Initiative and Clarification in Web-Based Surveys

Michael F. Schober
Department of Psychology
New School for Social Research
65 Fifth Avenue
New York, NY 10003
+1 212 229 5787
schober@newschool.edu

Frederick G. Conrad
Institute for Social Research
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106
+1 734 936 1019
fconrad@isr.umich.edu

Patrick Ehlen
Laura H. Lind
Tania Coiner
Department of Psychology
New School for Social Research
ehlenp@newschool.edu
laurza@yahoo.com
coint445@newschool.edu

ABSTRACT

Respondents to computer-administered surveys don't always interpret ordinary words and phrases uniformly; clarification dialog can help improve uniformity of interpretation, and thus survey data quality. Here we explore two approaches to the interfaces for 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, suggesting that respondents recognize the potential for conceptual misalignment. Another approach is to build respondent models so that a mixed-initiative 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. Our data show both potential benefits and pitfalls to this approach, and they begin to map out the territory for dialog in systems that collect data from users, as opposed to systems that provide data to users. Our data also begin to help determine the circumstances under which more sophisticated survey dialog systems with Natural Language generation and comprehension abilities are needed.

Keywords
Survey Data Collection Interfaces, User Models, Initiative, Question Comprehension, Clarification, Data Quality

INTRODUCTION

We have been exploring how best to design interfaces for computer systems that collect survey data from users. Such systems belong to a growing class of applications that collect information from users: on-line job applications, forms for e-commerce, on-line tax forms, electronic standardized tests, electronic voting systems, etc. They contrast with the more frequently-studied systems from which users extract information, like web search engines and other database query applications, and they may have quite different properties.

In current practice, computer-administered surveys present questions or probes to all users in the same format, leaving the interpretation of those questions up to the user. This is done in the name of standardizing the materials—making sure all respondents are exposed to the same stimuli. But this practice is problematic because different users can interpret the same questions quite differently (see, e.g., [11], [40]), which can lead to poor response accuracy under certain circumstances (see [16], [37], [38]). Even if a survey system offers clarification for users, many of them may not recognize that they need it, or they may be unwilling to request it ([2], [13], [14]).

Here we describe studies that examine which sorts of interfaces to survey systems promote more uniform interpretations of questions—and thus higher quality data. We contrast interfaces that require users to educate themselves about how the questions should be taken with interfaces that diagnose users' need for clarification; we also contrast systems that provide clarification only when users request it (user-initiated clarification) with systems that also provide clarification even without being asked (mixed-initiative clarification). These studies focus on desktop (windows, keyboard entry, and mouse clicking) survey interviewing systems that government agencies could plausibly implement in the future, to gather the data for influential statistics like the unemployment rate or the Consumer Price Index. These surveys ask questions about verifiable facts and behaviors, rather than about opinions and attitudes. As such, they provide an excellent research opportunity: if the accuracy of users' responses is verifiable, we can objectively evaluate how different dialogue system designs affect the uniformity of users' comprehension and the ultimate data quality. Because many government agencies have created official definitions for concepts in their survey questions—for example, defining what kinds of activities count as "work for pay"—we can determine which kinds of dialogue systems promote question interpretations that match the official definitions.

The purpose of the project is not to implement actual working systems, but rather, using the tools of experimental psychology, to determine what kinds of dialogue systems will be most effective, as determined by measuring response accuracy, duration of interactions, and user satisfaction. The findings will thus provide guidelines for future development of interfaces for collecting information from users, both for official statistics and for commercial purposes. More generally, the findings will contribute to the understanding of how design for systems that collect information should differ from design for systems that provide information to users.

The potential practical value of improving survey interfaces is enormous. The kinds of large-scale government surveys we are examining (e.g., the Current Population Survey, which is currently administered to 40,000 respondents per month, or the Decennial Census, which is supposed to be administered to all U.S. citizens) are currently carried out by large retinues of paid interviewers. Training and monitoring interviewers on this scale is costly and difficult, and there is no guarantee that interviewers follow procedures consistently and uniformly; in fact, some evidence suggests that interviewer practices vary enormously, despite uniform training (see [19], [39]). Another practical concern in gathering important official statistics is maintaining high response rates and appropriate sampling procedures as surveys migrate to on-line administration. This will almost certainly be affected by how bearable it is to respond to them.

