CLARIFYING QUESTION MEANING IN A WEB-BASED SURVEY[1]

Laura H. Lind, New School for Social Research
Michael F. Schober, New School for Social Research
Frederick G. Conrad, Bureau of Labor Statistics
Michael F. Schober, Dept. of Psychology AL-330, New School University,
65 Fifth Ave., New York, NY 10003

Key Words: web surveys, question clarification, data quality, response accuracy, conversational interviewing

INTRODUCTION

In carrying out a web survey—or any survey—getting accurate responses to survey questions depends on respondents being able to interpret the questions as the survey designers intend. However, we know from our previous research that people can interpret concepts in ordinary web-based or interviewer-administered questions in a variety of ways. Take, for example, the following question from the Current Point of Purchase survey, "During the past year, have you purchased or had expenses for home maintenance and repair?" When answering this question, some people count only work that they paid others to do, and some include expenses for work they did themselves (Conrad & Schober, 2000). Another example is this question from the Tobacco Supplement to the Current Population Survey, "Have you smoked at least 100 cigarettes in your entire life?" Some people include only tobacco cigarettes, and others include cloves, marijuana, and cigars (Suessbrick, Schober, & Conrad, 2000). And yet another example of how ordinary questions can be interpreted differently is this question from the Consumer Price Index–Housing survey, "How many people live in your house?" When answering this question, some people include college students away at school, but others don't (Conrad & Schober, 2000; Schober & Conrad, 1997).

One strategy for clarifying question meaning and increasing uniformity of interpretation in web surveys is to provide respondents with definitions of key concepts. Definitions are especially helpful when respondents' circumstances do not map onto survey concepts in a straightforward way. In fact, our earlier research has shown that respondents interpret questions more uniformly when they get definitions (Conrad & Schober, 2000; Schober & Conrad, 1997; Schober, Conrad, & Bloom, 2000; Schober, Conrad, & Fricker, 1999; Suessbrick, Schober & Conrad, 2000). The trouble is that these definitions can be rather long and complex. For example, this is the definition from the Consumer Price Index–Housing survey for who should be counted as "living in a house":

- A person is considered to be living in a housing unit even if the person is not present at the time of the survey. Live-in servants or other employees, lodgers, and members of the household temporarily away from the unit on business or vacation are included in the count.
- 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).
- Do not count overnight lodgers, guests and visitors.
- Do not count day employees who live elsewhere.

Although we know that providing such definitions to respondents improves response accuracy (Conrad & Schober, 2000; Schober & Conrad, 1997, 1998), it is unclear what the best way is to present them to respondents. We see several options.

One option is to rely on respondents to ask for clarification (for example, clicking on highlighted text) when they think they need it. Although this is relatively easy to implement, our previous research shows that relying on respondents to ask for clarification doesn't work. Respondents often don't recognize when they need clarification (Bloom & Schober, 1999; Conrad & Schober, 2000; Schober & Conrad, 1997). And when we told respondents that, for some questions, they would need to ask for clarification in order to answer accurately, they asked for it every time, even when they didn't need it, unnecessarily increasing survey duration, and possibly reducing their likelihood of completing a longer survey (Schober, Conrad & Bloom, 2000).

[1] We thank Jim Kennedy at the Bureau of Labor Statistics, Jeff Lind, and members of the Psycholinguistics Laboratory at the New School University for their assistance. This material is based upon work supported by the National Science Foundation under grants No. SBNR-9730140 and IIS-0081550. The opinions expressed are those of the authors and not of the Bureau of Labor Statistics.

A second option is to decompose each question into a series of questions that addresses each component of the definition, and therefore each potential ambiguity the respondent may face. For example, one might turn our example question into the following series of questions:

- How many people live in your house?
- Did you count any people who are living away on business, in the armed forces, or attending school (such as boarding school or college)?
- Did you count any overnight lodgers, guests and visitors?
- Did you count day employees who live elsewhere?

