CLARIFYING SURVEY QUESTIONS WHEN RESPONDENTS DON'T KNOW THEY NEED CLARIFICATION

Frederick G. Conrad
Bureau of Labor Statistics
Michael F. Schober
New School University

Abstract

Enabling respondents to obtain clarification about the intended meaning of survey questions can dramatically improve response accuracy. However, respondents often do not recognize that their conceptions of terms in a question can differ from the question author's conception. As a result they do not request clarification on many occasions when it could improve their understanding and, thus, the accuracy of their responses. Similarly, survey interviewers often do not recognize that conceptions are misaligned and so do not volunteer clarification when it may be most helpful. We have been exploring two approaches to the design of web-based questionnaires that could increase the number of occasions when helpful clarification is provided. One approach is to increase respondents' sensitivity to the possibility of conceptual misalignment by rewording questions so that they include part of the definition of the key concept(s). We found that under some circumstances this led to increased requests for the full definition. Another approach is to build respondent models into the questionnaires so that the survey system can determine when respondents are confused and volunteer clarification. Such models can vary in their specificity from generic respondents to groups of respondents (stereotypic models) to models of individual respondents. We discuss ongoing work on respondent modeling.

Keywords: Survey Measurement Error, Data Quality, Question Comprehension, User Interfaces, Respondent Modeling

Introduction

For survey data to be accurate, respondents must understand questions as intended by their authors. Over the last several years we have explored techniques that give respondents the opportunity to clarify the meaning of questions both in interviews and in self-administered data collection. We have shown that clarification can dramatically improve response accuracy under certain situations (e.g. Conrad & Schober, 2000; Schober & Conrad, 1997). However, in virtually every study we have conducted, respondents given the option to request clarification do not obtain it on many occasions when it would help them respond more accurately. It seems they do not recognize that their interpretations of the ordinary words used in survey questions may differ from the senses intended by the sponsoring agencies. We have also found that interviewers often fail to volunteer clarification when it might help, and at other times volunteer it when it is not necessary. It seems that interviewers are not as skilled as one would hope at distinguishing the occasions on which respondents misunderstand from the occasions on which their understanding is aligned with the authors' intentions. In this paper, we will present evidence that respondents' and interviewers' understanding are often misaligned without either participant's recognizing it, and we discuss two ongoing efforts to address the problem.

Unrecognized conceptual misalignment

In one set of studies (Schober & Conrad, 1997; Schober, Conrad & Fricker, under review), interviewers telephoned respondents in the laboratory at the Bureau of Labor Statistics and asked behavioral questions from several ongoing Federal surveys. Respondents answered on the basis of fictional scenarios for which the correct answers were known. Half of the scenarios were designed so that questions were hard to interpret without clarification (complicated mappings between questions and scenarios) and half were designed so that interpretation was easy (straightforward mappings). For example, when asked "Has Kelly purchased or had expenses for household furniture?" a scenario depicting a receipt for the purchase of a floor lamp makes it hard to interpret the question without knowing whether a floor lamp should be considered furniture, while a receipt for the purchase of an end table makes it easy to interpret the question without clarification. In this study, respondents rarely asked for clarification when it was not available (1.6% of the questions) and while they asked more often when it was available (32.9% of the time), this was not as often as they needed it: for complicated mappings, their accuracy was not perfect (61.7% correct), yet they only requested clarification 47.0% of the time. Respondents were apparently certain of their answers on many occasions when they should not have been. It is surprising they did not take greater advantage of the opportunity to obtain clarification considering that they received instructions from the experimenters and the interviewers about the possibility that questions might use ordinary words in non-standard ways.