SURVEY SYSTEMS

The usual approach to creating an information-collection system is to more or less directly transfer a paper-and-pencil questionnaire (job application, ballot, order form) or telephone interview script to a web page (e.g., [18]) or to a voice menu system (e.g., [30]). 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 [18], and how to structure the flow of the questionnaire on desktops (see papers in [17]) or in speech systems [5]. This is an important part of creating a well-designed user-centered system.

But user interfaces to current survey systems do not support clarification dialog with the user, and thus they do not take advantage of the full potential of interactive interfaces. In fact, by embodying the standardized interviewing techniques that human interviewers are often trained to use—leaving the interpretation of questions entirely up to respondents—current interfaces may promote poor data quality, because respondents' interpretations of the meanings of questions can vary so much. Our preliminary studies on survey interfaces ([2], [4], [13], [14]) support the notion that more collaborative interfaces—interfaces that allow users to engage in dialog with systems to learn what questions mean—can improve data quality without decreasing user satisfaction. This is consistent with proposals in other domains of human-computer interfaces (e.g., database query systems, advice-giving systems, help systems) that task performance and user satisfaction will improve when users engage in dialogs (linguistic or graphical) to correct misconceptions on either end (e.g., [6], [8], [9], [10], [25], [28], [29], [31], [44], among others).

It is not a foregone conclusion that every increase in a survey interface's sophistication brings a corresponding increase in data quality (for similar arguments about data retrieval applications, see [33], p. 331, among others). It may be that quite simple innovations provide the greatest payoff, and that more costly and difficult innovations (for example, using sophisticated Natural Language Generation techniques) may not bring sufficient improvement to be worth the effort and expense. It is just these tradeoffs that we examine here.

This is important to do because survey systems differ in important ways from information retrieval systems [40]. Beyond the obvious difference that in surveys users provide information to systems rather than retrieving it, another important feature is that users are often less motivated to participate in surveys than to participate in the kinds of tasks that they typically engage in with computers. Users may continue interacting with a less-than-ideal data retrieval system if the system ultimately promises to provide them with information they need, but it is far less likely that users will continue providing answers to unbearable survey systems. The user's aims in interacting with a survey system—to finish quickly, perhaps providing adequate answers—can run counter to the survey designers' aims, which are to collect the most accurate data possible. Thus more sophisticated survey systems that (for example) engage users in clarification subdialogs may increase the accuracy of the data only for the few users who can bear to complete the survey.

Our fundamental question is this: how intelligent should survey interfaces be? Should research effort be expended on implementing systems that can interpret users' natural-language input and reason about when and how to provide appropriate clarification? Are sophisticated user models necessary for improving data quality? Should effort be expended on designing multimedia survey interfaces that allow textual, graphical and spoken input and output? The programming effort involved would be formidable, and the reasoning rules might have to be developed separately for different content domains. (A single large-scale government survey can involve hundreds of question concepts, each quite complicated[1]). And the maintenance burden would be substantial, as surveys are revised regularly to keep up with societal changes. The point is that *if* sophisticated survey interfaces prove to be useful, it would need to be demonstrated that their added utility was sufficient to justify the costs of building and maintaining them. Our studies begin to provide a basis for calculating this equation.

Our studies are informed by longstanding debates in the social sciences about whether words contain meaning or whether meaning is more collaborative. We believe our studies can contribute to these debates as well, as the different interfaces we examine embody different assumptions about the nature of communication, just as different interviewing methods do (see [36]).

Our studies also provide further evidence (one way or another) for proponents of strict standardization [20] or for their critics in the survey world ([42]) and in other disciplines (e.g., [7], [23], [26]), who argue that rigid adherence to survey scripts jeopardizes validity ([42], p. 233) because there is no guarantee that all respondents will interpret questions the same way (see [1], for a history of the debate). A related debate in the HCI world is over the extent to which one-size-fits-all labels in interfaces are helpful (e.g., [32]) or harmful (e.g. [21], [22]). As part of this debate, our studies are informed by our parallel line of research on dialog in human-human survey interviews ([3], [15], [16] [38], [41], [43]).