This approach is impractical, because it drastically increases survey duration even for respondents who don't need clarification, and survey definitions can include too many components to list.

A third option is to present the definition along with the question. However, presenting the entire definition is often not feasible; many definitions are simply too long. If respondents bother to read them, doing this will simply increase survey duration for all respondents, and it may lead to lower survey completion rates.

A fourth option is to create web systems that mimic human "conversational interviewers" (see Conrad & Schober, 2000; Schober & Conrad, 1997), presenting only the parts of definitions that are relevant to a given respondent's circumstances, through judicious probing. Doing this in a web survey would require complex dialog systems and advanced software techniques such as those used in artificial intelligence, which are likely to be costly to develop and administer – if they can be developed at all. Although in other studies we are examining the feasibility and potential benefits of such systems, we are also interested in examining lower-tech ways of improving data quality. The approach we report here is to examine how, with existing technologies, we can improve uniformity of interpretation (and thus response accuracy) in a web-based survey.

Our strategy is to reword web-based survey questions such that they include parts of complex definitions, within a system where respondents can click on highlighted terms to get the full definitions. The hypothesis is that conveying the complexity of the concept to the respondent might motivate the respondent to request the full definition. When definitions are long and complex, this is a simple alternative to presenting the entire definitions along with the questions, and can be easily implemented in a web browser. Obviously, this should improve response accuracy when the part of definition that is included is directly relevant to respondents' circumstances. But what will respondents do when they are presented with irrelevant parts of definitions? Will it lead them to think harder about concepts? Will it inspire respondents to ask for clarification? Or will they simply ignore the additional information and answer inaccurately?

EXPERIMENTAL DESIGN

To test the effect of presenting parts of definitions along with questions, we constructed the following experimental design. Respondents were randomly assigned to one of three experimental groups. In the first group, respondents answered questions as they were originally worded – that is, without any part of the definition. In the second group, respondents answered questions that were reworded to include a part of the definition that was directly relevant to their circumstances, i.e., that directly resolved an ambiguity in their circumstances (which we controlled because they answered on the basis of fictional scenarios). In the third group, respondents answered questions that were reworded to include a part of the definition that was irrelevant to their circumstances, i.e., that did not resolve an ambiguity in their circumstances. The way in which we determined what information would be relevant or irrelevant to a particular ambiguity for a particular respondent will be explained shortly.

Questions. All respondents answered the same ten questions from two ongoing government surveys (from Conrad & Schober, 2000). There were five questions about purchases from the Current Point of Purchase Survey, and five questions about housing from the Consumer Price Index–Housing survey. Half of the respondents answered the housing questions first, and the other half answered the purchasing questions first. The order of questions within each domain (purchases and housing) remained the same for all respondents.

Interface. The questions were presented to respondents on a computer screen using a web-browser interface. Respondents answered questions by clicking on radio buttons with a mouse, for questions that required a 'yes' or 'no' response, and typing with the keyboard, for questions that required a numerical response. All questions contained a hyperlink (highlighted and underlined text) for the key concept of that question. Respondents in all three groups could always click on this text to see the full definition for that concept, but they were not required to do so.

Scenarios. Respondents answered these questions on the basis of fictional scenarios. This allowed us to measure response accuracy and know what part of a

definition would be relevant or irrelevant for a particular respondent. These scenarios were presented in a paper packet, and consisted of floor plans, pictures of receipts from purchases, and short stories. The respondents received a total of ten scenarios upon which to base their answers, one per question.

Respondents answered half of the survey questions on the basis of straightforward scenarios and half on the basis of complicated scenarios. Straightforward scenarios were designed to map onto the survey definitions in a typical way. In other words, they were designed to be relatively easy to answer correctly even without knowing the official definition. In contrast, complicated scenarios were designed to map onto the survey definitions in an atypical way, making them difficult to answer accurately without knowing the official survey definitions.

