer & Maynard, 2008, for a discussion of such "reports"). Schober and Bloom (2004) determined that a number of paralinguistic phenomena, in addition to reports, were related to the need for clarification, with pauses and fillers ('um's and 'uh's) being the most predictive. However, Schober and Conrad did not instruct conversational interviewers to specifically attend to or respond to these particular cues of communication difficulty.[2] Their thinking was that instructing interviewers to monitor for specific utterances requires considerable attention that may compromise how well interviewers perform their many other tasks and may inhibit their detection of other cues of comprehension difficulty. Moreover, people can recognize at least to some degree that a partner in everyday conversation does not understand or misunderstands what has been said and so interviewers should be able to use those skills in conversational interviews. How much variation exists is in these skills is part of what the current study investigates. Respondents' understanding of the survey questions have been assessed in different ways in different studies. In Schober and Conrad (1997) respondents answered based on fictional scenarios—not their own lives—for which there were clear correct and incorrect answers. For example, respondents were asked, "Has Dana purchased or had expenses for household furniture?" and asked to answer based on a scenario consisting of a receipt for the purchase of a floor lamp. The correct answer was "no" given the US Bureau of Labor Statistics' definition (used by Schober & Conrad, 1997): floor lamps are considered "lamps and lighting fixtures" not "household furniture." Schober and Conrad called this kind of ambiguity a "complicated mapping."[3] In contrast to complicated mappings, Schober and Conrad identified straightforward mappings in which the correspondence of the survey concept to the respondent's situation is not ambiguous. For example, based on a receipt for an end table the correct answer is clearly "yes" as an end table is a prototypical piece of household furniture. They reported nearly perfect accuracy by this measure for both strictly standardized and conversational interviews when scenarios were straightforward. However, when the scenarios were complicated the authors reported substantially greater response accuracy for conversational interviews, in which interviewers could help resolve the ambiguity, than for standardized interviews, which afford interviewers no tools for clarifying question meaning. A less direct—but presumably more realistic—method of assessing whether conversational interviewing improves data quality involves administering the same questionnaire twice, first in a standardized interview and second in either a standardized or a conversational interview. The logic of this approach, developed by Conrad and Schober (2000), relies on the empirical findings that, when true values are known (as in Schober & Conrad, 1997), conversational interviewing helps respondents interpret the question consistently with researchers' intentions, resulting

[2] Schober et al. (2012) identified visual cues of comprehension difficulty exhibited by respondents in face-to-face conversational interviews, in particular averting the interviewers' gaze while answering. As in the studies of telephone interviews, the authors instructed interviewers to use their judgment about the respondents' need for clarification rather than attending to this specific behavior.

[3] Note that complicated mappings do not indicate the question is badly written but that the way its words correspond to the respondent's circumstances is ambiguous—is a floor lamp furniture? Across the Conrad and Schober studies the survey questions had been previously pretested for production data collection so, presumably, problems related to question wording had been largely resolved.

in more accurate answers. Standardized interviews do not authorize interviewers to address conceptual misalignment directly in this way. Thus, in a re-interview design that includes an initial standardized interview, Conrad and Schober predicted more response change when the second interview was conversational than when it was standardized because the respondents' interpretations were brought into alignment with those of the researchers' by conversational interviewers but standardized interviewers were not able to effectively address respondents' misunderstandings. The authors did in fact observe significantly more response change between the initial standardized and a subsequent conversational interview (22%) than between two standardized interviews (11%).

It is possible that the greater response change when the second interview was conversational was unrelated to improved understanding of the critical terms in the questions, but Conrad and Schober's rationale was supported by a follow-up analysis. When respondents answered "yes" to questions asking whether they had made a particular purchase, the interviewers probed for more detail about the purchases on which respondents based their affirmative answers. The analysis showed that "yes" responses were based on legitimate inclusions, e.g., the category of "telephone purchases" legitimately includes buying cell phones but not paying for telephone service, substantially more often when the second interview was conversational (95%) than standardized (57%). The follow-up analysis strongly suggests that the increased response change in conversational than standardized re-interviews was due to interpretations that were more aligned with the researchers' intended meaning. A similar advantage for conversational interviewing was observed in other studies in which the authors directly measured response accuracy using scenarios (Schober et al., 2004) and indirectly using response change (Schober et al., 2012).[4] No matter how response quality was assessed, the benefits were not without cost: providing clarification took time, leading to increased duration for conversational interviews compared to standardized interviews. The tradeoff between increased interview duration and response accuracy for complicated situations is important for practitioners to weigh when determining whether to administer conversational interviews in a particular study. Budget, likelihood of complicated situations for the particular questions asked, and which interviewers might be available for the study, are among the relevant considerations.

There are still unanswered questions about the conversational interviewing technique. The two that we address here are (1) whether conversational interviewing improves respondents' understanding of opinion questions as it does for factual (behavioral) questions, and (2) whether all interviewers are equally effective in using conversational interviewing.

1.1 Effectiveness of conversational interviewing for opinion questions

Because the previous studies of conversational interviewing focused on factual questions, the notion of response accuracy was more straightforward than it is for opinion questions. Much has been written and debated about the nature of attitudes, for example, whether they are stable trait-like dispositions or are constructed in the moment and in a particular context (e.g., Schwarz, 2007). For current purposes, the exact definition of an attitude is not critical. Our focus is on how well a researcher is able to communicate to a respondent what he or she means when referring to the object of the attitude. For example, when a respondent is asked if "during the next 12 months we'll have good times financially, or bad times, or what?" what exactly is intended by terms like "good times" and "bad times?" By allowing interviewers to communicate the researchers' intended meaning to respondents who request it or seem to need it, conversational interviewing should help interviewers to resolve misconceptions in much the same way for attitude questions that it does for behavioral questions.

It is common to measure opinions on response scales such as favor-oppose, agree-disagree, satisfied-not satisfied, etc. While questions about behaviors and facts can also make use of response scales for dimensions such as frequency, likelihood, and educational attainment, it is probably more common to collect a number from respondents or the category into which they place their behavior (e.g., "employed"). This raises the question of whether providing definitions for values in a response scale is in any way different from defining terms in the question stem, e.g., the word "vote" in "How likely are you to vote?" We explore this distinction in the current study.