Finally, our studies not only have obvious practical implications for the survey world, but they also expand the range of task domains that test human-computer interaction theories. Applications that collect data from users have different characteristics than applications that provide users with information, and users may need to interact with such systems in different ways. Ultimately, such studies may lead to a broader vision of the interplay between task and dialog in HCI.

STUDIES

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 human-human interviews and in survey data collection via computers. We have shown that clarification can dramatically improve response accuracy under certain situations (e.g. [16], [38]). 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. Here we 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 in survey interfaces.

Prior studies

In an earlier set of studies ([38], [41]), interviewers telephoned respondents in a laboratory 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 (2% of the questions) and while they asked more often when it was available (33% of the time), this was not as often as they needed it: for complicated mappings, their accuracy was not perfect (62% correct), yet they only requested clarification 47% 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.

[1] Consider, for example, the Current Population Survey’s definition of what counts as having a “job”: A job exists when there is a definite arrangement for regular work every week, or every month, for pay or other compensation (e.g., profits, anticipated profits, or pay in kind, such as room and board). A formal, definite arrangement with one or more employers to work on a continuing basis for a specified number of hours per week or days per month, but on an irregular schedule during the week or month, is also a job…It is possible for individuals to have more than one employer, but only one job. If an individual does the same type of work for more than one employer in an occupation where it is common to have more than one employer, do not consider the individual a multiple jobholder. Examples include private household or domestic workers including babysitters, chauffeurs, gardeners, handypersons, cooks, and maids. (US Department of Commerce, 1994, p. B1-6)

In telephone interviews of a national household sample [16], 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 [43], 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.

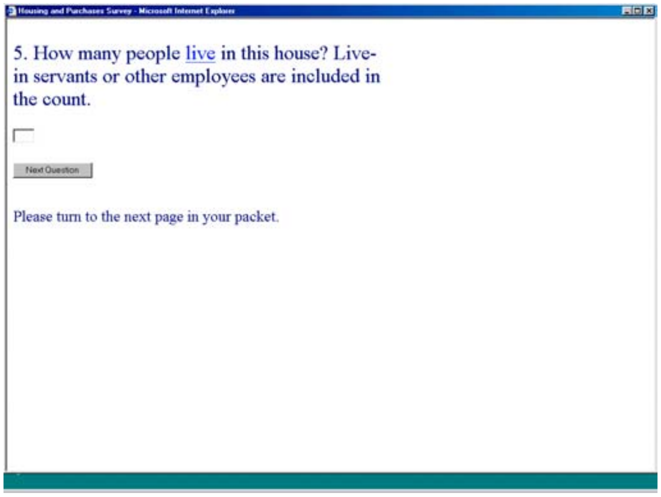
In a laboratory study [14], respondents interacted with a computer assisted self-administered interviewing (CASI) tool that enabled them to obtain definitions by clicking on highlighted text. They answered the same questions used in [38] using the same scenarios. One group of participants was told that 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 increased help requests, this did not seem to make respondents more sensitive to actual misalignment of interpretations. Both groups obtained definitions about as often 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 have been exploring two possibilities in browser-based survey interfaces: (1) give users more reason to suspect misalignment, and (2) automatically provide clarification when users' behavior indicates it could be helpful.

Give users reason to suspect misalignment

We [27] have begun exploring the first approach in a laboratory study by presenting parts of definitions along with the questions. The 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 [16]) to 3 groups of respondents. Some respondents were presented the questions as 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 (see Fig. 1) to obtain the full definitions. All respondents answered on the basis of fictional scenarios so we could determine the accuracy of their responses.



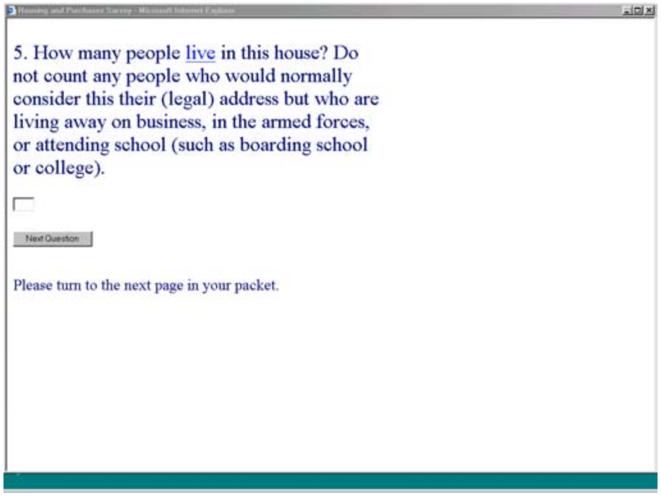
5. How many people live in this house? Live-in servants or other employees are included in the count.