The following is an example of a complicated scenario that a respondent might see in their paper packet and which they would use to answer "How many people live in your house?":

> The Gutierrez family owns the 3-bedroom house at 4694 Marwood Drive. The family has four members: Maria and Pablo Gutierrez, and their two children Linda and Marta. There is one bedroom for Maria and Pablo, one for Marta, and one for Linda. Linda is a college student. Although her legal address is still 4694 Marwood Drive, she stays at the college dorms all year, except for holidays and vacations.

Figures 1 through 3 are examples of how the computer screen would look to respondents in the three experimental groups when answering the question on the basis of this scenario. A respondent in the first group would answer as it was originally worded, i.e., without any part of the definition (see Figure 1). A respondent in the second group would answer a question that has been reworded to include the part of the definition that is directly relevant to the ambiguity presented in the scenario (see Figure 2). And a respondent in the third group would answer a question that has been reworded to include a part of the definition that is not relevant to the ambiguity presented in the scenario (see Figure 3). Notice that all three question versions contain a hyperlink (in this case, the term "live" is underlined and highlighted). This indicates to respondents that they may click on this portion of text to see the full definition for that term. If they choose to do this, the full definition appears on the right hand side of the computer screen (see Figure 4).

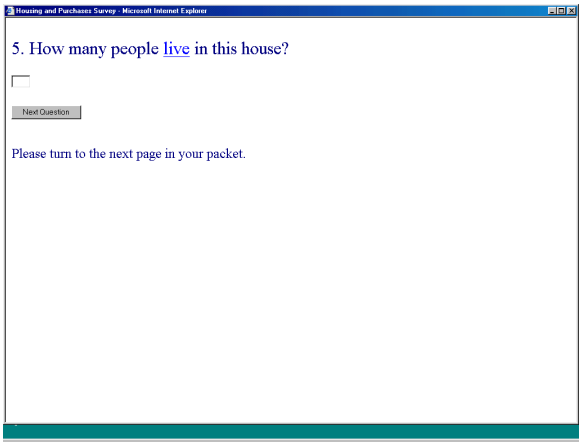

Figure 1. Survey question as originally worded

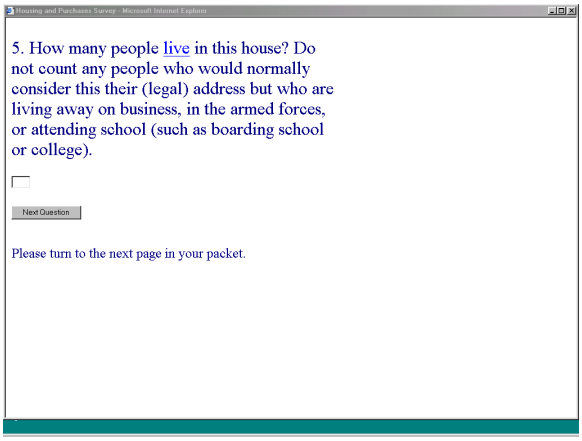

Figure 2. Survey question with relevant part of definition

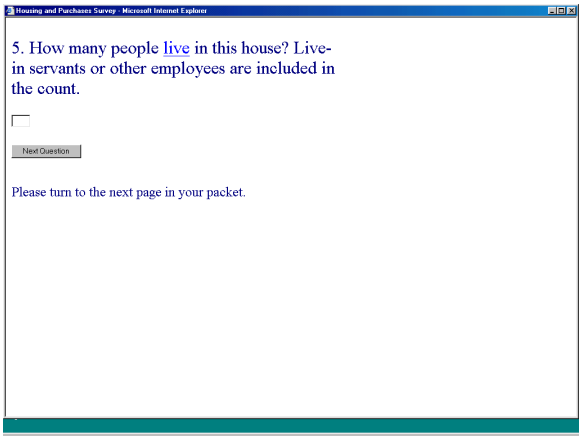

Figure 3. Survey question with irrelevant part of definition

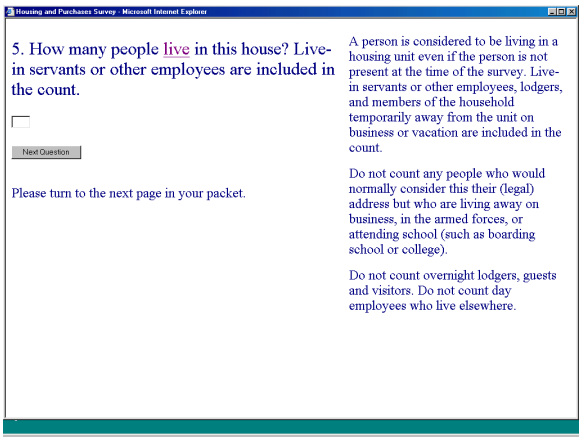

Figure 4. Full definition appears when hyperlink is clicked

Participants. The participants were 48 paid respondents recruited from the New York City area and the New School University community by means of an ad in the *Village Voice*. There were 27 women and 21 men. The mean age of the participants was 29.8 years old. Ethnicities, educational backgrounds, and experience with computers were balanced across the three experimental groups.

