Article Type: Original Article

When do misunderstandings matter? Evidence from survey interviews about smoking

Michael F. Schober, Anna L. Suessbrick

New School for Social Research

Frederick G. Conrad

University of Michigan

Author Note

Michael F. Schober and Anna L. Suessbrick, Department of Psychology, New School for Social Research; Frederick G. Conrad, Institute for Social Research, University of Michigan.

This material is based upon work supported by the National Science Foundation under grants No. SBR-9730140 and ITR-0081550 and by The New School for Social Research. Earlier versions of the material reported in this paper were presented at the annual meetings of the American Association for Public Opinion Research in Portland, OR (2000) and Miami Beach, FL (2005). We thank Tom Nardone, Cathy Dippo, and Bill Mockovak at the Bureau of Labor Statistics; Kimberly Clark, Ron Tucker, Mike Haas, and Fran Faull at the Bureau of the Census; Patrick Ehlen, Rene Holl, Alyssa Monnie, Maile O'Hara, and Gina Turner at the New School for Social Research; and the reviewers for thoughtful comments and questions.

Correspondence concerning this article should be addressed to Michael F. Schober, Department of Psychology, 80 Fifth Avenue, Room 710, New York, NY 10011, schober@newschool.edu

Abstract

This paper examines when conceptual misalignments in dialogue lead to consequential miscommunication. Two studies explore misunderstanding in survey interviews of the sort conducted by governments and social scientists, where mismeasurement can have real social costs. In 131 interviews about tobacco use, misalignment between respondents' and researchers' conceptions of ordinary expressions like "smoking" and "every day" was quantified by probing respondents' interpretations of survey terms and re-administering the survey questionnaire with standard definitions after the interview. Respondents' interpretations were surprisingly variable, and in many cases did not match the conceptions that researchers intended them to use. More often than one might expect, this conceptual variability was consequential, leading to answers (and, in principle, to estimates of the prevalence of smoking and related attributes in the population) that would have been different had conceptualizations been aligned; for example, fully 12% of respondents gave a different answer about having smoked 100 cigarettes in their entire life when later given a standard definition. In other cases misaligned interpretations did not lead to miscommunication, in that the differences would not have led to different survey responses. Although clarification of survey terms during the interview sometimes improved conceptual alignment, this was not guaranteed; in this corpus some needed attempts at clarification were never made, some attempts did not succeed, and some seemed to make understanding worse. The findings suggest that conceptual misalignments may be more frequent in ordinary conversation than interlocutors know, and that attempts to detect and clarify them may not always work. They also suggest that at least some unresolved misunderstandings don't matter in the sense that they don't change the outcome of the communication—in this case, the survey estimates.

When do misunderstandings matter? Evidence from survey interviews about smoking

Consider the survey question *Have you smoked at least 100 cigarettes in your entire life?*).[1] At first glance, the question seems to consist of ordinary, non-technical words that should be

[1] This is the first question in the Tobacco Use Supplement to the Current Population Survey (CPS), a major US government survey administered to a sample of 60,000 households per month from which important national statistics like the unemployment rate are derived.

easy for speakers of the language to understand; it is hard to imagine how the question could be misunderstood. But is this really the case? How often do survey respondents—and, more generally, people in conversations —interpret utterances in the same way as each other and as they were intended?

At least in the domain of survey interviews, several studies (e.g., Belson, 1981, 1986; Martin, Campanelli & Fay 1991; Wentland, 1993) have raised the possibility that people can interpret key terms in seemingly straightforward questions in surprisingly varied ways. From in-depth re-interviews about how respondents had interpreted questions in a recent interview, for example, Belson (1981) found that 16% of respondents had interpreted "you" in *How many hours of television do you watch each weekday?* to include other people, and 61% counted days other than the five weekdays. For another question, 7 of 59 respondents interpreted "over the last few years" to mean no more than two years, while 19 of 59 included ten or more years. Martin et al. (1991) found that a sample of nearly 2000 US survey respondents ranged substantially in how inclusive their conceptions of "work" were, as evidenced by their judgments of whether activities in vignettes should count as "work." For example, 38% of respondents considered that unpaid volunteer work at a local hospital qualified as work—and the rest did not.

In our own laboratory studies of interpretation in survey interviews (Schober & Conrad, 1997; Schober, Conrad & Fricker, 2004), we have demonstrated that misinterpretation of survey terms is particularly likely when the circumstances that respondents are answering about don't map neatly onto how researchers define the survey concepts. In these studies, we have relied on the fact that surveys for official government statistics develop definitions for key terms in their questions—e.g., for what they count as a "bedroom," or "work for pay," or "household furniture" for their purposes. (Everyday conversations generally don't have published definitions that elucidate a questioner's intended meaning!). The fact that there are definitions allows us to directly measure misunderstanding in this context: we can use respondents' answers to the survey questions, when we know the circumstances about which they are answering, as evidence about whether they interpreted the survey questions in ways that correspond with the survey designers' definitions. Explicit definitions also make it possible to investigate the extent to which respondents agree with *each other's* interpretations, which is independent of whether they agree with researchers' (potentially counterintuitive) definitions. Respondents might all share an

interpretation of a key term in a survey question that is different from or the same as the survey designers' interpretation, or they might vary amongst themselves in their interpretations.

These studies have demonstrated, for example, that respondents are more likely to misinterpret "bedroom" in a question about how many bedrooms are in a fictional house if one of those rooms was originally designed as a den and is now being used as a bedroom. Some respondents answer in ways consistent with the official definition (the survey from which the question is drawn defines rooms based on what they were originally designed for), but a worrisome percentage do not. Similarly, a substantial proportion of respondents (mis)interpret purchases of "household furniture" to include the purchase of a floor lamp (which for the survey from which the question was drawn should be excluded), while other respondents' interpretations are consistent with the survey designers' intentions.

How often does this kind of misinterpretation occur in survey interviews in more naturalistic conditions, and to what extent does it affect the population estimates that the surveys produce—e.g., of smoking prevalence, or employment rates, or crime victimization frequency? A methodological study using a US national telephone sample of 227 respondents (Conrad & Schober, 2000) provides some initial evidence. In that study, respondents were asked the same questions about housing and purchases excerpted from US government surveys on two different occasions; whenever respondents said "yes" to a question about purchases (e.g., "In the past five years have you purchased or had expenses for moving?"), they were asked to list the purchases on which they had based their answer. Because the first interview was strictly standardized (interviewers read the question as worded but never provided clarification even if it was requested, to avoid presenting different stimuli to different respondents), respondents could only answer based on their own interpretation of the questions. In the second interview, half the respondents participated in more collaborative interviews in which interviewers could clarify terms in the survey questions as needed. This allowed us to assess how often responses in the first interview had been based on misconceptions, and (in comparison with the baseline rate of response change for respondents interviewed twice in strictly standardized interviews) allows an estimate of how often misconceptions in the first interview had led to incorrect answers.

The evidence suggested that there was indeed substantial variability in interpretation across the sample: no question was uniformly interpreted by all respondents. The rate of response change was 11% greater for respondents in interviews in which they could obtain clarification

(above the 11% baseline rate of response change between the two strictly standardized interviews); this suggests that at least 11% of questions had been misinterpreted, leading to incorrect (later changed) answers, in the initial interviews. Analyses of the listed purchases demonstrate that a surprisingly high 43% of the listed purchases in the strictly standardized interviews did not meet the criteria for inclusion based on the official definition. Of course, this doesn't mean that 43% of the answers to the "yes-no" purchase questions were wrong; some "yes" answers no doubt should still have been "yes" because of other appropriate inclusions. But there may well also have been inappropriate omissions that we couldn't measure hidden in the "no" responses.

In the current study, our aim was to explore the prevalence of different types of misunderstandings and the consequences for survey estimates, as well as how attempts to repair misunderstanding succeed and fail, in a full-length US government survey. Our strategy was to ask respondents in the laboratory to participate in a telephone interview using an actual questionnaire deployed in its entirety (rather than excerpting questions from multiple surveys as in, e.g., Conrad & Schober, 2000; Lind et al., 2013; Schober et al., 2015), and answering questions about their own lives rather than fictional scenarios (as in, e.g., Conrad et al., 2016; Schober & Conrad, 1997; Schober, Conrad & Fricker, 2004). Using an actual complete questionnaire that included not only behavioral but also opinion questions allowed us to better quantify the potential consequences of misunderstanding in real-world social measurement, where the outcomes of the survey can have major implications for policy (Schober & Conrad, 2015). The laboratory setting allowed us to examine respondents' interpretations of survey concepts in a more detailed way than has been done previously (e.g., in Conrad & Schober, 2000), using post-interview measures that (1) assess respondents' conceptualizations and (2) how their answers might change if they used a standard definition.