□

Please turn to the next page in your packet.

Figure 1: Survey question containing clickable term ("live")

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 they had been presented with (see Fig. 2). For example, if the



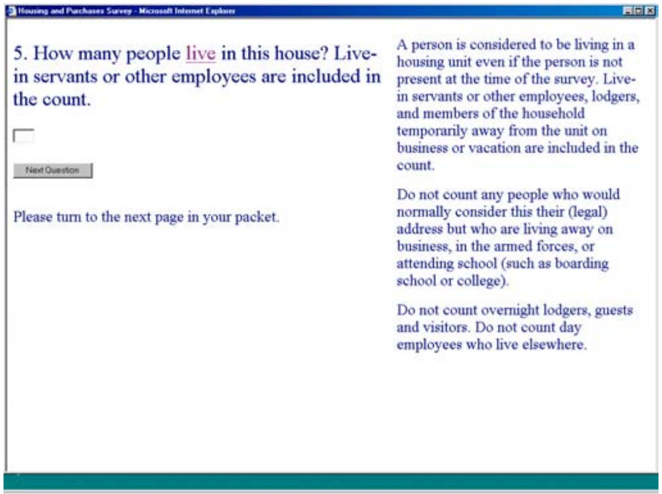
5. How many people live in this house? Do not count any people who would normally consider this their (legal) address but who are living away on business, in the armed forces, or attending school (such as boarding school or college).

□

Please turn to the next page in your packet.

Figure 2: Survey question with relevant part of definition

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 the complicated mappings they had been presented with (see Fig. 3, which includes the full definition that appears when "live" is clicked), for example, “How many people live in this house? Live-in servants or other employees are included in the count.”



5. How many people live in this house? Live-in servants or other employees are included in the count.

□

Please turn to the next page in your packet.

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.

Figure 3: Survey with irrelevant part of definition and the full definition produced by clicking

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 43% of the time when irrelevant information was included in the question, while they clicked 25% 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% 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% of the time) as they do for complicated mappings (43%).

Second, including irrelevant information in the question did not lead to a reliable increase in response accuracy for complicated mappings (47% correct with irrelevant information versus 42% 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 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 present definitions so that it is easier to extract their content. It may be that without resorting to designing more complex NLG systems, the use of bullets, indentation and other formatting features, could make the topics and structure of the definitions more transparent so that respondents can more quickly hone in on the information they need. With the addition of NLG capabilities, more interactive and adaptive systems might be able to diagnose the details of the respondent’s confusion and present just the relevant content of the definitions at the appropriate moment in the dialog. Future research, thus, ought to explore the development of such systems.

Automatically provide clarification when users’ 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 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 [14], 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 for complicated mappings (83%), while high, was no better than when the system could not volunteer definitions (79%). 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%) or could not (49%) volunteer definitions.

We [12] have carried out a laboratory experiment using stereotypic respondent models in an effort to increase the value of system initiated clarification. We focused 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. [35]). Fifty-seven younger respondents (younger than age 35, averaging 26.8 years of age) and 57 older respondents (older than age 65, averaging 72.4 years of age) interacted with one of several user interfaces to the CASI system. The critical interfaces in this study were those that allowed respondents to request clarification and (1) did not volunteer any clarification, (2) volunteered clarification based on a generic respondent model, and (3) volunteered clarification based on a stereotypic respondent model, i.e. inactivity thresholds were longer for old than young respondents. The results showed that increasingly specific user models led to improved response accuracy, although the story isn't entirely straightforward. Younger users seemed to benefit from the more specific models more than older users; older users initiated requests for clarification far less often than younger users, and they found system-initiated clarification more irritating than the younger users did. We conclude that survey interviewing systems and the dialog principles they embody may need to be tuned differently for different populations of users.