1.2 Interviewer nonverbal aptitude

In previous comparisons of standardized and conversational interviewing, trained, professional interviewers from the public, private, and academic sectors administered both techniques. Thus, conducting conversational interviews seemed not to require special skills beyond being able to

[4] In Schober et al. (2012), the second administration of the questions was self-administered on paper and included definitions of the key concepts, leading to a slightly different logic and pattern of results than in Conrad and Schober (2000), which also measured accuracy with response change. The authors expected more response change in the standardized than conversational interview condition because misconceptions that were not corrected in the standardized interview could be corrected through exposure to the definitions in the subsequent questionnaire leading respondents to change their earlier answers; because conversational interviewers could detect and correct those misconceptions during the interview, respondents would be less likely to change their answers when exposed to definitions in the later questionnaire. This is exactly what was observed.

judge whether respondents understand the question (Conrad & Schober, 2000). However, while most speakers of a language are able to engage in conversation, some may be better attuned to how well their conversational partners understand what they are saying and how well they are following the discourse. The ability to judge the listener's grasp of what one has just said falls into the broad category of nonverbal sensitivity, which Carney and Harrigan (2003) define as "the ability to accurately assess others' abilities, states, and traits from nonverbal cues." Some of the earliest social psychology experiments on nonverbal sensitivity investigated how accurately people "decoded" others' expressions and judged others' emotions (Feleky, 1914). Since the 1930s, social psychologists have considered nonverbal sensitivity to be an important skill in everyday functioning (Kanner, 1931; Vernon, 1933), and the topic continues to be widely studied (e.g., Hall, Andrzejewski, & Yopchick, 2009; Hall, Bernieri, & Carney, 2005; Pickett, Gardner, & Knowles, 2004). If some interviewers are higher in nonverbal sensitivity than others, they may more effectively conduct conversational interviews than others because they will better judge how well respondents understand questions and be able to provide clarification when it is needed and not when it will be superfluous. Survey respondents, like most listeners, sometimes signal their understanding through their nonverbal behaviors (e.g., Moore & Maynard, 2002; Schaeffer & Maynard, 2002, 2008; Schober & Bloom, 2004; Schober et al., 2012). Thus, the current study investigates whether interviewers who are more sensitive to the nonverbal behavior of others are more adept at recognizing when respondents are confused or have misinterpreted a word or phrase in a question and so are more likely to provide help at the right times.

Nonverbal sensitivity may be particularly important because respondents rarely ask for clarification even when they are explicitly encouraged to do so. Hence, instead of simply relying on respondents to state explicitly that they are confused, successful conversational interviewers presumably detect respondents' implicit indications of confusion or misunderstanding and provide clarification accordingly. The spoken nonverbal evidence that respondents might benefit from clarification include pauses of longer than one second as well as fillers like "um" and "uh". Interviewers who are relatively skilled in the detection of these cues may be better at providing clarification when it is needed and foregoing it when it is not needed. By contrast, interviewers who are lower in nonverbal sensitivity may fail to act on evidence of comprehension difficulty or may compensate for being relatively weak at assessing respondents' comprehension by providing frequent but indiscriminate clarification in conversational interviews. This could lead to reduced response accuracy but is likely, at least, to inflate completion times.

We tested the following hypotheses in our study:

H1a Comprehension of factual questions will be more accurate with conversational than standardized interviewing.

H1b Comprehension of opinion questions will be more accurate with conversational than standardized interviewing.

H2a Interviewers who are more sensitive to respondents' nonverbal behavior will administer conversational interviews more effectively than those who are less sensitive, producing more improvement in comprehension.

H2b Interviewers who are more sensitive to respondents' nonverbal behavior will administer conversational interviews more efficiently than those who are less sensitive, producing improved comprehension in less time.

By testing H1a, we will see if the effects reported in the conversational interviewing literature, which were produced with factual questions, will replicate in a nationally representative sample on the telephone for these particular factual items. H1b tests whether comprehension is similarly improved by conversational interviewing for opinion questions. For H2a, our question is whether interviewers higher in nonverbal sensitivity are able to more effectively administer definitions (i.e., communicate the intended question meaning to respondents) better than interviewers lower in nonverbal sensitivity, leading to more accurate responses. H2b asserts that interviewers higher in nonverbal sensitivity are able to deliver clarification more strategically than interviewers lower in nonverbal sensitivity leading to faster, though not necessarily more accurate, conversational interviews.

2 Methods

The experiment was implemented as a rider in the June 2011 administration of the Surveys of Consumers (SCA), a monthly, centralized telephone survey about the U.S. economy conducted by the University of Michigan Survey Research Center (SRC). Each month, the SCA selects a nationally representative, random digit dial (RDD) sample of landline and cell phone numbers and then completes approximately 500 interviews, 60% with newly selected telephone households and 40% with respondents first interviewed six months prior. The June 2011 SCA completed 506 interviews, resulting in an AAPOR RR2 of 42.4%.

In the June 2011 survey, respondents were asked to answer questions about the economy (i.e., core SCA questions) and a set of supplementary questions about drowsy driving (i.e., driving a vehicle while tired). They were then re-asked ten questions from the first section in the same order the questions were asked in the first section. These questions were re-administered using either standardized interviewing, as in the original administration, or conversational interviewing. The

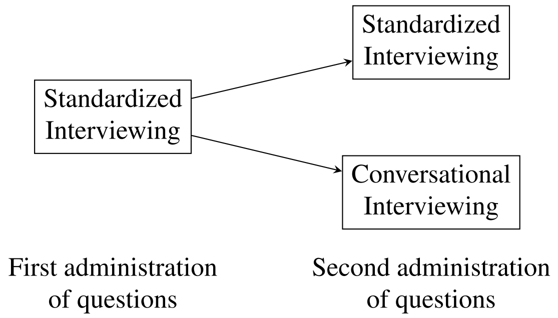
Standardized Interviewing
Standardized Interviewing
Conversational Interviewing
First administration of questions
Second administration of questions

Figure 1. Study design

critical measure was response change between the two administrations (see below). Note that in the re-administration section interviewers administered exactly the same questionnaire irrespective of the interviewing technique with the exception of definitions being available to the conversational interviewers. This led to the same question context in all interviews, and rules out question order (e.g., Sudman, Bradburn, & Schwarz, 1996, chapters 4 and 5; Schwarz, Strack, & Mai, 1991; Strack, Schwarz, & Wänke, 1991; Tourangeau, Rips, & Rasinski, 2000, chapter 7) as an explanation for any effects of interviewing technique.

Prior to data collection, all selected cases were randomly assigned to one of the two techniques. All interviewers were blind to the experimental assignment of each case until the start of the second section.[5] The interviewer did not know which interviewing technique would be used to re-administer the ten items until the sample member agreed to participate; at this point the assignment was communicated to the interviewer by the instructions (to be read aloud) that appeared on the interviewer's screen at the start of the second section[6] (see Figure 1).

Data quality—in particular comprehension accuracy—for the ten re-administered questions was operationalized as the amount of response change between the first and second administration of the questions based on the following logic. By re-administering the ten questions with standardized interviewing techniques, we established a baseline level of response change, which is akin to random response variance. Response change in the conversational re-administration over and above the levels observed in standardized re-administration of the questions can be directly attributed to the conversational interviewing technique. The assumption is that higher levels of change reflect a revised understanding of the questions' meanings. Following the logic of Conrad and Schober (2000), if there is more response change when the second interview is conducted with a technique that has been shown to improve the understanding of key concepts in survey questions (conversational interviewing), it is reasonable to attribute the increased change in answers to increased changes in interpretation. Moreover, these changed interpretations are likely to be more aligned with the intentions of the researchers who designed the questions. As shown by Conrad and Schober 2000, when respondents were asked to explain their "yes" answers to questions about whether they had made certain purchases, their explanations were substantially more likely to include legitimate purchases, i.e., those consistent with official definitions, in conversational than standardized interviews, strongly supporting the link between increased response change and improved comprehension. Despite improving comprehension, clarifying questions requires additional time leading conversational interviews to be longer than standardized interviews (e.g., Conrad & Schober, 2000; Schober & Conrad, 1997; Schober et al., 2004). Thus, we also examine the duration[7] of the re-administered questions by interviewing technique.