Study 1: Conceptual variability and its consequences

In the study, US Census Bureau interviewers at the Hagerstown, MD, telephone center called respondents in our laboratory in New York City and carried out a strictly standardized interview in which the interpretation of survey terms was left entirely up to respondents. Then respondents filled out two self-administered (paper-and-pencil) questionnaires. The first of these assessed conceptual variability through a series of multiple-choice questions that also allowed us to

determine the extent to which respondents' interpretations matched official survey definitions. The second self-administered questionnaire re-asked the original survey questions. For half the respondents (the Definitions-in-Reinterview group) each question was accompanied by definitions of key concepts; for the other half (the No-Definitions-in-Reinterview group) the questions were simply asked with the same wording as in the interview, with no definitions. To the extent that respondents change their answers more when the re-administration of the questions included a definition than when it did not, this would suggest that respondents' conceptual variability (as measured in the first post-interview questionnaire) has consequences for survey measures. In other words, the second questionnaire allowed us to assess the extent to which conceptual variability led to misinterpretation and thus inaccurate answers in the initial survey administration.

Survey questionnaire. The survey we focused on, the Tobacco Use Supplement to the Current Population survey (CPS), is sponsored by the US National Cancer Institute and is administered occasionally (in some years) by Census Bureau telephone interviewers to all CPS households. It assesses respondents' current and previous smoking and tobacco use, as well as their opinions about related topics. Respondents answer from twelve to thirty-six questions in the same fixed order, with the only variability depending on "skip patterns" (some questions are only asked if respondents answer an earlier question in a particular way). All respondents are asked the initial behavioral "filter question" *Have you smoked at least 100 cigarettes in your entire life?* and a similar question later in the survey about *pipes, cigars, chewing tobacco, and snuff*. Only those respondents who answer "yes" to at least one of these questions are asked additional behavioral questions (e.g., *Have you ever stopped smoking for one day or longer because you were trying to quit smoking?*). All respondents are then asked a number of opinion questions (e.g., *In restaurants, do you think that smoking should be allowed in all areas, allowed in some areas, or not allowed at all?*). For the complete list of questions, see Supplementary Materials.

Although the questions in this survey all seem quite straightforward, there is no guarantee that respondents might interpret them in the same way. Has one "stopped smoking" if one temporarily stops during an illness? What counts as a "restaurant" in a question that asks whether restaurants should include outdoor seating areas and restrooms? Even the first filter question *Have you smoked at least 100 cigarettes in your entire life?* might be difficult to answer for a respondent who isn't sure whether to include clove or marijuana cigarettes, cigarettes that have

never been inhaled, or cigarettes from which only a puff or two were taken. The potential consequences of misinterpreting a filter question are particularly severe, in that this could lead a respondent to be asked the wrong questions or not asked the right questions later on.

Definitions of survey concepts. We used the survey sponsors' definitions when they existed. These included, for example, "*Past 12 months* means 12 months from today, NOT from the first of the month and not just the last calendar year." For those concepts for which the sponsors had not provided definitions, we created definitions that either conformed with the survey designers' intent, to the extent that we had evidence of it, or that seemed reasonable to us based on lab discussions of possible interpretations of key terms in the questions that led to the Conceptualization Questionnaire (see below). Each definition of a key survey term corresponded with a particular choice of possible responses for each question on the conceptualization questionnaire. An example of a definition we created is: "By *smoked* we mean any puffs on any cigarettes, whether or not you inhaled AND whether or not you finished them." Our definition of *cigarettes* included hand-rolled cigarettes as well as manufactured ones, but not cigars or non-tobacco cigarettes, like cloves or marijuana cigarettes. For the complete set of definitions, please see Supplementary Materials.

Participants. Fifty-three paid respondents (27 Female, 26 Male) were recruited, using newspaper advertising and word-of-mouth, from the New York City area and The New School community. Subjects were randomly assigned to the Definitions-in-Reinterview (n = 27) or No-Definitions-in-Reinterview (n = 26) group. Subjects ranged in education, with 16 of 53 not having a college degree, 23 with college degrees, and 14 with graduate degrees; subjects ranged in self-identified race, with 25 subjects identifying as White, 12 as Black, 4 as Hispanic, 6 as Asian, and 6 in other groups (see Supplementary Table 1 for more demographic details). The sample included smokers in all categories as assessed by the first questions in the telephone interview[2]: 17 non-smokers, 15 former smokers, 4 some-days smokers, and 15 daily smokers. Subjects in the two groups did not differ in the percentage of smokers in different categories, $X^2(3) = 1.12$, $p = 0.772$.

[2] Of course, the topic of our research and our findings lead us to question the exact accuracy of these categorizations. Nonetheless, any error in our classification of smokers in the sample should be independent of the experimental conditions to which participants were assigned because these answers preceded the experimental treatment.

Ten interviewers (8 Female, 2 Male) were recruited from the Hagerstown, MD, Bureau of the Census telephone facility. Interviewers averaged 58.4 months of interviewing experience at the Census Bureau, and there was no difference in the average experience of interviewers whose respondents were assigned to either the Definitions-in-Reinterview and No-Definitions-in-Reinterview respondents, $F(1, 8) = 0.02$, $p = 0.886$, partial $\eta^2 = 0.003$. Respondents were randomly assigned to interviewers based on interviewer availability; each interviewer conducted five or six interviews.

Interviewer training. Before the experiment was conducted, interviewers were trained on the survey concepts for about two hours. Interviewers studied the key survey concepts and then took a quiz, followed by a group discussion. Although these interviewers were instructed not to provide clarification to respondents during the survey (that is, they would be administering strictly standardized interviews), concept training allowed interviewers to know when to probe and ensured comparability with interviews in Study 2.

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

Conceptualization questionnaire. In the first of two paper-and-pencil questionnaires immediately following the telephone interview, respondents were asked their interpretations of the concepts in the survey questions they had just answered. For each survey question that they had answered (anywhere from 12 to 36 questions, depending on their answers to smoking history questions), they were asked from one to seven multiple-choice questions that asked about how they had interpreted key survey terms when they had participated in the interview. This conceptualization questionnaire was designed to explore respondents' range of interpretations, rather than to have uniformly structured response options or an equal number of concepts to be probed for each question. Thus the number of components that the conceptualization questionnaire tested varied for each question, with some items asking about more concepts

within a survey question and others asking about fewer; and what was probed varied in its complexity, so that for some concepts there were simple binary distinctions (e.g., when answering about "smoking" did you consider "all puffs whether or not you inhaled" or "only puffs on which you inhaled"), and for others there were multiple features probed in a "pick-all-that-apply" format. Respondents were presented with from 37 to 90 conceptualization items depending on how many survey questions they had answered.

The instructions on the conceptualization questionnaire asked respondents to select the response option that most closely matched what they were thinking at the time when they answered these questions on the phone, rather than what they now thought (although there is little reason to imagine that their thinking would have changed in the brief interval since they answered the telephone survey questions). They were instructed that there were no right answers, because we were testing how differently people think about these questions, and we were very interested in *their* interpretations; this was not a test of their abilities. All respondents were given the conceptualization questionnaire, and nothing in the conceptualization questionnaire instructions (nor in any other instructions throughout the study) suggested that their receiving the conceptualization questionnaire was in any way based on their performance or the quality of their answers during the interview.

Figure 1 shows the conceptualization questions about the first question in the survey questionnaire; see Supplementary Materials for the entire conceptualization questionnaire.