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 [34] has argued they will; see also [24]). 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 users' 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.

The text- and mouse-based dialog systems we have implemented here are quite rudimentary relative to what is possible. In particular, we have not yet looked at dialog systems with NLG capabilities. We *have* experimented with speech-based survey interfaces using Wizard-of-Oz-based speech generation ([2],[4]), and although the results are similar to those for the text-based interfaces, they are not identical. Our hope is that the work reported here can help determine when more complex systems need to be designed and when simpler ones will do. To the extent that more sophisticated systems with NLG capabilities are needed, we hope our data can ultimately form a basis for implementing systems that help respondents interpret survey questions more accurately

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants No. SBR9730140 and ISS0081550 and by the Bureau of Labor Statistics. The opinions expressed here are the opinions of the authors and not those of the Bureau of Labor Statistics.

REFERENCES

1. Beatty, P. Understanding the standardized/non-standardized interviewing controversy. *Journal of Official Statistics, 11,* (1995), 147-160.
2. Bloom, J.E. (1999). *Linguistic markers of respondent uncertainty during computer-administered survey interviews.* Unpublished doctoral dissertation, New School for Social Research.
3. Bloom, J.E., and Schober, M.F. Respondent cues that survey questions are in danger of being misunderstood. In *Proceedings of the American Statistical Association, Section on Survey Research Methods, 1999* (Alexandria VA, 2000), American Statistical Association.
4. Bloom, J.E., and Schober, M.F. (in preparation). A speech interface for collaborating with a computer-administered survey system.
5. Blyth, B. 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.* John Wiley & Sons, Inc., New York, 1997, 249-266.
6. Brennan, S.E., and Hulteen, E. Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems, 8* (1995), 143-151.
7. Briggs, C.L. *Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research*. Cambridge University Press, Cambridge UK, 1986.
8. Cahn, J.E., and Brennan, S.E. 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* (Menlo Park CA, 1999), AAAI Press, 25-33.
9. Cawsey, A. *Explanation and interaction: The computer generation of explanatory dialogues.* MIT Press, Cambridge MA, 1992.
10. Cawsey, A. User modelling in interactive explanations. *User Modeling and User-Adapted Interaction, 3* (1993), 221-247.
11. Clark, H.H., and Schober, M.F. Asking questions and influencing answers. In J.M. Tanur (ed.), *Questions about questions: Inquiries into the cognitive bases of surveys*, Russell Sage Foundation, New York, 1991, 15-48.
12. Coiner, T.F., Schober, M.F., Conrad, F.G., & Ehlen, P. Assessing respondents' need for clarification in web surveys using age-based user modeling. In *Proceedings of the American Statistical Association, Section on Survey Research Methods, 2002* (Alexandria VA), American Statistical Association.
13. Conrad, F.G., and Schober, M.F. A conversational approach to text-based computer administered questionnaires. In *Proceedings of the Third International ASC conference*. (Chichester UK, 1999), Association for Survey Computing, 91-101.
14. Conrad, F.G., and Schober, M.F. A conversational approach to computer-administered questionnaires. In *Proceedings of the American Statistical Association, Section on Survey Research Methods, 1998* (Alexandria VA, 1999), American Statistical Association, 962-967.
15. Conrad, F.G., and Schober, M.F. Costs and benefits of standardized and conversational interviewing. In *Proceedings of the Federal Committee on Statistical Methodology Research Conference (FCSM)* (1999), FCSM, 21-30.
16. Conrad, F.G., and Schober, M.F. Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly*, 64 (2000), 1-28.
17. Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nicholls II, W.L, and O'Reilly, J.M.

(eds). *Computer assisted survey information collection*. John Wiley & Sons, New York, 1998.