2.1 Questions and definitions

The ten re-administered questions concerned several topics (see Appendix A for a list of the questions and definitions). Three opinion questions about the U.S. economy were selected because in previous SCA interviews respondents indicated higher levels of confusion when answering these items compared to others. Thus, they seemed like good candidates for improved understanding when interviewers are authorized to clarify the underlying concepts.[8] Three fac-

[5] We saw no evidence that the interviewers had trouble with the transition between techniques, either within or between interviews, as has been suggested might be the case (Schober et al., 2004). A review of the audio-recorded interviews revealed that interviewers successfully used the interviewing technique which they were randomly asked to use in all but four of the 466 interviews: these four cases were conducted as conversational interviews although they were assigned to the standardized interviewing condition and were excluded from our analyses.

[6] We did not to use conversational interviewing during the first section of the survey for two reasons. First, if respondents had received clarification (e.g., definitions) during the first administration, this knowledge about the intended meanings of the question would have likely carried over into the second administration of the questions, even if additional clarification was not provided in the second administration. In addition, using conversational rather than standardized techniques in the first administration, i.e., in the production interview used to collect data for the Index of Consumer Sentiment, could have affected the data contributing to the economic indices.

[7] The median duration of the initial interview for the 490 English language interviews that completed at least one question of the re-administration section was 34.3 minutes. Given this, some respondents may have been fatigued by the start of the re-administration section. While it is possible that the results may have looked different had respondents been less fatigued, all respondents irrespective of the experimental condition would have been equally fatigued on average, having just completed the same (standardized) SCA interview when beginning the re-administration section.

[8] This conclusion was based on a 2010 unpublished study of approximately 100 SCA interviews. We transcribed and coded the in-

tual questions were selected about the respondent's household. These questions concerned everyday concepts such as cell phones that were nevertheless complicated and potentially ambiguous. Finally, four questions (two factual, two opinion) were selected from the supplement about drowsy driving; our intuition was that these items contained concepts likely to be ambiguous for some respondents and which were therefore at risk of being misunderstood if interviewers were not authorized to clarify the intended meaning (as was the case in standardized interviews). We classified as "factual" the five questions that asked about information that could, in theory, be verified from an external source, e.g., "How many working cell phones do you (and your family living there) have in your household?" We classified as "opinion" the five questions that could not be externally verified (i.e., that measured a latent characteristic) and were fundamentally subjective, e.g., "Now turning to business conditions in the country as a whole--do you think that during the next 12 months we'll have good times financially, or bad times, or what?"

In conjunction with the SCA and Drowsy Driving investigators, we developed definitions for all of the questions. For the attitude questions some definitions concerned the attitude object, e.g., "better off financially" and "worse off financially", and others concerned the response scale, e.g., "extremely risky" and "not at all risky."

All interviewers completed a two-hour training session prior to data collection consisting of instruction about both the survey concepts, i.e., the definitions and scope of the concepts for all ten questions including details about the response categories, and the interviewing techniques. After they were trained in the concepts, the interviewers completed, and passed, a written test on the concepts. In the techniques training session, they were taught how to conduct both conversational and standardized interviews. Interviewers engaged in "paired-practice," taking turns as interviewer and respondent; they received feedback from supervisors and project staff and were able to ask any questions.

2.2 Measuring Nonverbal Sensitivity

The interviewers, drawn from the pool of regular SCA interviewers, completed a test of nonverbal sensitivity, allowing us to partition them into "high" and "low" sensitivity group. We used the Profile of Nonverbal Sensitivity (PONS), (Rosenthal, Hall, Di Matteo, Rogers, & Archer, 1979), because it is widely used and has strong predictive validity across a variety of research domains. The PONS tests both visual and auditory cues of nonverbal sensitivity, whereas many other measures only test visual cues. In fact, we used a version that tests sensitivity to only auditory cues[9] on the assumption this would best identify interviewers who were more and less able to detect relevant respondent cues over the telephone.[10] As the PONS is typically administered, test-takers listen to a recording, then are immediately presented with two situations (e.g., "helping a customer" or "talking to a lost child"; "asking forgiveness" or "expressing jealous anger"), and must choose which one is a better characterization of the situation. One is "correct" and the other is not. Identifying the correct scenario requires a relatively astute social judgment. To score highly on the PONS test-takers need to be relatively perceptive, interpersonally. Sixteen out of 24 SCA interviewers completed the PONS at a workstation in the centralized telephone facility while wearing a headset. The PONS testing took place after all interviews had been conducted.

2.3 Interviewers

Twenty-four interviewers completed at least one interview during the June 2011 SCA data collection. Seven interviewers who scored above the median audio PONS score of 30 were classified as "high" in nonverbal sensitivity group, and nine interviewers who scored 30 or below on the audio PONS were classified into the "low" nonverbal sensitivity group;[11] eight interviewers did not take the audio PONS test. Interviewers in the low and high sensitivity groups did not differ reliably in education, gender, race, or interview tenure.[12] Low sensitivity interviewers, however, tended to be older (40.7 years on average) than high sensitivity interviewers (26.4 years on average),[13] so we conducted two versions of analyses pertaining to nonverbal sensitivity, one version controlling for interviewer age and the other version not controlling for age. We focus on analyses that do not control for age, but at the end of the Results section, we report analyses that do control for age.

terviewers' and respondents' behavior in these interviews and then examined the prevalence of respondent behaviors that may indicate higher levels of confusion.

[9] The audio PONS can be found online: http://hdl.handle.net/2047/D20194665.

[10] The creators of the PONS captured 40 audio recordings, each of no more than five seconds, in which a female speaker of American English acts out interpersonal situations. The creators removed a listener's ability to understand any of the words in the recordings by either cutting the recordings into very short segments and then rearranging the order, or filtering the highest bands of frequencies of the recording so that the voice is muffled and the words cannot be recognized.

[11] A perfect score was 40. The audio PONS asks 40 questions and we assigned one point per correct answer. The interviewers' scores ranged from 23 to 33.

[12] The average interviewer tenure was 3.4 years (median: 2.8 years). There was no correlation between PONS score and interviewer tenure ($r = -0.13$, $p = 0.62$). This suggests that nonverbal sensitivity is not a matter of training but is more of an enduring characteristic.

[13] The average age of interviewers who did not complete the PONS was 42.2 years.

2.4 Analyses

The first set of results reported here is based on the responses of the 490 completed English language interviews (232 standardized and 258 conversational). The results concerning nonverbal sensitivity are based on the 367 interviews (195 conversational, 172 standardized) conducted by the 16 interviewers who took the PONS test. In addition to t-tests and chi-square tests, we fit two logistic regression models at the question level with response change as the outcome variable. In the first model, the predictor variables were interviewing technique (conversational, standardized), question type (factual, opinion) and the interaction of interviewing technique and question type. For the second model, the predictor variables were interviewing technique, interviewer nonverbal sensitivity group (low, high), and the interaction of interviewing technique and interviewer nonverbal sensitivity group. For the second model, in which nonverbal sensitivity group was a predictor, we would ideally have been able to treat interviewers as a random effect in order to generalize beyond SCA interviewers to the larger population of interviewers. Unexpectedly, treating interviewers as a random effect created a negative covariance structure within the nonverbal sensitivity group, making the models unstable. Thus, this model does not include interviewer as a random effect.