RESULTS

We examined how the different question wordings affected 1) rates of clarification seeking, 2) response accuracy, and 3) interview duration.

Requests for full definition. In all three conditions, people clicked for definitions more often with complicated mappings (for 29.7% of the questions) than for straightforward mappings (for 23.3% of the questions), showing that there was an effect of mapping, $F(1,45) = 4.34$, $p < .05$. As Figure 5 shows, when answering on the basis of complicated scenarios, respondents who received originally worded questions requested the full definition for 25.0% of questions. Respondents who received reworded questions that contained a part of the definition that was relevant to their scenarios did not click for definitions any more often than respondents who received originally worded questions (21.4% of the time). Presumably, this is because the ambiguity that is present in the scenario is resolved when the relevant information is included, making it unnecessary to obtain the full definition. However, respondents who received a part of the definition not relevant to their scenarios clicked about twice as often as they did in the other groups (42.7% of the time), as if they recognized that they needed clarification, $F(1,45) = 4.35$, $p < .05$.

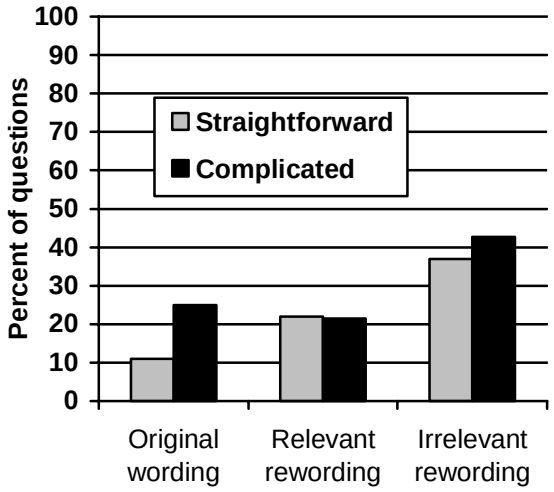

Figure 5. Requests (clicks) for a full definition

Response accuracy. As can be seen in Figure 6, when answering questions based on straightforward scenarios, accuracy was uniformly good for all three respondent groups (93.7% of questions answered correctly). When answering questions based on complicated scenarios, accuracy was poor for respondents who received originally worded questions (42.2% correct). Presumably this is because they did not realize that their understanding of the question concepts differed from the survey designers' definitions. Accuracy was much better for respondents who received questions that included a relevant part of the definition (84.9% correct). But when respondents answered questions that included only an irrelevant part of the definition, their accuracy was no better than for respondents who answered originally worded questions (47.4% correct).