18. Dillman, D.A. *Mail and Internet surveys: The tailored design method (2nd ed.).* John Wiley & Sons, New York, 2000.
19. Dykema, J., Lepkowski, J.M., and Blixt, S. The effect of interviewer and respondent behavior on data quality: Analysis of interaction coding in a validation study. In *Survey measurement and process quality*, L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), John Wiley & Sons, New York, 1997, 287-310.
20. Fowler, F.J., and Mangione, T.W. *Standardized survey interviewing: Minimizing interviewer-related error*. SAGE Publications, Inc., Newbury Park CA, 1990.
21. Grudin, J. The case against user interface consistency. *Communications of the ACM, 32* (1989), 1164-1173.
22. Gulliksen, J., and Sandblad, B. Domain-specific design of user interfaces. *International Journal of Human-Computer Interaction, 7*(2) (1995), 135-151.
23. Holstein, J.A., and Gubrium, J.F. *The active interview*. SAGE Publications, Inc., Thousand Oaks CA, 1995.
24. Kay, J. Vive la difference! Individualized interactions with users. In C.S. Mellish (ed.), *Proceedings of the 14th International Conference on Artificial Intelligence* (San Mateo, CA, 1995), Morgan Kaufmann, Publishers , 978-984..
25. Kobsa, A., and Wahlster, W. (eds.) *User models in dialog systems.* Springer-Verlag, New York, 1989.
26. Kvale, S. Ten standard objections to qualitative research interviews. *Journal of Phenomenological Psychology,* 25 (1994), 147-173.
27. Lind, L.H., Schober, M.F., & Conrad, F.G. Clarifying question meaning in a web-based survey. *In Proceedings of the American Statistical Association, Section on Survey Research Methods* (Alexandria, VA, 2001), American Statistical Association.
28. Moore, J.D. *Participating in explanatory dialogues: Interpreting and responding to questions in context*. MIT Press, Cambridge MA, 1995.
29. Moore, J.D., and Paris, C.L. Exploiting user feedback to compensate for the unreliability of user models. *User Modeling and User-Adapted Interaction, 2* (1992), 287-330.
30. Nicholls II, W.L., Baker, R.P., and Martin, J. 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.* John Wiley & Sons, New York, 1997, 221-248.
31. Paek, T., and Horvitz, E. 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* (Menlo Park CA, 1999), AAAI Press, 85-92.
32. Pangalo, G.J. Designing the user-interface. *Computers in Industry, 22* (1993), 193-200.
33. Rich, E. Stereotypes and user modeling. In A. Kobsa & W. Wahlster (eds.), *User models in dialog systems.* Springer-Verlag, New York, 1989, 35-51.
34. Rich, E. Users are individuals: Individualizing user models. *International Journal of Human-Computer Studies*, 51 (1999), 323-338.
35. Salthouse, T.A. The processing-speed theory of adult age differences in cognition. *Psychological Review*, 103 (1996), 403-428.
36. Schober, M.F. Conversational evidence for rethinking meaning. *Social Research* (special issue on Conversation), *65*(3) (1998), 511-534.
37. Schober, M.F. Making sense of questions: An interactional approach. In M.G. Sirken, D.J. Hermann, S. Schechter, N. Schwarz, J.M. Tanur, & R. Tourangeau (eds.), *Cognition and survey research* John Wiley & Sons, New York, 1999, 77-93.
38. Schober, M.F., and Conrad, F.G. Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61 (1997), 576-602.
39. Schober, M.F., and Conrad, F.G. Response accuracy when interviewers stray from standardization. In *Proceedings of the American Statistical Association, Section on Survey Research Methods,* (Alexandria VA, 1998), American Statistical Association, 940-945.
40. Schober, M.F., Conrad, F.G., and Bloom, J.E. Enhancing collaboration in computer-administered survey interviews. In Proceedings of the American Association for Artificial Intelligence Fall Symposium *Psychological models of communication in collaborative systems* (Menlo Park CA, 1999), AAAI Press, 108-115.
41. Schober, M.F., Conrad, F.G., and Fricker, S.S. (in press). Misunderstanding standardized language in research interviews. *Applied Cognitive Psychology*.
42. Suchman, L., and Jordan, B. Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association, 85*(409) (1990), 232-253.
43. Suessbrick, A., Schober, M.F., and Conrad, F.G. Different respondents interpret ordinary questions quite differently. In *Proceedings of the American Statistical Association Section on Survey Methods Research*. (Alexandria VA, 2000), American Statistical Association.
44. Traum, D.R. (1994). *A computational theory of grounding in natural language conversation.* Unpublished Ph.D. dissertation, University of Rochester.