In a telephone interview of a national household sample (Conrad & Schober, 2000), respondents requested clarification only 4% of the time. This could suggest they underappreciated the potential value of clarification, but we argue that the low rate of help requests seems more likely to result from interviewers' poor discrimination between situations where clarification was needed and those where it was not: on 56% of the occasions that interviewers provided clarification, respondents did not show any evidence – that we could discern – of uncertainty or misconception. In a laboratory study in which respondents were interviewed by phone about their smoking behavior and opinions (Suessbrick, Schober & Conrad, 2000), they virtually never asked for clarification of everyday terms like "smoking" and "cigarettes." Yet a post-interview questionnaire indicated that they misunderstood the key survey concepts 49% of the time, on average. Unlike in the field study just described, these interviewers almost never offered clarification. We propose that it simply didn't occur to respondents or interviewers that their interpretations did not match, and therefore they did not recognize the need for clarification.

In a laboratory study (Conrad & Schober, 1999), respondents interacted with a computer assisted self-administered interviewing (CASI) tool that enabled them to obtain definitions by clicking on highlighted text. They answered the same questions used in the first laboratory study using the same scenarios. One group of participants was told that they were free to obtain clarification but that it was not required -- it was up to them. Another group was told that they might not understand the question and would be likely to respond inaccurately if they did not obtain clarification. The first group clicked for definitions relatively infrequently, about 23% of the time on average; the second group obtained the definitions much more often, about 81% of the time. Even though explicit instructions were effective in increasing help requests, this did not seem to make respondents more sensitive to misalignment of interpretations. Both groups obtained

definitions at about the same rate regardless of whether the question and scenario corresponded in a straightforward or complicated way.

Overcoming Unrecognized Misalignment

How can we get necessary information to respondents if (1) they don't realize when they need it and (2) the "keepers" of the information (interviewers or interviewing systems) don't recognize when respondents need it? We are currently exploring two possibilities in CASI applications: (1) give respondents more reason to suspect misalignment, and (2) automatically provide clarification when respondents' behavior indicates it could be helpful.

Give respondents reason to suspect misalignment.
We (Lind, Schober & Conrad, 2001) have begun exploring the first approach in a laboratory study by presenting parts of definitions along with the questions. The basic idea is that giving respondents a glimpse of the definition might make them aware that the concepts in the question are more complicated than they originally seemed to be and might inspire them to request clarification more than they otherwise would. To test this we presented 10 questions about housing and purchases from ongoing US government surveys (also used in Conrad & Schober, 2000) to three groups of respondents. Some respondents were presented the questions as they were originally worded and others were also presented parts of the critical definition. The questions were presented in a web browser and respondents could click on highlighted text to obtain the full definitions. All respondents answered on the basis of fictional scenarios so we could determine the accuracy of their responses.

One group of respondents was presented with the original questions, i.e. without any part of the definition, e.g. "How many people live in this house?". Another group was presented with questions that included a part of the definition that clarified the complicated mappings. For example, if the ambiguity concerned whether or not to count a child living away at college, the question might read "How many people live in this house? Do not count any people who would normally consider this their (legal) address but who are living away on business, in the armed forces, or attending school (such as boarding school or college)." The third group was presented with questions that included a part of the definition that was irrelevant to the ambiguity in complicated mappings, for example, "How many people live in this house? Live-in servants or other employees are included in the count." The basic issue is whether respondents request clarification (click on highlighted text) more often when presented with an irrelevant piece of the definition than when presented the question as originally worded.

In fact, irrelevant parts of the definition did lead to more requests for clarification than originally worded questions for complicated mappings: respondents clicked for the full definition 42.7% of the time when irrelevant information was included in the question, while they clicked 25.0% of the time when no part of the definition was included. Not surprisingly, relevant information did not increase respondents' rate of requesting clarification: they clicked for the definition on 21.4% of occasions, no different than the rate for the originally worded questions. Presumably the

potential confusion in complicated mappings is resolved when questions are supplemented with relevant information, making it unnecessary to obtain the full definition.

Including irrelevant information is potentially promising as a way to sensitize respondents to possible mismatches between their understanding of a concept and the question authors' intended interpretation. In addition, because it promotes voluntary requests for clarification, it may be more practical than instructing respondents to request clarification: in actual survey conditions, respondents may well ignore such instructions. However, in its current form the technique has several problems. First, it leads to an indiscriminate increase in requests for clarification. When presented with irrelevant information, respondents click almost as often for straightforward mappings (37.0% of the time) as they do for complicated mappings (42.7%).