Finally, the results concerning the details of the interaction between respondents and interviewers are based on the 466 interviews (243 conversational, 223 standardized) that were audio-recorded; some interviews were not recorded because of technical errors and others because the respondent declined to be audio-recorded. For this final analysis, in order to quantify the interactions between interviewers and respondents in the recordings, we developed a coding scheme consisting of 43 "move" codes. The codes classified basic functional units of speech for respondents and interviewers (e.g., "respondent asks for clarification," "interviewer provides clarification") along with codes for fluency of speech, such as pauses and fillers.[14] When analyzing the relationship between interviewing technique and interviewer nonverbal sensitivity for both response change and re-interview duration (our second dependent variable), we created three groups of interviews:

1. conversational—high nonverbal sensitivity,
2. conversational—low nonverbal sensitivity interviewer, and
3. standardized.

We did not divide the standardized interviews into those completed by high nonverbal sensitivity or low nonverbal sensitivity interviewers as we had no reason to expect nonverbal sensitivity to affect when and how often interviewers provided clarification in standardized interviews. Standardized interviewing affords the interviewer few methods for clarifying the meaning of survey questions. The only tool at their disposal is neutral probes, which are not designed to clarify concepts. In these interviews, the PONS scores should not affect how respondents understand the questions.[15] Indeed, in the standardized interviews, section timings and response change did not reliably differ by interviewers' level of nonverbal sensitivity.

3 Results

We first examine overall response change. After controlling for the type of question (factual vs. opinion), significantly more responses changed when the re-administered questions were asked with conversational (18.3% of 2,508 question administrations, $n = 458$) than standardized (13.1% of 2,240 question administrations, $n = 294$) methods (OR = 1.505, $p < 0.0001$; see effect of Interviewing Technique in Table 1). [16] This pattern is similar to that found in Conrad and Schober (2000), and thus, as in that study, suggests that conversational interviewing helped respondents interpret questions as intended. More evidence that conversational interviewing performed as it has in almost all studies comparing the two techniques comes from the overall duration of the re-administered questions. The re-administration section took longer to complete when it was conversational (3.42 minutes) than when it was standardized (2.41 minutes), (Wilcoxon-Mann-Whitney test: z = -11.8, p < 0.001).[17] As in earlier studies, the longer duration of conversational interviews is due primarily to the time required to provide definitions. This helps confirm that the interviewing techniques were implemented as we intended: because providing definitions-as-needed leads to different wording for different respondents, interviewers were trained not to provide definitions when following standardized practice; because conversational interviewing is designed to promote conversa-

[14] Three people transcribed the recorded interviews and one person coded the interviews in Sequence Viewer 5.1 (http://www.sequenceviewer.nl). To assess the quality of the codes, we enlisted a second person to code 25 randomly selected conversational and 25 randomly selected standardized interviews. The Cohen's kappa was 0.85, indicating "Almost Perfect" agreement (Everitt & Haye, 1992).

[15] Interviewers classified into the "high" nonverbal sensitivity group might more accurately recognize that respondents are confused or have misinterpreted the question, but when they conduct standardized interviews there is little they can do to improve the respondent's understanding.

[16] The pattern of results for interviewing technique remained the same after also controlling for gender, age and education.

[17] In the re-administration section duration analysis, 41 of the 490 completed interviews were excluded because either the respondent indicated earlier in the initial interview that he/she did not drive and thus did not receive four of the ten re-administration questions ($n = 39$) or the respondent did not complete more than two of the re-administration questions (n = 2).

tional grounding, interviewers were trained to provide clarification when respondents requested it or the interviewers thought it was needed. If clarification improved comprehension, responses will change more when these questions are re-administered with conversational techniques.

Comprehension of both factual and opinion questions was indeed improved by conversational interviewing. Supporting Hypothesis 1a (and replicating previous studies), there was significantly more response change in conversational than standardized re-administrations for factual questions (9.0% of 1,272 conversational factual questions vs. 6.8% of 1,140 standardized factual questions, $\chi^2(1) = 4.29$, $p = 0.038$). In addition, there was significantly more response change in conversational than standardized re-administrations for opinion questions (27.8% of 1,236 conversational opinion questions vs. 19.7% of 1,110 standardized opinion questions, $\chi^2(1) = 20.56$, $p < 0.001$), supporting Hypothesis 1b. This suggests that when one clarifies the intended meaning of opinion questions (as is possible in conversational but not standardized interviewing), comprehension benefits much as it does for factual questions. After controlling for the interviewing technique, respondents changed their answers to opinion questions more than to factual questions (24.0% of 2,336 opinion questions vs. 8.0% of 2,412 factual questions, OR = 3.667, $p < 0.001$; see effect of Question Type in Table 1).[18] There was no interaction between interviewing technique and question type, indicating that conversational interviewing had the same effect on opinion questions and factual questions. Therefore, only the results of a main effects model, with no interaction term included, are presented in Table 1.[19] Although conversational interviewing increased response change for both opinion and factual questions compared to standardized interviewing, there was substantially more response change in opinion questions in general. People are likely to know with near certainty how many landline phones they have or their highest level of education, but are probably less certain about whether they think the economy is going to improve, stay the same or get worse. Factual questions by their nature concern more concrete concepts than opinion questions and this may increase respondents' confidence in their original answer. In contrast, the imprecision of opinion questions—if their object is not well defined—likely reduces respondents' confidence in their interpretation and leaves them more susceptible to change when provided a definition, even a few minutes later.

We created two types of definitions for the five opinion questions in this experiment: ones that defined the attitude object (e.g., "drowsy driving") mentioned in the question stem and others that concerned the response scale labels (e.g., "extremely risky" and "not at all risky"). There was some reason, a priori, to believe that clarifying respondents' interpretation of the response scale might have more impact on their answers than clarifying terms in the stem of opinion questions. In particular, if an interviewer defined a value in the response scale and it was at odds with the how the respondent had interpreted the scale value, the respondent could change his/her answer in light of the definition without re-executing the entire response process—he or she would just select a different answer. However, if the interviewer-provided definition corrected a respondent's misconception about the wording in the question stem, this would almost surely require more mental work by the respondent, potentially reformulating his/her attitude. Thus, clarifying aspects of the response scale could lead to more response change than defining a concept in the question stem because the former is easier to do than the latter.