Reinterview questionnaire. The second paper-and-pencil questionnaire assessed the extent to which respondents' variable interpretations actually affected responses. In this self-administered "re-interview" respondents answered exactly the same questions they had answered in the original telephone interview, in the same order. Respondents in the Definitions-in-Reinterview group were instructed to answer using the official definitions; respondents in the No-Definitions-in-Reinterview group were presented with the identical questions without definitions. Figure 2 shows a sample item from the self-administered reinterview with definitions; see Supplementary Materials for the entire questionnaire. The comparable item from the self-administered re-interview without definitions simply presented the question and the response alternatives again.

RESULTS

Conceptual variability. Consistent with prior findings, our respondents varied substantially in how they interpreted the key concepts in the survey questions, but quite differently for different items. One way of looking at this variability is as the extent to which responses to each item on the conceptualization questionnaire fit (or deviated from) the standard definition used in the re-interview questionnaire. Table 1 presents examples for the items that correspond with two of the primary survey questions asked of all respondents (the smoking history questions were only asked of those who answered "yes" to these two questions).

First, it is clear that respondents' judgments agreed with the standard definitions significantly more than would occur by chance (chance levels differed for each item depending on the number of response options and combinations possible for "pick-all-that-apply" items), 42.1% agreement vs. 18.0% for chance $F(1,44) = 55.18$, $p < .001$, partial $\eta^2 = .556$ (focusing on the items that all respondents answered). This suggests that the definitions were plausible. It is also clear that agreement with the standard definition ranged enormously for different items, from none at all to 92.5%. There was no evidence that agreement with the standard definition was any different for respondents in the Definitions-in-Reinterview group (40.7%) than the No-Definitions-in-Reinterview group (43.5%), $F(1,44) = 1.78$, $p = .189$, partial $\eta^2 = .039$—which makes sense because at this point in the study no one had been presented with any definitions.

The variability in interpretation could be surprisingly large. Consider the conceptualization questionnaire items related to the first survey question *Have you smoked at least 100 cigarettes in your entire life?*. Respondents were evenly split on whether they reported having thought of "smoking" as including any puffs whether or not they inhaled (54%) or only thinking of puffs that had been inhaled (46%). For the second item, 23% of the respondents considered "smoking" to include only cigarettes that were finished, another 23% also included partly smoked cigarettes, and 54.7% also included cigarettes that they only took a puff or two from. (See Table 3 and Supplementary Table 2 for further details about the range of interpretations on items in the conceptualization questionnaire answered by participants in both studies reported here). In both cases a slight majority of respondents interpreted these concepts in ways that conformed with the standard definition for the survey, but the fact that so many respondents did not is troubling from

the perspective of survey measurement.[3] It also raises the possibility that conceptual variability is greater in ordinary conversations than speakers and listeners realize (Schober, 2005).

Reliability of responses. To what extent did this variable interpretation affect respondents' answers to the telephone survey questions? Would their answers have been different if they had been thinking about the concepts in a more uniform way—on the basis of a standard definition? The evidence shows that conceptual variability did indeed affect survey responses: Respondents who were presented with definitions in the response change questionnaire changed their answers to a significantly greater percentage of the survey questions (averaging 12.8% of those questions they answered) than the No-Definitions-in-Reinterview respondents (5.3%), $F(1,51) = 12.26$, $p = .001$, partial $\eta^2 = .194$. This base rate of 5.3% response change may seem high for an almost-immediate re-interview in which forgetting of previous answers is unlikely to be the explanation; as we see it this suggests a surprising fluidity of conceptualization, though that may have been heightened by having just spent time on a conceptualization questionnaire that suggested a range of possible interpretations for ordinary concepts. (Note that this rate of change is not unusual for a survey re-interview study [McGovern & Bushery, 1999], although those studies usually involve longer intervals between interviews). In any case, the fact that Definitions-in-Reinterview respondents changed more than twice as many of their answers as the No-Definitions-in-Reinterview respondents demonstrates that conceptual variability can have practical consequences: inaccurate answers that are almost sure to affect the population estimates derived from those answers in "production" surveys.

Can we attribute the change in answers to improved understanding? For those respondents who changed at least one answer (23 of 27 Definitions-in-Reinterview respondents and 19 of 26 No Definitions-in-Reinterview respondents), we can compare the extent to which their conceptualizations fit the standard definitions for survey questions on which their answers subsequently changed or remained the same. For each respondent we calculated conceptual fit at

[3] We do not assume that the response options in the items in our conceptualization questionnaire necessarily cover the full range of interpretations that our respondents naturally came to the study with or that the response options corresponded to all considerations relevant to their responses in the telephone interview. A think-aloud study in this line of research (see Suessbrick, 2004) demonstrated that another set of 17 similar respondents given open-ended post-interview prompts about their interpretations of these survey questions reproduced a large percentage of the conceptual distinctions tapped by our questionnaire, and demonstrated similar variability and idiosyncratic patterns of interpretation.

the question level by averaging the agreement with the standard definition for each concept in the question as tested in the conceptualization questionnaire; for example, for the first survey question this meant averaging rates of agreement with the standard definition for the three component concepts. Based on this approach, the total conceptual fit per survey question was lower for questions to which respondents later changed their answers on re-interview (average of .356) than for questions that respondents didn't change their answers to (.460), $F(1,40) = 6.32$, $p = .016$, partial $\eta^2 = .136$. This suggests that response change was more likely to occur when conceptual fit with our standard definition was poorer—i.e., that respondents corrected their initial misunderstanding.

Perhaps surprisingly, No-Definitions-in-Reinterview respondents seemed to show exactly the same effect (.366 conceptual fit on questions for which respondents later changed their answers on re-interview vs. .482 for those for which they didn't) as Definitions-in-Reinterview respondents (.346 conceptual fit on questions for which respondents later changed their answers on re-interview vs. .438 for those for which they didn't)—i.e., the relationship between conceptual fit and response change did not interact with Definitions condition, interaction $F(1,40) = .085$, $p = .773$, partial $\eta^2 = .002$. Why the No-Definitions-in-Reinterview respondents should change those answers more when they weren't presented with definitions in the re-interview is unclear; perhaps these concepts were less stable in the first place, or perhaps the conceptualization questionnaire had prompted No-Definitions-in-Reinterview respondents to question or rethink their interpretations enough to change their answers.

In any case, the findings demonstrate not only that respondents interpret ordinary terms in survey questions in substantially variable ways, but that some percentage of the time (for 7.5% of the questions they answered, here) this variability is consequential enough to lead respondents to provide different answers than they would if they were answering according to a standard definition. The findings also suggest that people's commitment to any one interpretation of concepts in survey questions is less than total; the fact that without definitions a substantial number of respondents were willing to change their answers suggests that their conceptualizations are not necessarily permanent or stable. We return to this point in the general discussion.

We propose that the phenomenon demonstrated here—that conceptual misalignment between conversing parties is only sometimes consequential—is a feature of referring in ordinary

conversation that interlocutors only sometimes notice. Consider, for example, a host who invites a dinner guest who explains to her that he is vegetarian, and as a result she intends to cook him a vegetarian meal; in their conversation using the word "vegetarian" they both felt they had understood each other. But it turns out they are conceptually misaligned about what "vegetarian" means: she happens to think that while "vegetarian" means "no meat" it includes eating fish, but this vegetarian considers "vegetarian" not to include fish-eating. If the host happens to cook only vegetables for the guest, this conceptual misalignment will likely be undetected and will have no consequences. If the host cooks fish, however, the misalignment will indeed be detected through the socially awkward outcome.

As we see it, misunderstanding in survey interviews has the same structure: some conceptual misalignments have consequences in the world (the host's preparing fish for a non-fish-eating guest, an answer being given in the survey that would have been different if the respondent's and interviewer's conceptions were aligned), and some do not. Another way of saying this is that some conceptual misalignments lead to misunderstanding and others do not. So, for example, a misalignment on what counts as "smoking" for someone who has never touched a cigarette in their lives, or for a chain-smoker, is unlikely to change her answer to *Have you smoked at least 100 cigarettes in your entire life?*. But a misalignment could well change the answer for someone who never buys cigarettes but occasionally smokes part of a friend's, or for someone who has only ever smoked marijuana but not tobacco cigarettes.

Undetected conceptual misalignments that don't have consequences probably aren't important to repair, or even notice. Undetected conceptual misalignments with consequences (uncomfortable dinners, survey responses that would have been different) *are* important to repair, in order to avoid the consequences. The next study explores how this kind of repair does—and doesn't—work in the survey interview setting, where misalignment can potentially be repaired (and response accuracy improved) when interviewers provide clarification about key survey concepts *during* the interview (Conrad & Schober, 2000; Schober & Conrad, 1997; Schober, Conrad, & Fricker 2004).