Second, including irrelevant information in the question did not lead to a reliable increase in response accuracy for complicated mappings (47.4% correct with irrelevant information versus 42.2% correct with original question wording). In all our previous studies, whether with interviews or self-administered, computerized questionnaires, when respondents obtained clarification, their response accuracy was much better than when they did not obtain it. The difference in this case may be have been that respondents did not read the full definition (or enough of it to respond correctly). This might have occurred because the definitions themselves contain long passages that are irrelevant to the particular ambiguity involved, and having already encountered irrelevant information in the question, these respondents may not have been willing to read more.

This interpretation is bolstered by the response time data. Reading full definitions should increase response times, but respondents who were given irrelevant information answered at the same speed as respondents who were given relevant information and who had no reason to read the full definition (35.1 versus 33.7 seconds per question for complicated mappings). Both of these groups responded more slowly than the group that was presented original wording (25.9 seconds per question for complicated mappings), which presumably reflects the time required to read longer questions, i.e. questions that also include parts of definitions.

Assuming that these respondents did not read the definitions because they seemed largely irrelevant to their particular confusion, subsequent research ought to explore how to display definitions so that it is easier to extract their content. It may be that by using bullets, indentation and other formatting features, the topics and structure of the definitions can be made more transparent so that respondents can more quickly hone in on the information they need. In addition, more interactive and adaptive systems should be able to diagnose the details of the respondent's confusion and display just the relevant content of the definitions. Future research, thus, ought to explore the development of such systems.

Automatically provide clarification when respondents' behavior indicates it could be helpful. Another possible advantage of more interactive and adaptive survey systems is that they may be able to diagnose when respondents are uncertain about how to interpret questions and, thus, when clarification would be most helpful. This approach involves building into the CASI system a model of the respondent that could trigger the presentation of definitions. One potential piece

of evidence of user uncertainty that can be modeled in conventional desktop interfaces (graphical user interfaces with mouse and keyboard input) is periods of inactivity (no user action) that exceed some threshold.

The way the inactivity threshold is established depends on the type of respondent model. A generic model involves a single threshold for all respondents based on average response time for complicated mappings when no clarification is requested. A more specialized type of model would set different thresholds for groups of users who are known to differ in their overall response times – a so-called stereotypic respondent model. Models can also be developed for individual respondents. The specific threshold is set on the basis of response time to a small number of questions that are known to be good predictors of individual response times. Regardless of the type of model, inactivity thresholds must be adjusted upward for questions that typically take a long time to respond to (e.g. relatively long questions or particularly difficult ones) and downward for questions that typically produce quick responses (e.g. relatively short questions or particularly easy ones). A final consideration in creating user models is to identify thresholds that, on the one hand, are not too brief – which would lead to clarification that may not be needed – and, on the other hand, not too long – which might allow users to respond without obtaining clarification they need.

In our earlier work (Conrad & Schober, 1999), we tried to improve response accuracy with a generic respondent model in some versions of a CASI system, although the generic user model was apparently not tuned as well as it could have been. When respondents were told that they might need clarification to understand as intended, they generally clicked for definitions before the inactivity threshold was exceeded; the result was that response accuracy was very good for complicated mappings (83.3%) but no better than when the system could not volunteer definitions (78.8%). When respondents were not instructed about possible conceptual misalignment they generally responded quickly – well before the inactivity timeout – without requesting definitions. The result was low response accuracy for complicated mappings whether the system could (55.0%) or could not (49.4%) volunteer definitions.