Indeed, conversational interviewing produced more response change than standardized interviewing when interviewers clarified response scale labels (40% vs. 23%, $\chi^2(1) = 29.6$, $p < 0.001$) than question stems (20% vs 17%, n.s.). We recognize that this result is based on only five opinion questions, and so may depend on the particular definitions. In fact, some of our definitions for scale labels may have been less intuitive than the definitions for terms in the question stem. Specifically, the definition for Question 10 (see Appendix A) may assign more extreme values to the response scale labels than respondents would spontaneously assign ("By 'extremely risky' we mean it will cause an accident every time you do it. By 'not at all risky' we mean it never causes an accident."). Nevertheless, this definition was not responsible for the pattern of response change we observed: when answers to Question 10 were removed from the data set, the pattern of response change (i.e., more for conversational than standardized interviews) remained the same.

Our second question is whether interviewers' nonverbal sensitivity has an impact on response accuracy, interview duration, or both. If interviewers who are relatively sensitive to nonverbal cues are more adept at detecting that respondents are confused, we would expect them to either generate more response change (Hypothesis 2a) or be more efficient, i.e., take less time (Hypothesis 2b), in administering conversational interviews than their less nonverbally-sensitive counterparts.

Turning first to response change as a function of nonverbal sensitivity, there was little support for this hypothesis (2a). While interviewers who scored higher on nonverbal sensitivity produced slightly more response change in conversational

[18] The pattern of results for interviewing technique and question type remained the same after controlling for respondents' gender, age and education. There was more response change among the oldest (70 years of age and older) respondents compared to the youngest (18-44 years of age), $p = 0.01$.

[19] The main effects of question type and interviewing technique are significant when examined for respondents of different ages, different levels of education, and gender. See Table B1 in the Appendix.

Table 1
Odds Ratios of Logistic Regression Predicting Response Change

		95% C.I.	
Predictor	Odds Ratio	Lower	Upper
Conversational interviewing (vs. standardized interviewing)	1.505***	1.278	1.771
Opinion question (vs. factual question)	3.667***	3.075	4.371

*** $p < 0.001$ (two-tailed tests)

interviews (18.4% of 860 questions) than their counterparts who scored lower on nonverbal sensitivity (17.9% of 1,090 questions), this difference is not significant (OR = 1.033, $p = 0.784$).[20] However, interviewers' nonverbal sensitivity did affect the increased duration in conversational interviews relative to standardized interviews, supporting Hypothesis 2b. Interviewers high in nonverbal sensitivity completed conversational interviews faster (2.99 minutes) than did their colleagues who scored low in nonverbal sensitivity (3.49 minutes) (Wilcoxon-Mann-Whitney test: $z = -2.96$, $p = 0.003$). Therefore, although both groups of interviewers produced the same level of response change, the interviewers high in nonverbal sensitivity did this in significantly less time than low nonverbally sensitive interviewers.[21]

Why did interviewers high in nonverbal sensitivity take less time than their low nonverbal sensitivity counterparts in conversational interviews? Since providing definitions takes time, it is possible that the number of definitions provided by interviewers differed by level of nonverbal sensitivity. To examine this, when we coded the interactions between interviewers and respondents, we distinguished between two methods for delivering definitions. We refer to the first as "responsive definitions." In these situations, the interviewer provided the definition after the respondent provided some evidence of misunderstanding or confusion such as a pause, a disfluent response (e.g., containing fillers like "um" and "uh"), an unreasonably quick response, or an explicit request for help. The key point is that interviewers provided definitions when, based on respondent behavior, they judged the respondent needed help. We refer to the other method of providing clarification as "preemptive strikes" (Conrad & Schober, 2000; Mittereder, Durow, West, Kreuter, & Conrad, 2018). When interviewers engaged in preemptive strikes, they provided definitions immediately after reading the question without giving the respondent a chance to indicate that they might need help—if in fact they did. Preemptive strikes do not require the interviewer to make a judgment about the respondent's state of understanding based on nonverbal cues.

Overall, interviewers low in nonverbal sensitivity provided more definitions (42% of 1,050 conversationally administered questions) than did interviewers high in nonverbal sensitivity (33% of 740 conversationally administered questions), $\chi^2(1) = 14.35$, $p < 0.001$. Far more of the definitions were delivered as preemptive strikes by interviewers low (16% of 1,050 questions) than high (1% of 740 questions) in nonverbal sensitivity, $\chi^2(1) = 109.84$, $p < 0.001$ (see Table 2). This could well account for the shorter conversational interviews conducted by interviewers high in nonverbal sensitivity. It seems the interviewers high in nonverbal sensitivity administered conversational interviews more efficiently than those low in this ability. Consistent with this idea, interviewers high in nonverbal sensitivity provided a higher proportion of responsive definitions (32% of 740 questions) than did those low in nonverbal sensitivity (26% of 1,050 questions), $\chi^2(1) = 7.83$, $p = 0.005$.

This raises the question of whether preemptive strikes produced as much response change as responsive definitions. Because interviewers high in nonverbal sensitivity almost never administered preemptive strikes, we focused our analysis on the behavior of interviewers low in nonverbal sensitivity, in particular for the 417 question administrations in which they provided a definition of either kind. For these interviewers, respondents changed more answers after a responsive definition (35% of 257 questions) than after a preemptive strike (24% of 160 questions), $\chi^2(1) = 5.23$, $p = 0.022$. Presumably, preemptive strikes were relatively ineffective at improving respondents' understanding of questions because they were often not needed, i.e., respondents understood as intended without the preemptively provided definition and so did not change their answer.

As noted in the methods section, interviewers lower in nonverbal sensitivity tended to be older than those higher in

[20] Because interviewers were not permitted to—and rarely did—provide clarification in standardized interviews, it is not surprising that when conducting conversational interviews, each group of interviewers (high and low nonverbal sensitivity) produced more response change than was observed in the standardized interviews (12.6%), (OR = 1.556, $p < 0.001$ for high nonverbal sensitivity interviewers, and OR = 1.506, $p < 0.001$ for low nonverbal sensitivity).

[21] Even though conversational interviews conducted by high nonverbal sensitivity interviewers were shorter than those conducted by low nonverbal sensitivity interviewers, the duration of both sets of conversational interviews was longer than were the standardized interviews (2.40 minutes): low nonverbal sensitivity interviewers (3.49 minutes, $z = 8.85$, $p < 0.001$) and high nonverbal sensitivity interviewers (2.99 minutes, $z = 6.61$, $p < 0.001$).

Table 2
Type of Definition Provided in Conversational Interviews by Level of Interviewer Nonverbal Sensitivity

	Sensitivity		
Questions administered in conversational interviews with ...	Low NV %	High NV %	p value of χ^2
... any definition provided	42	33	< 0.001
... preemptive strike provided	16	1	< 0.001
... responsive definition provided	26	32	0.005
Number of questions	1,050	740	-

nonverbal sensitivity. While none of the results displayed above control for interviewer age, we reran all analyses reported here that compare the high and low nonverbal sensitivity interviewer groups controlling for interviewer age. The only result that changed was that interviewers high in nonverbal sensitivity no longer provided significantly more responsive definitions than interviewers low in nonverbal sensitivity. It seems that increased age may reduce interviewers' ability to detect respondents' need for clarification but those higher in nonverbal sensitivity still exercised more discretion in how they provided definitions producing significantly fewer preemptive strikes.