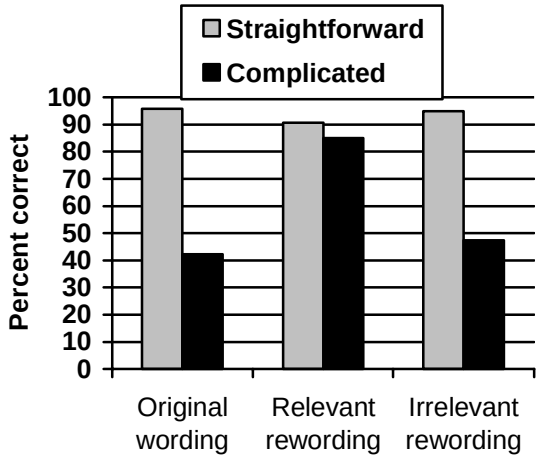

Figure 6. Response Accuracy

So although these respondents requested clarification more often, receiving that clarification did not help them answer more accurately. Perhaps when respondents got definitions they may not have read them or may not have understood them sufficiently to improve their accuracy.

Response time per question. In Figure 7 we can see that response times were longer for reworded questions (whether they included relevant information, 32.3 seconds, or irrelevant information, 33.5 seconds) than for originally worded questions (23.4 seconds), Helmert contrast $F(1,45) = 11.04$, $p = .007$. So it seems that including parts of definitions along with questions increases interview duration, but indiscriminately, and without necessarily improving response accuracy.

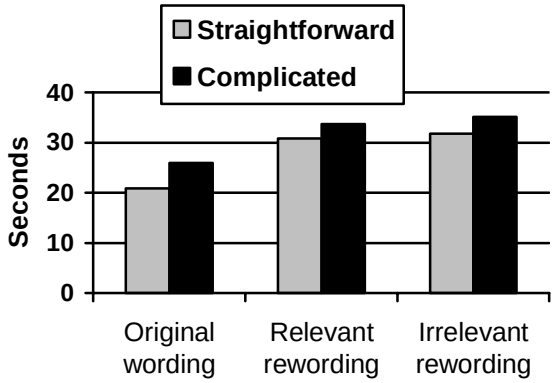

Figure 7. Response Time

DISCUSSION

The results of our study indicate that including parts of definitions in questions has both benefits and drawbacks. On the one hand, this strategy seems like a good idea. If you happen to include a part of the definition that directly addresses the ambiguity in the respondent's situation, then accuracy will be improved. And even if the part of the definition that you choose to include is irrelevant to the respondent's circumstances, you may succeed in prompting the respondent to ask for clarification. On the other hand, this technique could be seen as a bad idea. Including parts of definitions in questions does increase survey duration fairly substantially, and for a survey with many items, this could prove to be especially problematic. In addition, while irrelevant information seems to inspire respondents to ask for clarification, the results of our study indicate that this did not lead to greater accuracy.

In our previous studies, we have found that respondents answered survey questions more accurately when they obtained definitions than when they did not. In the current study, this was not the case. A possible reason for this may be the definitions themselves. Many of the definitions contained lengthy passages, and information that was irrelevant to a respondent's situation. For respondents who had already encountered irrelevant information in the question, they may have been discouraged when, at first glance, the definitions seemed to consist of even more irrelevant information. Therefore, they may have been unwilling to read the definitions thoroughly enough to extract the information that they needed to answer the questions accurately. If this is the case, then subsequent research should examine how to display definitions so that it will be easier for respondents to zero in on the information they need.

More broadly, the current results suggest the importance of investigating more intelligent and adaptive survey interfaces that don't rely on respondents (or interviewers, for that matter) to determine when they need clarification. Based on respondent textual or vocal cues (Schober, Conrad, & Bloom, 2000), survey systems of the future might be able to diagnose when clarification is relevant for a particular respondent even when the respondent hasn't requested clarification.