Study 2: When do clarification attempts resolve misunderstanding?

Our view of conceptual (mis)alignment and its potential consequences suggests several possibilities for how clarification efforts can result in successful or unsuccessful communication. Table 2 lays out our view of the logical possibilities in the context of survey interviews, where respondents and researchers[4] can start out aligned or misaligned on a survey concept, and respondents can either get additional evidence about the researcher's definition (attempted clarification) during the interview or not. Depending on the circumstances about which a respondent is answering (e.g., whether they unambiguously smoke tobacco cigarettes daily or not), the definition could be relevant and helpful or not. The different "paths" laid out in the table lead to end states of alignment or nonalignment by the time the survey question is answered, and to survey responses that are accurate (match what the survey designers' definitions require) or not. The key takeaway point from Table 2 is that a number of alternative clarification pathways (Rows 2-8) are possible beyond the prototypical cases (Rows 1 and 9) that easily come to mind.

In Study 2, we used the same experimental setup as in Study 1 to conduct additional interviews in which the telephone interviewers could provide definitions *during* the initial interview. The intention was to gather enough cases, using a full-length survey questionnaire, of conceptual misalignments that would and would not be addressed by definitions during the course of the interview, and to assess the impact of those definitions on ultimate alignment and the responses. In other words, we hoped to gather enough data using this paradigm to be able to statistically compare the prevalence of the different paths in Table 2 in an actual survey, and to estimate the effects on accuracy of measurement of clarification during an interview through the lens of conceptual misalignment.

Procedure. Almost everything about Study 2's materials and procedure—the laboratory setting for the telephone surveys, the survey questions, conceptualization questionnaire, response change questionnaire—was the same as in Study 1. What was different were the two interviewing procedures that telephone interviewers were trained in and that respondents were instructed to participate in. In Respondent-Initiated Clarification interviews, interviewers were trained to define a survey term only when clarification was explicitly requested by the

[4] This account doesn't distinguish between the potentially differing conceptualizations of survey designers and the interviewers who administer those surveys, though of course they might also be misaligned. So there are further layers of complexity to consider in examining misunderstanding in survey interactions. For current purposes we assume the interviewers implement the researchers' intentions and so are aligned.

respondent (e.g., *What do you mean by "every day"?*). In Mixed-Initiative Clarification interviews, interviewers were empowered to provide clarification whenever they believed that the respondent was in danger of misinterpreting an important survey term, even if the respondent hadn't explicitly requested it. In this study all respondents' reinterview (self-administered) questionnaires included the standard definitions, as in the Definitions-in-Reinterview condition in Study 1.

Participants. Nine interviewers (7 female), all new to this study, averaging 59.1 months interviewing experience, were recruited from the same Census Bureau telephone facility. They were randomly assigned to one of the two interviewing techniques; each interviewer conducted five or six interviews, except for one interviewer who conducted ten. An additional four novice interviewers (3 female), also new to this study, were subsequently recruited from the Psychology graduate student population at The New School to administer Mixed-Initiative Clarification interviews once it became clear that the Census Bureau professional interviewers had not provided clarification at a sufficiently high rate for the study's purposes.

78 paid respondents were recruited from the New York City area and The New School community using the same methods as in Study 1. These participants did not differ demographically (in age, gender, education, ethnicity, or smoking status) from the participants in Study 1 (see Supplementary Table 1). The 51 respondents interviewed by the professional interviewers were randomly assigned either to Respondent-Initiated Clarification interviews (n=25) or Mixed-Initiative Clarification interviews (n=26); the remaining 27 respondents interviewed by the novice interviewers were all assigned to Mixed-Initiative Clarification interviews. We do not distinguish between the type of interviewer (experienced or novice) conducting Mixed-Initiative Clarification interviews in the analyses we report.

RESULTS

Transcripts of all the audiorecorded interviews in Study 1 and Study 2 (except one with recording failure, for a total of 2567 question-answer sequences across 130 interviews) were coded in Sequence Viewer (http://www.sequenceviewer.nl/) using a coding scheme developed for these kinds of survey interviews with and without clarification (see Schober et al. 2012, Appendix B, for details of the scheme).

In the Study 2 interviews, respondent requests for clarification were rare, despite the interviewer's instructions and encouragement that clarification could be helpful: 56 of the 78 Study 2 respondents never asked for clarification, 15 asked once during the entire interview, and only 7 asked for clarification on more than one question. Interviewers in the Respondent-Initiated Clarification interviews provided clarification correspondingly rarely, for 1.1% of question-answer sequences. Mixed-Initiative Clarification interviewers presented clarification substantially more often, for 30.2% of question-answer sequences for each respondent, on average (ranging from 0% to 82.4% of the questions in an interview), $F(1, 72) = 42.67$, $p < .001$, partial $\eta^2 = .378$.

Conceptual variability. The corpus of 131 interviews in Studies 1 and 2 allows us to observe a larger sample of the range of conceptual variability across the concepts in this survey, as measured by our conceptualization questionnaire. Table 3, which combines data from both studies for the same conceptualization questions in Table 1, demonstrates that the range observed in Study 1 is robust, with a degree of interpretative variability and levels of agreement with the standard definitions that were quite similar across both studies—even though some definitions were given during some of the interviews. (Supplementary Table 2 presents the data for all questions answered by at least half the respondents).

Table 3 also demonstrates that for some survey concepts a large majority of respondents in our sample agreed on one interpretation (whether that was the same as the standard definition or not), but that for others the range of possible interpretations was large. For example, on the very first question, all respondents considered *cigarettes* to include cigarettes that were finished, but only 72.3% included partially smoked cigarettes, and only 53.8% included cigarettes with only a puff or two taken; 98.5% of respondents counted manufactured cigarettes, but only 52.3% considered hand-rolled cigarettes, and a notable proportion counted marijuana cigarettes (16.9%), non-tobacco cloves cigarettes (26.9%), and even cigars (23.8%). Almost all respondents (97.7%) considered *pipe smoking* to include smoking pipe tobacco, but a number also reported having considered pipe smoking to include smoking hashish (10.9%), crack (11.6%), and marijuana (17.1%). As shown in Supplementary Table 2, in answering the question about whether smoking should be allowed in hospitals, most respondents reported having considered waiting rooms (85.3%) and patient rooms (82.2%), but fewer considered other public areas like elevators (59.7%) and rest rooms (47.3%), and many included non-public areas

like staff lounges (65.9%). Of course, just from these data we can't tell whether this range of interpretation actually affected the responses to the survey questions--e.g., whether responses in the interview would have been different if respondents had considered elevators and rest rooms in hospitals, or not included hashish in their answer about pipe smoking; this is what our reinterview questionnaire can assess.

To what extent did definitions presented during the interviews reduce this conceptual variability? Contrary to our expectations, we saw no overall evidence that post-interview conceptual fit with the standard definition, as measured in the conceptualization questionnaire, was any better for questions where a definition had been given during the telephone interview than for questions where a definition had not been given, nor that conceptual fit was better in Mixed-Initiative Clarification interviews than in Respondent-Initiated Clarification interviews (where definitions were almost never presented). Closer inspection of the conceptualization questionnaire results when paired with the interview transcriptions made clear that the component of a survey definition that was presented in an interview was only sometimes relevant to the respondent's misaligned conceptualization or the respondent's circumstances. So even though we had intended to build a sufficiently large corpus of interviews to cleanly test our hypotheses about the effects of definitions during interviews, we seem to have ended up with a relatively sparse data set once one considers how respondents' interpretations on the many concepts we were testing combined with the autobiographical circumstances about which they were answering (their smoking behaviors and opinions), crossed with whether they happened to receive a definition during the interview or not and whether the component of the definition they received was relevant or not.

Although we therefore have too few cases of definition-giving where the definition that was given unambiguously (for us as researchers) corresponds with the particular respondent's circumstances and misaligned interpretation, our data set does allow further exploration of our various hypothesized clarification-during-interview pathways.