4 Discussion

There are three main findings from our study. First, we replicated the major results of previous research about conversational interviewing, only this time the technique was implemented in a production setting with SCA interviewers. Specifically, we found that a conversational re-interview produced more response change than did a standardized re-interview, reflecting improved question interpretation and, as a result, improved response accuracy compared to standardized interviewing. In addition, the conversational interviews took longer than the standardized interviews because providing clarification took time. Thus, as earlier studies have also suggested, investigators will need to weigh the costs (more time) and benefits (more accurate answers) of conversational interviewing when deciding whether and when to use the method. The familiar pattern of results in the context of a well-known production survey speaks to the robustness and viability of conversational interviewing for producing population estimates—estimates based on more accurate responses than those collected in comparable standardized interviews.

Second, respondents' interpretation of both factual and opinion questions benefitted from conversational interviewing (responses changed more for both types of questions in conversational than standardized re-administrations) suggesting that conversational interviewing is equally effective for both types of questions. Prior research on conversational interviewing has focused on factual/behavioral questions so this finding expands the set of potential items whose responses can benefit from conversational administration. As with factual questions, the interpretation of opinion questions can be standardized and consistent with what the authors had in mind if interviewers are able to ground the meaning of those questions. Additionally, defining the values in response scales appeared to be at least as effective as defining key terms in the question stem: we observed more response change for questions with definitions of response scales than attitude objects in the question stem. The finding that opinion questions produced more response change than factual questions suggests that respondents' interpretation of opinion questions may be even more variable than their interpretation of factual questions. This probably reflects people's more fluid and less concrete representation of opinions than autobiographical events and behaviors.

Third, although typical, professional interviewers are able to implement conversational interviewing effectively, some do it more efficiently than others: interviewers high in nonverbal sensitivity produced the same amount of response change as less nonverbally sensitive interviewers but they conducted the conversational interviews in less time. The difference in the duration of the conversational interviews is at least partly due to the more frequent delivery of preemptive strikes by interviewers lower in nonverbal sensitivity: their relatively indiscriminate provision of definitions suggests they were less attuned to respondents' misunderstanding and confusion than were interviewers who scored higher on nonverbal sensitivity. Indeed, interviewers low in nonverbal sensitivity delivered preemptive strikes for nearly 16% of questions whereas their counterparts delivered them for only 1% of questions. These preemptive definitions improved comprehension (i.e., led to response change) some of the time but not as high a percentage of the time as did definitions provided in response to evidence of comprehension or response difficulty, the clarification approach followed far more often by interviewers higher in nonverbal sensitivity. Apparently, interviewers low in nonverbal sensitivity provided many definitions that respondents did not need, leading to extra time with little payoff in the form of improved response accuracy.

If additional research replicates the relationship between nonverbal sensitivity and conversational interviewing efficiency, it may be possible to reduce the increased cost associated with conversational interviewing. By deploying interviewers who score higher on nonverbal sensitivity measures, interviewers should advance more quickly through the instrument because they are more likely to clarify questions only when necessary. It is an open question whether perceptual skills involving nonverbal information can be improved with training and practice, and whether trainability cuts across both spoken (e.g., disfluencies) and visual (e.g., facial expressions, gestures) indications of response difficulty. There is preliminary evidence that these skills can be improved with training: research participants trained in person perception more accurately judged others' emotions, personality traits, status, and intentions (Blanch-Hartigan, Andrzejewski, & Hill, 2016). The ability to train interviewers in nonverbal sensitivity is certainly worth exploring further.

A valuable follow-up study would investigate the connection between nonverbal sensitivity and conversational interviewing efficiency in a face-to-face environment, in which interviewers can detect audio and visual cues of confusion from respondents. While the literature on nonverbal sensitivity (e.g., Hall et al., 2005) demonstrates that individuals higher in nonverbal sensitivity are more receptive to visual cues of confusion than those lower in nonverbal sensitivity, it remains to be seen if and how these results may extend face-to-face survey interviews. For example, it is possible that visual cues of confusion are potentially more palpable than audio cues and thus may help level the playing field between high and low nonverbal sensitivity interviewers in detecting confusion when conducting conversational interviews. On the other hand, the addition of visual cues could exacerbate the difference in detecting respondents' confusion between interviewers who are high and low in nonverbal sensitivity.

A final methodological lesson: In this study, interviewers conducted both standardized and conversational interviews, shifting as needed to conversational techniques for the re-administration of questions after a standardized main interview. This contrasts with previous comparisons of conversational and standardized interviews in which interviewers conducted one technique or the other, but not both. The demonstration here that interviewers can shift between techniques suggests it may be possible to more precisely target the use of conversational interviews. For example, survey organizations could use the technique for particular respondents or particular questions for which there is reason to believe misunderstanding may be more likely. This may help reduce costs associated with conversational interviewing while maintaining the benefits.

In conclusion, we find that conversational interviewing is a broadly applicable method. It can be used when asking both opinion and factual questions, in a production telephone setting, and in combination with standardized interviewing. There is also promise in the current findings that the efficiency of conversational interviewing can be increased if survey organizations are able to recruit interviewers with above average nonverbal sensitivity or develop methods to strengthen these skills through new training methods. Overall, this is good news for survey practitioners.

5 Acknowledgement

We thank the University of Michigan Survey Research Center for supporting this research. In addition, we thank Allison Bowers, Joe Matuzak, Dylan Volans and Jennifer Kelley for their invaluable help.

References

Blanch-Hartigan, D., Andrzejewski, S. A., & Hill, K. M. (2016). Training people to be interpersonally accurate. In J. A. Hall, M. S. Mast, & T. V. West (Eds.), *The social psychology of perceiving others accurately* (pp. 253–269). Cambridge University Press.

Carney, D. R., & Harrigan, J. A. (2003). It takes one to know one: Interpersonal sensitivity is related to accurate assessments of others' interpersonal sensitivity. *Emotion*, *3*(2), 194.

Clark, H. H. (1996). *Using language.* Cambridge University Press.

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

Everitt, B., & Haye, D. (1992). *Talking about statistics: A psychologist's guide to data analysis.* New York: Halstead Press.

Feleky, A. (1914). The expression of the emotions. *Psychological Review*, *21*(1), 33–41.

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

Hall, J. A., Andrzejewski, S. A., & Yopchick, J. E. (2009). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior*, *33*(3), 149–180.

Hall, J. A., Bernieri, F. J., & Carney, D. R. (2005). Nonverbal behavior and interpersonal sensitivity. In J. A. Harrigan, R. Rosenthal, & K. R. Scherer (Eds.), *Handbook of nonverbal behavior research methods in the affective sciences*. New York: Oxford.

Kanner, L. (1931). Judging emotions from facial expressions. *Psychological Monographs*, *41*(3), i.

Mittereder, F., Durow, J., West, B. T., Kreuter, F., & Conrad, F. G. (2018). Interviewer–respondent interactions in conversational and standardized interviewing. *Field Methods*, *30*(1), 3–21.

Moore, R. J., & Maynard, D. W. (2002). Achieving understanding in the standardized survey interview: Repair sequences. In D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, & J. van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview* (pp. 281–311). New York: Wiley.