The usual approach to creating survey systems is to more or less directly transfer a survey from another mode to a web page (e.g., Dillman, 2000) or to a voice menu system (e.g., Nicholls, Baker & Martin, 1997). Thus far, the attention in designing the appropriate interfaces has gone into determining the appropriate layout—for example, how to set up navigation and data entry buttons so that they are intelligible (Dillman, 2000), and how to structure the flow of the questionnaire on desktops (see papers in Couper et al., 1999) or in speech systems (Blyth, 1997). This is an important part of creating a well-designed user-centered system.

In our view, the fact that user interfaces to current survey systems do not support clarification dialog with the user means that they do not take advantage of the full potential of interactive interfaces. If one can generalize from proposals in other domains of human-computer interfaces (database query systems, advice-giving systems, help systems, e.g., Brennan & Hulteen, 1995; Cahn & Brennan, 1999; Cawsey, 1992, 1993; Kobsa & Wahlster, 1989; Moore 1995; Moore & Paris, 1992; Paek & Horvitz, 1999; Traum, 1994, among others) to survey settings, it may well be that task performance and user satisfaction will improve when users engage in dialogs (linguistic or graphical) to correct misconceptions on either end.

REFERENCES

Bloom, J.E., & Schober, M.F. (1999). Respondent cues that survey questions are in danger of being misunderstood. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 992-997. Alexandria, VA: ASA.

Blyth, B. (1997). Developing a speech recognition application for survey research. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (eds.), *Survey measurement and process quality* (pp. 249-266). New York: Wiley.

Brennan, S.E., & Hulteen, E. (1995). Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems, 8*, 143-151.

Cahn, J.E., & Brennan, S.E. (1999). A psychological model of grounding and repair in dialog. In Proceedings of the American Association for Artificial Intelligence Fall Symposium *Psychological models of communication in collaborative systems* (pp. 25-33). Menlo Park, CA: AAAI Press.

Cawsey, A. (1992). *Explanation and interaction: The computer generation of explanatory dialogues.* Cambridge, MA: MIT Press.

Cawsey, A. (1993). User modelling in interactive explanations. *User Modeling and User-Adapted Interaction, 3*, 221-247.

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

Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nicholls II, W.L, & O'Reilly, J.M. (eds). (1998). *Computer assisted survey information collection*. New York: Wiley.

Dillman, D.A. (2000). *Mail and Internet surveys: The tailored design method (2nd ed.)*. New York: Wiley.

Kobsa, A., & Wahlster, W. (eds.) (1989). *User models in dialog systems.* New York: Springer-Verlag.

Moore, J.D. (1995). *Participating in explanatory dialogues: Interpreting and responding to questions in context*. Cambridge, MA: MIT Press.

Moore, J.D., & Paris, C.L. (1992). Exploiting user feedback to compensate for the unreliability of user models. *User Modeling and User-Adapted Interaction, 2*, 287-330.

Nicholls II, W.L., Baker, R.P., & Martin, J. (1997). The effect of new data collection technologies on survey data quality. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (eds.), *Survey measurement and process quality* (pp. 221-248). New York: John Wiley & Sons.

Paek, T., & Horvitz, E. (1999). Uncertainty, utility, and misunderstanding: A decision-theoretic perspective on grounding in conversational systems. In Proceedings of the American Association for Artificial Intelligence Fall Symposium *Psychological models of communication in collaborative systems* (pp. 85-92). Menlo Park, CA: AAAI Press.

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

Schober, M.F., Conrad, F.G., & Bloom, J.E. (2000). Clarifying word meanings in computer-administered survey interviews. In L.R. Gleitman & A.K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 447-452). Mahwah, NJ: Lawrence Erlbaum Associates.

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: ASA.

Suessbrick, A., Schober, M.F., and 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: ASA.

Traum, D.R. (1994). *A computational theory of grounding in natural language conversation.* Unpublished Ph.D. dissertation, University of Rochester.