Reliability of responses. 65 of the 74 Study 2 respondents for whom we had response change data changed their answers to at least one survey question when presented with a definition on the re-interview questionnaire, on average changing 14.5% of the answers to the questions they had answered (max change 46.2%). The rate of response change varied substantially by survey question, averaging 14.9% of responses changing per question when definitions were later

provided (max change 50%, for *About how long has it been since you last smoked cigarettes EVERY DAY?* and *When you last smoked every day, on average how many cigarettes did you smoke daily?).* For 8 questions there was no response change on the re-interview questionnaire at all, for example for *Have you EVER stopped smoking for one day or longer because you were TRYING to quit smoking?*, *In the PAST YEAR have you SEEN a medical doctor?*).

Perhaps most strikingly, 12.3% of Study 2 respondents changed their answer to the very first question in the survey, *Have you smoked at least 100 cigarettes in your entire life?*, from "yes" to "no" or from "no" to "yes" when presented with a standard definition in the re-interview questionnaire. (This is in clear contrast to the 0% of respondents in Study 1 who were not given a definition in the re-interview questionnaire changing their answers). The extent of unreliability on this first "filter" question has consequences that go beyond potential mismeasurement of smoking prevalence, in that respondents whose answers to this question are unreliable were likely sent down the wrong path of the questionnaire (nonsmoker instead of smoker or vice versa) and thus were probably asked a number of additional questions for which their answers may not be relevant or not asked questions for which their answers would have been relevant. This level of misunderstanding at a crucial point in a survey (as in any conversation) can snowball into downstream consequences for what is and is not further taken up that is likely to affect other estimates. In more complex questionnaires with more complex branching structures, this could lead to a combinatorial explosion of measurement error.

Unlike in our prior studies in which respondents answered questions about their own lives (Conrad & Schober, 2000; Schober et al., 2012), we did not see significant differences in response change on the post-interview questionnaire with definitions for question-answer sequences in which definitions had been given in the telephone interview relative to question-answer sequences in which definitions had not been given. That is, response change was equivalent for the standardized interviews in Study 1, the Respondent-Initiated Clarification interviews in Study 2, and the Mixed-Initiative Clarification interviews in Study 2. We do not know whether this is because clarification was rarer than in our earlier studies, because the clarification was provided with less sensitivity to respondents' needs, or because our sample size was too small given the range of respondents' circumstances relevant to these particular survey questions and the range of conceptualizations they came to the study with.

Clarification pathways during the interviews. Although we did not have direct evidence about survey response accuracy, we did have evidence in the interview transcripts about responses and the presentation of definitions. This evidence could be connected with each respondent's answers on the conceptualization questionnaire and on the re-interview questionnaire. We therefore had the basis for making plausible (not definitive) inferences about which sequences were consistent with at least a subset of our hypothesized pathways.

Table 4 presents examples of transcript excerpts with our reasoning about which alignment paths they could reflect given the evidence available in this study. Because we don't have evidence about pre-interview conceptualizations, our evidence leaves some ambiguity about whether, for example, what looks like a successful clarification during the interview was actually a presentation of a superfluous (unneeded) definition; in both cases the post-interview conceptualization is fully aligned with the researchers' definition and the re-interview response doesn't change. Similarly, it is unclear whether what looks like an unsuccessful needed clarification attempt (missing out on the component that needed clarification) should instead count as a harmful definition that "spoiled" a good answer; in both cases the re-interview response is different from the response during the interview and the interviewer had provided a definition. Consider this question-answer sequence, where the interviewer interjects an unsolicited definition ("and that includes only indoor areas") while asking the respondent's opinion about where smoking should be allowed:

I: How about bars and cocktail lounges. And that includes only indoor areas. Would that be allowed in s- all areas? allowed in some areas, or not allowed at all.

R: Some areas.

In this case, the re-interview response changes to "not allowed at all," and the conceptualization questionnaire evidence shows that the discrepancy is in what the respondent counted as "smoking" rather than what areas of bars and cocktail lounges should be included. The interviewer's clarification failed to address the critical relevant conceptual misalignment, but without our having evidence on the respondent's pre-interview conceptualization we can't rule out that the presentation of the definition might actually have changed or harmed the respondent's initial aligned conceptualization.

Even though we do not have access to pre-interview conceptualizations that could disambiguate such cases, our available evidence still allows us to quantify the distribution of different pathways in our corpus. Table 5 shows the distribution of pathways of the 1,922 interview sequences for all respondents in Studies 1 and 2 whose reinterview questionnaire included definitions. As Table 5 shows, not only are these pathways logically possible but they *do* seem to occur in an actual survey. The good news for survey researchers is that a majority of question-answer sequences here (87.7%) are likely unproblematic, in the sense that the final interview response reflects conceptualizations sufficiently aligned with the researchers' for the purposes of the survey—even if irrelevant misalignments were not corrected. Another practical implication that may be reassuring for survey researchers is that, on balance, interviewers' providing definitions helps more than it hurts: of the 300 question-answer sequences in which definitions were provided, responses were reliable for 83.3% of them (250). Nonetheless, the fact that 12.2% of responses were unreliable raises the concern that important conceptual misalignments can remain undetected, and that not all attempts to improve matters succeed.

Discussion

Taken together, the findings in these two studies demonstrate that people can interpret at least some of the ordinary words in a survey in surprisingly variable ways, far beyond what is apparent to them and to researchers. And some percentage of the time this conceptual variability can lead to consequential misunderstanding: to interpretations of survey questions and of how the terms in those questions map onto the survey respondent's circumstances that differ from what a survey designer intends, and thus to inaccurate answers—and, in turn, to inaccurate summary descriptions of the population based on those answers. But conceptual misalignment doesn't *necessarily* lead to problematic (unreliable, inaccurate) responses: sometimes respondents' interpretations can differ from survey designers' but in ways that would not change the respondent's answers if their interpretations were aligned, and so the misalignment is not always (functionally) a problem.

To put it another way, conceptual misalignment doesn't necessarily lead to a (functional) misunderstanding: a survey response can be perfectly adequate—accurate—despite survey respondent and researcher holding different interpretations. Whether this should be considered "misunderstanding" because both parties aren't perfectly aligned, or successful understanding

despite misalignment, in that the dialog task or project was achieved sufficiently for current purposes, depends on how one defines misunderstanding. On the one hand, misalignment that doesn't have dire consequences for the task at hand may not qualify as misunderstanding; on the other hand, misalignment (despite task success) can be considered, technically, misunderstanding, in that it may contain the seeds for potential future task failure.

The findings also demonstrate that attempts at clarification—attempts to ground understanding (Clark & Brennan, 1991; Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989)—in survey interviews can work quite differently depending on (a) which aspects of respondents' conceptualizations of survey terms misalign with researchers' and (b) their actual circumstances or opinions. While many clarification attempts are successful or at least not harmful (they don't "spoil" a good answer), some needed attempts are never made, some attempts do not succeed, and some may actually make understanding worse (that is, they lead to an unreliable answer). In particular, clarification can only be helpful if it pertains to the dimension of a question concept that is relevant to the respondent's circumstances (the behavior or opinion about which she is reporting). Irrelevant clarification might even lull respondents into the belief that their interpretation of anything that wasn't clarified must be accurate.

Methodological questions

The findings and conclusions presented here are based on the particular methods we chose, which give one view of how respondents were thinking during interviews and how far off their conceptualizations might be from researchers'. Our strategy was to allow the interviews to proceed as naturally as possible in the lab setting, with respondents interviewed by professional telephone interviewers, answering about their own lives and opinions (rather than experimentally designed vignettes), and following an actual survey's skip patterns. By design, we did not probe respondents' conceptualizations of survey terms before the interview, to avoid the possibility that such probing could affect their ordinary thinking about the survey terms during the interview. But the post-interview measures—conceptualization questionnaire and re-interview-with-definitions questionnaire, administered in a different mode (self-administered paper-and-pencil) than the telephone interviews—are likely far outside the ordinary interview experience for most respondents, and very likely led to non-typical reflection about question meaning. This raises the

question of whether something about the method led to overstatement of respondents' conceptual variability or the effectiveness of clarification during an interview.