Pickett, C. L., Gardner, W. L., & Knowles, M. (2004). Getting a cue: The need to belong and enhanced sensitivity to social cues. *Personality and Social Psychology Bulletin*, *30*(9), 1095–1107.

Rosenthal, R., Hall, J. A., Di Matteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test.* Baltimore, MD: The Johns Hopkins University Press.

Schaeffer, N. C., & Maynard, D. W. (2002). Occasions for intervention: Interactional resources for comprehension in standardized survey interviews. In D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, & J. van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview* (pp. 261–280). New York: John Wiley & Sons.

Schaeffer, N. C., & Maynard, D. W. (2008). The contemporary standardized survey interview for social research. In F. G. Conrad & M. F. Schober (Eds.), *Envisioning the survey interview of the future* (pp. 31–57). New York: John Wiley & Sons.

Schober, M. F., & Bloom, J. E. (2004). Discourse cues that respondents have misunderstood survey questions. *Discourse processes*, *38*(3), 287–308.

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

Schober, M. F., Conrad, F. G., Dijkstra, W., & Ongena, Y. P. (2012). Disfluencies and gaze aversion in unreliable responses to survey questions. *Journal of Official Statistics*, *28*(4), 555–582.

Schober, M. F., Conrad, F. G., & Fricker, S. S. (2004). Misunderstanding standardized language in research interviews. *Applied Cognitive Psychology*, *18*(2), 169–188.

Schober, M. F., Suessbrick, A. L., & Conrad, F. G. (2018). When do misunderstandings matter? Evidence from survey interviews about smoking. *Topics in Cognitive Science*, *10*(2), 452–484.

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition*, *25*(5), 638–656.

Schwarz, N., Strack, F., & Mai, H. P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly*, *55*(1), 3–23.

Strack, F., Schwarz, N., & Wänke, M. (1991). Semantic and pragmatic aspects of context effects in social and psychological research. *Social Cognition*, *9*, 111–125.

Suchman, L., & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association*, *85*(409), 232–241.

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.

Sudman, S., Bradburn, N., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass Publishers.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge University Press.

Vernon, P. E. (1933). Some characteristics of the good judge of personality. *The Journal of Social Psychology*, *4*(1), 42–57.

West, B. T., Conrad, F. G., Kreuter, F., & Mittereder, F. (2018). Can conversational interviewing improve survey response quality without increasing interviewer effects? *Journal of the Royal Statistical Society: Series A.*, *181*(1), 181–203.

Appendix A
Experimental Section Questionnaire

Conversational interviewing introduction

Thank you. For the remainder of this interview, I am going to ask some questions that I may have already asked you to help us conduct these interviews better in the future. This time, when you hear the question, we want you to be sure you understand exactly what the question means. Sometimes surveys use ordinary words to mean something different than they usually mean. It is very important that you ask me to explain the meaning of any words about which you are at all unsure or confused.

Standardized interviewing introduction

Thank you. For the remainder of this interview, we are going to ask some questions that we have already asked you to help us conduct these interviews better in the future. More specifically, we want to be sure that the questions are presented exactly the same way each time they are asked by any of the interviewers on this survey. This is a very important part of conducting high quality surveys. As before, we want you to take your time and answer as accurately and thoughtfully as possible.

Question 1 (opinion)

Now turning to business conditions in the country as a whole—do you think that during the next 12 months we'll have good times financially, or bad times, or what?

1. Good times
2. Good with qualifications
3. Pro-con
4. Bad with qualifications
5. Bad times
8. Don't know

Conversational interviewing definitions

When deciding if we are having good times financially or bad times, please consider all of the following: the economy's growth, stock prices, the level of unemployment, consumer prices, interest rates and home values.

In contrast to some other questions, this question is not asking you for an answer in comparative terms, like better or worse; instead, we would like an answer in terms of good or bad. Also, we are looking for your thoughts on the country as a whole, not regionally or locally.

QxQ (standardized and conversational interviewing)

This question asks about R's expectations for the economy as a whole for the next 12 months. *Note*: It is important that you do not accept comparative answers here (i.e., "better," "worse," "same," etc.).

We are not interested in hopes and dreams, but in what R thinks will be the case. In selecting the appropriate answer category you should first decide whether R's answer is "Good," "Pro-con," or "Bad" and then whether the good or bad is qualified.

Question 2 (opinion)

We are interested in how people are getting along financially these days. Would you say that you (and your family living there) are better off or worse off financially than you were a year ago?

1. Better now
2. Same
3. Worse
8. Don't know

Conversational interviewing definitions

By "better off financially" we mean that you and your family have more left over at the end of each month than you did a year ago. Or if you're spending more than you're earning, that you're adding less debt each month than you did a year ago.

By "worse off financially" we mean that you and your family have less left over at the end of each month than you did a year ago. Or if you're spending more than you're earning, that you're adding more debt each month than you did a year ago.

By "these days" we mean the last three months.

QxQ (standardized and conversational interviewing)

This question is about how the respondent/family living there is doing financially.

We are asking R to compare the current situation to a year ago.

If R says "better off" in some respects but "worse off" in others, probe by asking, "Overall, would you say you (and your family living there) are better off or worse off financially than you were a year ago?"

It is very important to get a personal response that represents the particular situation of the R/family, so, if R answers in terms of the general economy, please use a standard probe, such as RQ, TM, or MS. If necessary, probe the R with TM to clarify if the change is within the last year.

Question 3 (factual)

During the last few months, have you heard of any favorable or unfavorable changes in business conditions?

1. Yes
2. No; haven't heard
8. Don't know

Conversational interviewing definitions

When deciding if you have heard about any changes in business conditions, please consider the following sources of information: television, radio, the internet, newspapers, magazines, and conversations with other people.

By "business conditions" we mean the profitability and prosperity of business for the country as a whole, not locally or regionally. Indicators can include earnings, sales, taxes, inflation, unemployment rates, interest rates, and government policies, laws, and regulations about business.

By "favorable" we mean anything that would increase the profitability or prosperity of business for the country as a whole. By "unfavorable" we mean anything that would decrease the profitability or prosperity of business for the country as a whole.

By "last few months", we mean the last 4 months.

QxQ (standardized and conversational interviewing)

These questions seek to elicit any "business news" the respondent has heard in the last few months for the country as a whole. If necessary (after probing) you may accept a regional/local answer. "No, I haven't heard anything" is an acceptable answer and should not be probed.

Question 4 (factual)

What is the highest grade of school or year of college you completed?

- Grades of School

  0 1 2 3 4 5 6 7 8 9 10 11 12

- College

  13 14 15 16 17+

Conversational interviewing definitions

By "highest grade completed", we want to know the number of school years completed, not the total number of years it took you to complete them. For example, if you earned an undergraduate degree over a six-year period, your answer would be 16. That is, grade twelve plus four years of college, even though it would have actually taken you eighteen years.

In your count of school years, include graduate school years completed as well as years of professional education completed, such as mechanic school or a chef's certificate.

QxQ (standardized and conversational interviewing)

We want to know the educational status of the respondent.