We (Coiner, Schober & Conrad, in preparation) are currently experimenting with stereotypic respondent models in an effort to increase the value of system initiated clarification. We focus on respondents' *age*, rather than on other potentially modelable stereotypes (expertise, education, culture/ethnicity, gender) because our diagnostic indicator—inactivity—is likely to be influenced by the well-known slowing of behavior that occurs with age (see, e.g. Salthouse, 1996). Equal numbers of young respondents (younger than age 40) and old respondents (older than age 65) will interact with one of several user interfaces to the CASI system. The critical interfaces for current purposes are those that allow respondents to request clarification and (1) do not volunteer any clarification, (2) volunteer clarification based on a generic respondent model, and (3) volunteer clarification based on a stereotypic respondent model, i.e. inactivity thresholds are longer for old than young respondents. If the generic model can be adequately tuned through pretesting, we expect response accuracy to be higher for complicated mappings than when the system does not offer any clarification. In addition, we expect the stereotypic respondent model to increase accuracy above the generic model by interpreting inactivity more precisely for the different groups of respondents.

We have designed a follow-up study to examine whether individualized user models improve response accuracy above and beyond any improvements from generic or stereotypic user models (as Rich, 1999, has argued they will; see also Kay, 1995). The idea is that a system might improve its diagnostic abilities by observing each individual respondent's baseline performance (assessed on speed to answer practice questions), rather than basing its diagnoses on the less fine-grained criterion of group membership. In the previous experiment, we are administering a fairly large pool of practice items. We will select a small number of these for the current study that were most effective in predicting response times for the actual survey items. Because we will have data from the previous study on how response latencies to practice items are statistically related to response latencies to the survey questions, the system can adjust its thresholds for presenting clarification, tailoring the thresholds for each question for each respondent.

Conclusions

Differences in people's conceptions of particular terms surely go unnoticed in everyday conversation, and yet people still manage to adequately understand each other. In surveys, however, non-standard and technical meanings may lead to greater misalignment than in everyday conversation, so that overlooking the misalignment may lead to serious misunderstanding and, as a result, inaccurate responses. The techniques we have been exploring look promising, but both have limits. Including irrelevant parts of definitions does seem to increase respondents' sensitivity to the possibility of conceptual mismatch, but it is limited by their motivation to address the problem and, most likely, by the usability of definitions.

Modeling respondent behavior also seems likely to help, but the exact form in which the model-based clarification is delivered may be extremely important. Unobtrusive, embedded user models are increasingly common in software; for example, agents that learn the user's common typographical errors and correct them as soon as they occur are standard in word processing programs. They seem to improve the user's performance without disrupting the primary task. In fact users often don't notice that the correction has been made. On the other hand, animated agents like the Microsoft Office Assistant appropriate the system's focus from the user's primary task and often misdiagnose the user's actions. Many users report finding this irritating and hard to ignore. Finding the right balance between background and foreground operation for respondent models will be an essential step in their acceptance by respondents.

References

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

Conrad, F.G. & Schober, M.F. (1999). A conversational approach to text-based computer-administered questionnaires. In *Proceedings of the Third International ASC conference*. Chichester, UK: Association for Survey Computing, pp. 91-101.

Coiner, T.F., Schober, M.F. & Conrad, F.G. (in preparation). Age-based models of respondents' need for clarification in a web survey.

Kay, J. (1995). Vive la difference! Individualized interactions with users. In C.S. Mellish (ed.), *Proceedings of the 14th International Conference on Artificial Intelligence*, pp. 978-984. San Mateo, CA: Morgan Kaufmann, Publishers.

Lind, L. H., Schober, M. F. & Conrad, F.G. (2001*). Clarifying question meaning in a web-based survey*. Paper presented at the 56th Annual Conference of the American Association for Public Opinion Research. Montreal, Canada.

Rich, E. (1999). Users are individuals: Individualizing user models. *International Journal of Human-Computer Studies*, 51, 323-338.

Salthouse, T.A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, 103, 403-428.

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

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

Suessbrick, A., Schober, M.F. & Conrad, F.G. (2000). Different respondents interpret ordinary questions quite differently. In *Proceedings of the American Statistical Association, Section on Survey Methods Research*. Alexandria, VA: American Statistical Association