More specifically, the fact that the conceptualization questionnaire required respondents to think about conceptual distinctions that they may not have considered at all during the interview raises the possibility that respondents' choices on the conceptualization questionnaire reflect processes beyond those that were at play during the interview. It is also possible that responses to the reinterview-with-definitions questionnaire could have been affected by the extra doubts or considerations raised in the conceptualization questionnaire, or by the potential pragmatic implication from merely being asked the reinterview questions that one ought to change one's answer. As we see it, our method allows us to probe into conceptual discrepancies that are otherwise hard to uncover, but how general and stable the discrepancies we observe are is unknown. It is also unknown how the findings would extend beyond our convenience sample of New York City participants to a representative sample of the population. In any case, the fact that all respondents in both studies were given the *same* post-interview measures in the same order, in the same mode (paper-and-pencil), and with the same post-interview delay means that our pattern of findings—differences across interviewing conditions—cannot be attributed to differential administration of our (admittedly unusual) measures. And the fact that our participants were randomly assigned to experimental conditions also suggests that our pattern of findings is unlikely to have resulted only from the particular characteristics of our sample.

Generalizability and implications for theories of dialog

To what extent do our findings generalize to other communicative settings? Comprehension of terms in surveys is clearly related to comprehension of references in other arenas of interaction: terms are offered for an addressee—the respondent—to comprehend and use, just as references are offered in everyday dialog to be understood at the level of specificity that fits the interlocutors' current purposes (Clark, 1996; Clark & Wilkes-Gibbs, 1986). The phenomenon of interpretive variability we observe in these survey questions may well be related to interpretive variability of language more generally (see Kurtz and Schober, 2001, for evidence on how fiction readers' interpretation of themes in short stories can be surprisingly divergent). But survey interviews have particular features (Schober & Conrad, 2002; Houtkoop-Steenstra, 2000; Schaeffer, 2002; Schaeffer & Maynard, 2008) that make them distinct from other kinds of dialog

or language comprehension settings, beyond the fact that they are consequential for social measurement (Schober & Conrad, 2015). In particular, unlike in references to objects in physical settings, in which interlocutors can have immediate evidence about what their partner means and when understanding has gone wrong, the autobiographical circumstances (behaviors and opinions) about which survey respondents answer are not immediately visible to the researcher. Survey interviewers are intermediaries for survey researchers, which changes their responsibility for the meaning of what they say (Clark & Schober, 1992; Conrad, Schober & Schwarz, 2014). And survey respondents are not the initiators of the references in the questions, which can lead them to be less likely to question whether their own interpretations of terms might be different than their interlocutor's (Schober, et al., 2003). How comparable survey dialog is to the range of different kinds of everyday conversation is, as we see it, an open question.

Nonetheless, survey respondents *do* bring their ordinary linguistic and interactive repertoire to the survey setting, and so the kinds of misunderstandings observable in surveys are likely informative about misunderstanding in other dialog settings. As we see it, referential communication in most conversational settings that we can think of allows task success despite "undetected conceptual misalignment" (Schober, 2005) in much the way we observe here. As modeled in laboratory referential communication tasks (maze games, figure-matching tasks), interlocutors can succeed at the tasks (accurately finding their way, selecting the right ambiguous figures) without having to perfectly agree on the detailed conceptualizations underlying every term they use; local "conceptual pacts" that speakers in dialog establish (Brennan & Clark, 1996) aren't exhaustive, nor are they permanent or even necessarily generalizable to other conversational partners. Undetected conceptual misalignment may extend beyond spoken dialog to other forms of interdependent action; for example, jazz improvisers can play together without conceiving of all their individual contributions or their joint product in the same way as each other (Pras, Schober, & Spiro, 2017; Schober & Spiro, 2014, 2016).

The implication for models of dialog more generally is that, to the extent that the conceptual variability and varying effectiveness of clarification attempts observed here extends more broadly, our findings are more consistent with dynamic or situation-specific views of the nature of meaning in dialog (e.g., Larsson 2008) than views that assume stable representations that extend across circumstances. They are also consistent with demonstrations that judgments about category membership can differ between people and that, within people, concepts can be fluid

across different circumstances (e.g., Barsalou, 1983; Smith, 2005). For theories of dialog, our data raise questions about the extent of overlap of speakers' individual networks; in terms of Pickering and Garrod's (2004) model, dialog participants' lexical and semantic networks may be just different enough that what looks and feels like alignment may actually be farther off than either party can tell based on the dialog evidence. Beyond that, the "basic interactive repair mechanisms" of dialogue (Pickering and Garrod, 2004) do not necessarily detect everything that might need to be repaired, and they can address concerns that aren't relevant or important to the current task.

Practical implications for survey researchers

What are the practical implications for social research that administers standardized surveys? To the extent that our laboratory findings generalize to fully representative samples of the sorts targeted by survey researchers, the findings suggest that researchers and interviewers should be aware that conceptual variability may be greater than they assume and that there may well be response-relevant conceptual misalignments that need to be uncovered if they want data that fully embody their standard definitions. Our findings also suggest that extra attention to the potential for conceptual variability in responses to “filter” questions that are consequential in branching would be particularly useful. That said, as we have argued elsewhere (Conrad & Schober, 2000) the various possible solutions for addressing conceptual misalignment each have their own downsides. Including definitions that cover all possible misalignments in scripted question wording may well be impossible, though attending to the most frequent misalignments for filter questions may be feasible if survey researchers have the time and resources for careful pretesting. Simply providing definitions when they are requested, or even when the interview suspects they are needed even though the respondent hasn't asked, will only cover a subset of misalignments; respondents in our study almost never requested clarification.

And the effects of clarification attempts are more complex than they at first seem, with no absolute guarantee of success. As our findings show, clarification attempts can lead to extra dialog work without actually improving response accuracy, sometimes they can make things worse, and sometimes they may even mislead by suggesting that the only points that needed clarification were those that were addressed in the clarification attempt. Nonetheless, on balance

our evidence suggests that clarification attempts—and empowering interviewers to provide clarification—are worth it more often than not. Our evidence is also consistent with findings of objectively improved data quality (compared with administrative records) in national surveys that encourage clarification (e.g., West et al., 2018).

It may be that alerting survey participants—interviewers and respondents—that undetected conceptual misalignment can lead to misunderstanding, inaccurate responses, and ultimately inaccurate survey estimates is the most important intervention, along with giving them evidence about the ways clarification can go right and go wrong. The silver lining in the evidence from this study is that the consequences of misunderstanding are not always serious and that clarification on balance helps. More generally, even though clarification attempts don't guarantee full conceptual alignment, some misunderstandings simply don't matter.

References

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 271-227.

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

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

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493.

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

Clark, H. H., & Schober, M. F. (1992). Asking questions and influencing answers. In J. M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys* (pp. 15–48). New York: Russell Sage Foundation.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Conrad, F. G., & Schober, M. F. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly*, *64*(1), 1–28. http://doi.org/10.1086/316757

Conrad, F. G., Schober, M. F., Jans, M., Orlowski, R. A., Nielsen, D., & Levenstein, R. (2015). Comprehension and engagement in survey interviews with virtual agents. *Frontiers in Psychology*, *6*(October), 1578. https://doi.org/10.3389/fpsyg.2015.01578

Conrad, F. G., Schober, M. F., & Schwarz, N. (2014). Pragmatic processes in survey interviewing. In T. Holtgraves (Ed.), (pp. 420–437). New York: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199838639.013.005

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

Houtkoop-Steenstra, H. (2000). *Interaction and the standardized survey interview: The living questionnaire*. Cambridge: Cambridge University Press.

Kurtz, V., & Schober, M. F. (2001). Readers' varying interpretations of theme in short fiction. *Poetics*, *29*(3), 139–166. http://doi.org/10.1016/S0304-422X(01)00040-7

Larsson, S. (2008). Formalizing the dynamics of semantic systems in dialogue. In R. Cooper & R. Kempson (Eds.), *Language in flux: Dialogue coordination, language variation, change and evolution* (pp. 121–142). London: College Publications.

Lind, L. H., Schober, M. F., Conrad, F. G., & Reichert, H. (2013). Why do survey respondents disclose more when computers ask the questions? *Public Opinion Quarterly*, *77*(4), 888–935. https://doi.org/10.1093/poq/nft038

Martin, E.A., Campanelli, P.C., & Fay, R.E. (1991). An application of Rasch Analysis to questionnaire design: Using vignettes to study the meaning of `work' in the Current Population Survey. *Journal of the Royal Statistical Society. Series D (The Statistician), 40*(3), 265-276.