Grade of school or year of college: the total number of years of primary, secondary, or college completed. For example, if the R attended college part-time and took longer than a year to complete the work for a given status (freshman, sophomore, etc.), we want to know the years completed, not the total time spent.

College degree: A "degree from a junior college or community college (associate's degree) or four-year academic institution (a bachelor's degree or more)."

Specialty-type "degrees" or "certificates" (such as vocational training, mechanic's school, chef's certification, etc.) are excluded at the next question. You should select "NO" in these instances. However, we would like you to write this information in the F2 field. Please note that trade school is not college and repeat E4 if R mentions trade school.

Question 5 (factual)

How many working cell phones do you (and your family living there) have in your household? Please exclude cell phones that are for business use only.

____ Number of cell phones in HH

Conversational interviewing definitions

Please include cell phones used by persons who currently live in the household, regardless of who pays for use of the cell phone.

Exclude any children currently away at college.

Definitions of "working" cell phone: A cell phone that, if turned on, could be used at this moment to receive a phone call. Note that "working" defines whether a cell phone could be currently used to receive a call, not whether it is used for business or personal calls.

Cell phones that are placed in cars for emergency purposes only are defined as "working" if a person could receive and make a call on the phone when it is turned on. Cell phones that are not in service or are currently disabled (e.g., no SIM card) are not "working."

"Business use only" means that all calls, both outgoing and incoming, are for business purposes only and not personal use at all.

Skype, Google Voice or any other type of VOIP (internet phone) number should be counted as a landline phone number and not a cell phone number.

QxQ (standardized and conversational interviewing)

We want to include cell phones used by persons who currently live in the household, regardless of who pays for use of the cell phone.

Definition of "working" cell phone: A cell phone that, if turned on, could be used at this moment to receive a phone call. Note that "working" defines whether a cell phone could be currently used to receive a call, not whether it is used for business or personal calls.

Cell phones that are placed in cars for emergency purposes only are defined as "working" if the respondent could receive and make a call on the phone when it is turned on.

Cell phones that are not in service or are currently disabled (e.g., no SIM card) are not "working."

"Business use only" means that all calls, both outgoing and incoming, are for business purposes only and not personal use at all.

We ask these questions to try to measure how much opportunity each person has to be included in our research, and that opportunity is related to the number of telephone lines in the household. We exclude business numbers because this is a household study.

Question 6 (factual)

(In addition to your household's cell phone(s),) how many different landline telephone numbers are there in your home? Please exclude landline phone numbers that are for business use only.

0. zero
____ Number of landline phone numbers in HH

Conversational interviewing definitions

This question is about how many landline telephone numbers are in the housing unit. We are not interested in the number of telephone sets (extensions), but rather whether all the phones connect to the same telephone line (telephone number) or whether there is more than one line (telephone number) to the residence.

"Business use only" means that all calls, both outgoing and incoming, are for business purposes only and not personal use at all.

Exclude any numbers used exclusively for faxes or computers. Skype, Google Voice or any other type of VOIP (internet phone) number should be counted as a landline phone number and not a cell phone number.

QxQ (standardized and conversational interviewing)

This question is about how many landline telephone numbers are in the housing unit.

We are not interested in the number of telephone sets (extensions), but rather whether all the phones connect to the same telephone line (telephone number) or whether there is more than one line (telephone number) to the residence.

"Business use only" means that all calls, both outgoing and incoming, are for business purposes only and not personal use at all.

A VOIP (internet phone) number is a landline phone number.

We ask these questions to try to measure how much opportunity each person has to be included in our research, and that opportunity is related to the number of telephone lines in the household. We exclude business numbers because this is a household study.

Question 7 (opinion)

For the following questions, remember that by drowsy driving we mean times when someone is close to falling asleep or nodding off behind the wheel. Does driving long distances increase your own likelihood of drowsy driving these days? (Would you say yes or no?)

1. Yes
2. No

Conversational interviewing definitions

By "long distances" we mean driving 30 miles or more.

We are not asking about whether you drive more than 30 miles or more regularly, but instead are asking about your own likelihood of driving drowsy when you do drive 30 miles or more.

If you never drive long distances, we are asking about your own likelihood of driving drowsy if you were to drive 30 miles or more.

Question 8 (opinion)

Does driving in the morning increase your own likelihood of drowsy driving these days? (Would you say yes or no?)

1. Yes
2. No

Conversational interviewing definitions

By "morning" we mean from 4 AM to 10 AM.

We are not asking about whether you drive from 4 AM to 10 AM regularly, but instead we are asking about your own likelihood of driving drowsy when you do drive from 4 AM to 10 AM.

If you never drive in the morning, we are asking about your own likelihood of driving drowsy if you were to drive from 4 AM to 10 AM.

Question 9 (factual)

Over your lifetime, have you ever had an accident because of drowsy driving? (Would you say yes or no?)

1. Yes
2. No

Conversational interviewing definitions

By “accident” we mean colliding with an animal, road debris, road equipment such as a sign or light, another vehicle, or other objects; count incidents in which your vehicle hit the other object or was hit by the other object.

Question 10 (opinion)

In your opinion, how risky is drowsy driving? Would you say that it is extremely risky, very risky, somewhat risky, not too risky, or not at all risky?

1. Extremely risky
2. Very risky
3. Somewhat risky
4. Not too risky
5. Not at all risky

Conversational interviewing definitions

By “extremely risky” we mean it will cause an accident every time you do it. By “not at all risky” we mean it never causes an accident.

Appendix B
Tables

Table B1 presents odds ratios, confidence intervals around the odds ratios, and p-values for question type and interviewing technique for education, age, and gender groups. We initially constructed each of the seven models with the two main effect terms and an interaction term. None of the interaction terms were significant in any of the seven models, so we dropped the interaction term and present only the results from the seven main effects models. The main effects of Question Type and Interviewing Technique are significant for all levels of these subgroups.

Table B1
Odds Ratios of Question Type and Interviewing Technique Predicting Response Change, for Seven Separate Models with Different Cases Included

			95% Wald C.I.		
Respondents Used	*n*	Odds Ratio	Lower	Upper	*p*
Question type (reference group = factual)					
All respondents	4,748	3.67	3.08	4.37	< 0.001
College or more	2,118	4.09	3.11	5.37	< 0.001
Some college or less	2,630	3.38	2.69	4.26	< 0.001
Age 18-59	2,656	3.18	2.52	4.00	< 0.001
Age 60+	2,092	4.42	3.37	5.80	< 0.001
Female	2,598	3.31	2.62	4.17	< 0.001
Male	2,150	4.20	3.20	5.50	< 0.001
Interviewing technique (Reference group = standardized)					
All respondents	4,748	1.51	1.02	1.85	< 0.001
College or more	2,118	1.50	1.18	1.92	0.001
Some college or less	2,630	1.50	1.21	1.87	< 0.001
Age 18-59	2,656	1.61	1.29	2.01	< 0.001
Age 60+	2,092	1.39	1.09	1.77	0.008
Female	2,598	1.44	1.16	1.79	0.001
Male	2,150	1.59	1.25	2.04	< 0.001