McGovern, P., & Bushery, J.M. (1999). Data mining the CPS interview: Digging into response error. In *Proceedings of the Research Conference of the Federal Committee on Statistical Methodology, Monday B Sessions*, pp. 76–85. Washington, DC: Federal Committee on Statistical Methodology.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *The Behavioral and Brain Sciences*, *27*(2), 169-190. https://doi.org/10.1017/S0140525X04000056

Pras, A., Schober, M. F., & Spiro, N. (2017). What about their performance do free jazz improvisers agree upon? A case study. *Frontiers in Psychology*, *8*, 966. https://doi.org/10.3389/fpsyg.2017.00966

Schaeffer, N. C. (2002). Conversation with a purpose—or conversation? Interaction in the standardized interview. In D. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, & H. van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview* (pp. 95–124). Hoboken, NJ: Wiley.

Schaeffer, N. C., & Maynard, D. W. (1996). From paradigm to protoype and back again: Interactive aspects of cognitive processing in survey interviews. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey interviews* (pp. 65–88). San Francisco: Jossey-Bass.

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). Hoboken, NJ: Wiley.

Schober, M. F. (2005). Conceptual alignment in conversation. In B. F. Malle & S. D. Hodges (Eds.), *Other minds: How humans bridge the divide between self and others* (pp. 239–252). New York: Guilford Press.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232. http://doi.org/10.1016/0010-0285(89)90008-X

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, *61*(4), 576–602. http://doi.org/10.1086/297818

Schober, M. F., & Conrad, F. G. (2002). A collaborative view of 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. 67–94). Wiley.

Schober, M. F., & Conrad, F. G. (2015). Improving social measurement by understanding interaction in survey interviews. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 211–219. http://doi.org/10.1177/2372732215601112

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., Ehlen, P., & Fricker, S. S. (2003). How web surveys differ from other kinds of user interfaces. In *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp. 190–195).

Schober, M. F., Conrad, F. G., & Fricker, S. S. (2004). Misunderstanding standardized language in research interviews. *Applied Cognitive Psychology*, *18*(2), 169–188. http://doi.org/10.1002/acp.955

Schober, M. F., & Spiro, N. (2014). Jazz improvisers' shared understanding: A case study. *Frontiers in Psychology*, *5*, 808. http://doi.org/10.3389/fpsyg.2014.00808

Schober, M. F., & Spiro, N. (2016). Listeners' and performers' shared understanding of jazz improvisations. *Frontiers in Psychology*, *7*, 1629. https://doi.org/10.3389/fpsyg.2016.01629

Smith, L. B. (2005). Emerging ideas about categories. In L. Gershkoff-Stowe & D. H. Rakison, (Eds.), *Building object categories in developmental time*, 159-173. Mahwah, NJ: Erlbaum.

Suessbrick, A. (2004). *The limits of clarification in coordinating conceptualizations in discourse*. Doctoral dissertation, New School for Social Research, New York, NY.

Wentland, E. J. (1993). *Survey responses: An evaluation of their validity*. New York: Academic Press.

West, B.T. et al., 2018. Can conversational interviewing improve survey response quality without increasing interviewer effects? *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 181(1), pp.181–203.

Table 1. Percentage of respondents whose answers on conceptualization questionnaire matched the standard definition, relative to chance responding, for items related to two survey questions answered by all respondents. Response options corresponding to standard definition are in bold.

	No-Definitions-in-Reinterview	Definitions-in-Reinterview	Overall	Number of response options	Possible response combinations	Chance responding
S32. Have you smoked at least 100 cigarettes in your entire life?						
When you answered the question, did you interpret "smoking" to include:	52.0%	55.6%	53.8%	2	2	50.0%
(a) Only puffs that you inhaled						
(b) **Any puffs, whether or not you inhaled**						
How did you interpret "cigarettes"? (*Circle all*	57.7%	51.9%	54.7%	3	7	14.3%

that apply) **(a) Cigarettes that you finished** **(b) Cigarettes that you partially smoked** **(c) Cigarettes that you only took a puff or two from**						
Did you interpret "cigarettes" to include (*Circle all that apply*): **(a) Manufactured cigarettes** **(b) Hand-rolled cigarettes** (c) Marijuana cigarettes (d) Cigars (e) Cloves (or other non-tobacco) cigarettes (f) Something else. Specify: _____	11.5%	11.1%	11.3%	6	59	1.7%

S62a. Have you EVER used pipes, cigars, chewing tobacco or snuff?						
When you were answering this question, did you consider "ever used" to mean (*Pick one*):	69.2%	74.1%	71.7%	4	4	25.0%
(a) Trying even once (one puff, chew, or sniff)						
(b) Trying several times						
(c) Using regularly						
(d) Something else. Specify:_____						
Did you consider pipe smoking to include (*Pick all that apply*):	92.3%	66.7%	79.2%	5	31	3.2%
(a) Pipe tobacco						
(b) Hashish						
(c) Crack						
(d) Marijuana						

(e) Something else. Specify:_____						
Did you consider "using chewing tobacco" to mean (*Pick all that apply*): **(a) Placing it in the mouth** (b) Something else. Specify: _____	92.3%	92.6%	92.5%	2	3	33.3%
Did you consider "using snuff" to mean (*Pick one*): **(a) Only sniffing it** (b) Only placing it in the mouth (c) Either sniffing it or placing it in the mouth (d) Something else? Specify:_____	38.5%	33.3%	35.8%	4	4	25.0%

Table 2. Logically possible paths from initial to final state of conceptual alignment during a question-answer sequence.

Path	Starting conceptual alignment	Starting misalignment relevant to survey response?	Respondent receives definition?	Definition helpful?	Final conceptual alignment	Final alignment relevant to survey response?	Survey response accurate?
1. Prototypical Q-A sequence (no clarification)	**Aligned**	**--**	**No**	**--**	**Aligned**	**Yes**	**Yes**
2. Confirmatory (or superfluous) definition	Aligned	--	Yes	No	Aligned	Yes	Yes
3. Harmful definition	Aligned	--	Yes	No	Misaligned	Yes	No
4. Irrelevant definition	Aligned	--	Yes	No	Misaligned	No	Yes
5. Uncorrected misalignment irrelevant to response	Misaligned	No	No	--	Misaligned	No	Yes
6. Uncorrected misalignment relevant to response	Misaligned	Yes	No	--	Misaligned	Yes	No
7. Unsuccessful but unnecessary clarification attempt	Misaligned	No	Yes	No	Misaligned	No	Yes

8. Unsuccessful needed clarification attempt	Misaligned	Yes	Yes	No	Misaligned	Yes	No
9. Prototypical successful clarification	**Misaligned**	**Yes**	**Yes**	**Yes**	**Aligned**	**Yes**	**Yes**

Table 3. Conceptualization questionnaire responses to items related to two survey questions by all respondents in both studies. (See Supplementary Table 2 for details on all items answered by at least half the respondents). Response options that fit the standard definitions are in bold.

S32. Have you smoked at least 100 cigarettes in your entire life?

		Total respondents selecting option		"Pick all that apply" combinations	
When you answered the question, did you interpret "smoking" to include:		n	%	n	%
(a) Only puffs that you inhaled	a	60	46.2%		
(b) Any puffs, whether or not you inhaled	**b**	**70**	**53.8%**		
	Total	130	100.0%		

How did you interpret “cigarettes”? (*Circle all that apply*)							
(a) Cigarettes that you finished	**a**	**130**	**100.0%**	a	35	26.9%	
(b) Cigarettes that you partially smoked	**b**	**94**	**72.3%**	a,b	25	19.2%	
(c) Cigarettes that you only took a puff or two from	**c**	**70**	**53.8%**	**a,b,c**	**69**	**53.1%**	*
				a,c	1	0.8%	
	Total	130		Total:	130	100.0%	
Did you interpret “cigarettes” to include (*Circle all that apply*):							
(a) Manufactured cigarettes	**a**	**128**	**98.5%**	a	57	43.8%	*
(b) Hand-rolled cigarettes	**b**	**68**	**52.3%**	**a,b**	**26**	**20.0%**	

(c) Marijuana cigarettes	c	22	16.9%	a,b,c	1	0.8%
(d) Cigars	d	31	23.8%	a,b,c,d	3	2.3%
(e) Cloves (or other non-tobacco) cigarettes	e	35	26.9%	a,b,c,d,e	12	9.2%
(f) Something else. Specify: _____	f	5	3.8%	a,b,c,d,e,f	1	0.8%
				a,b,c,d,f	1	0.8%
				a,b,c,e	1	0.8%
				a,b,d	3	2.3%
				a,b,d,e	8	6.2%
				a,b,d,e,f	1	0.8%
				a,b,e	10	7.7%
				a,c,d	2	1.5%
				a,c,f	1	0.8%
				a,e	1	0.8%
				b	1	0.8%
				f	1	0.8%
				Total:	130	100.0%

S62a. Have you EVER used pipes, cigars, chewing tobacco or

snuff?

When you were answering this question, did you consider "ever used" to mean (*Pick one*):

(a) Trying even once (one puff, chew, or sniff)	**a**	**100**	**76.9%**			
(b) Trying several times	b	18	13.8%			
(c) Using regularly	c	9	6.9%			
(d) Something else. Specify: _____	d	3	2.3%			
	Total	130	100.0%			

Did you consider pipe smoking to include (*Pick all that apply*):

(a) Pipe tobacco	**a**	**126**	**97.7%**	**a**	**105**	**81.4%**
(b) Hashish	b	14	10.9%	a,b,c,d	8	6.2%
(c) Crack	c	15	11.6%	a,b,c,d,e	3	2.3%
(d) Marijuana	d	22	17.1%	a,b,d	2	1.6%
(e) Something else. Specify: _____	e	5	3.9%	a,b,d,e	1	0.8%
				a,c,d	3	2.3%
				a,d	3	2.3%
				a,e	1	0.8%
				c	1	0.8%
				d	2	1.6%

				Total	129	100.0%
Did you consider "using chewing tobacco" to mean (*Pick all that apply*):						
(a) Placing it in the mouth	**a**	**126**	**97.7%**	**a**	**122**	**94.6%**
(b) Something else. Specify:						
_____	b	7	5.4%	a,b	4	3.1%
				b	3	2.3%
				Total	129	100.0%
Did you consider "using snuff" to mean (*Pick one*):						
(a) Only sniffing it	**a**	**47**	**36.4%**	**a**	**46**	**35.7%**
(b) Only placing it in the mouth	b	24	18.6%	a,b,c	1	0.8%
(c) Either sniffing it or placing it in the mouth	c	43	33.3%	b	23	17.8%
(d) Something else? Specify:						
_____	d	13	10.1%	c	42	32.6%
	don't know	4	3.1%	d	13	10.1%
				don't know	4	3.1%
				Total	129	100.0%

Table 4. Examples of question-answer sequences (with answers underlined) in the corpus classifiable as fitting particular alignment paths, based on evidence available in this study. More than one possible path is listed when pre-interview conceptual alignment—not available in this study or perhaps even in principle—would distinguish the paths.

Q-A sequence	Survey response reliable? (Unchanged response on reinterview questionnaire)	Respondent received definition in interview?	Post-interview conceptual alignment	Possible path(s)
I: Which of these best describes your place of work's smoking policy for indoor indoor public or common areas, such as lobbies, restrooms, and lunchrooms. One, not allowed in any public areas, Two, allowed in some public areas, Allowed in all public areas. R: Uh one, it's <u>not allowed anywhere</u> actually. (Resp 71, Q69)	Yes	No	Aligned (100%)	**Prototypical Q-A sequence (no clarification)**
I: Have you EVER used pipes, cigars, chewing tobacco or snuff. And by ever used we mean took at least one puff, chew, or sniff? And that doesn't include um smoking hashish, marijuana, that's just a tobacco pipe.	Yes	Yes	Aligned (100%)	**Prototypical successful clarification** OR Confirmatory (or superfluous) definition

R: You mean, one puff ever? I: Ever. R: Yes. (Resp 1202, Q62)				
I: How old were you when you first started smoking cigarettes fairly regularly? R: Uh: sixteen. (Resp 84, Q33)	Yes	No	Misaligned (67%)	Uncorrected misalignment irrelevant to response
I: In HOSPITALS do you think that smoking should be allowed in all areas, allowed in some areas, or not allowed at all. R: Oh. Definitely: not allowed at all. I: okay and again I'm just gonna um kinda clarify the definition. Um we want to make sure you consider all public areas like the waiting rooms the cafeterias, and patient rooms. As well. R: Yeah, no. I: Okay (Resp 1206, Q72.2)	Yes	Yes	Misaligned (20%)	Irrelevant definition OR Unsuccessful but unnecessary clarification attempt
I: Have you smoked at least a hundred cigarettes in your entire life? R: Um: in my entire life, no, not really. Ha ha. I: Okay. I: And we want you to include any puffs on ANY cigarettes whether or not	No (Re-interview response with definition: "No")	Yes	Misaligned (67%)	Unsuccessful needed clarification attempt OR Harmful definition

you inhaled AND whether or not you finished them. R: Okay. I: And um so yo- have you smoked at least 100 cigarettes in your entire life? R: Um: I would say yeah! Ha ha. If you include that, yeah. I: Yeah? Okay. All righty. (Resp 1264, Q32)					
I: Do you think that ADVERTISING of tobacco products should be always allowed? Allowed under some conditions, or not allowed at all. R: Allowed under: some conditions. (Resp 264, Q77)	No (Re-interview response with definition: "Not allowed at all")	No	Misaligned (0%)	Uncorrected misalignment relevant to response	

Table 5. Distribution of 1,922 interview sequences for all respondents whose reinterview questionnaire included definitions, so that response reliability plausibly measures response accuracy during survey interview. This table omits 27 sequences (1.3% of the total) for which interview responses were unreliable despite perfect conceptual alignment with the definitions post-interview; whether these cases reflected sudden recall of circumstances overlooked during the interview or some other cause cannot be determined from these data.

Survey response reliable? (Unchanged	Respondent received definition in interview?	Post-interview conceptual alignment	Possible path(s)	Percent of sequences (number of cases)

response)

Yes	No	Aligned	**Prototypical Q-A sequence (no clarification)**	**15.1% (290)**
Yes	Yes	Aligned	**Prototypical successful clarification** OR Confirmatory (or superfluous) definition	**2.2% (42)**
Yes	No	Misaligned	Uncorrected misalignment irrelevant to response	59.6% (1147)
Yes	Yes	Misaligned	Irrelevant definition OR Unsuccessful but unnecessary clarification attempt	10.8% (208)
			Total reliable cases:	1687 (87.8%)
No	Yes	Misaligned	Unsuccessful needed clarification attempt OR Harmful definition	2.6% (50)
No	No	Misaligned	Uncorrected misalignment relevant to response	9.6% (185)
			Total unreliable cases:	12.2% (235)

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

When you answered this question, did you interpret "smoking" to include: *(Pick one)*

( ) Only puffs that you inhaled
( ) Any puffs, whether or not you inhaled

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

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

Figure 1. Conceptualization questions about "Have you smoked at least 100 cigarettes in your entire life?".

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

Definition:

- We want you to include any puffs on any cigarettes, whether or not you inhaled AND whether or not you finished them.
- We want you to include hand-rolled cigarettes as well as manufactured ones, and tobacco cigarettes with additives like cloves.
- We DON'T want you to include cigars or non-tobacco cigarettes, like marijuana cigarettes.

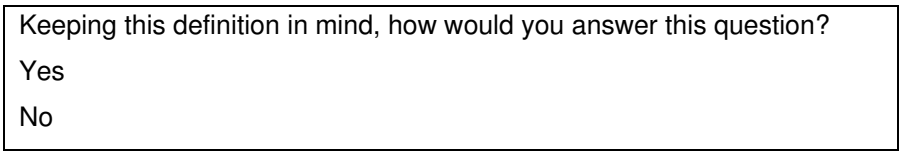

Figure 2. Re-administration of question "Have you smoked at least 100 cigarettes in your entire life?" with definitions for "smoked" and "cigarettes."