ADAPTIVE INTERFACES FOR COLLECTING SURVEY DATA FROM USERS

Michael F. Schober
New School University
schober@newschool.edu

Frederick G. Conrad
Bureau of Labor Statistics
conrad_f@bls.gov

Abstract

This project (IIS00-81550) consists of three sets of laboratory studies which will examine the degree to which different types of user interfaces to survey systems promote uniform interpretations of questions—and thus higher quality data. In particular, the studies focus on improving clarification dialogue and system diagnosis of user uncertainty. Users' response accuracy is assessed on the basis of government definitions for key survey concepts; in conjunction with time and satisfaction measures, accuracy indicates which approaches to interface design are most promising for web based surveys for collecting official statistics. The studies explore desktop (windows, keyboard entry, and mouse clicking) and speech survey interviewing systems, some of which are implemented as prototypes while others are simulated. The primary goals of the project are to inform the design of computerized survey systems, but the results apply to the 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 systems for retrieving information and may have quite different properties.

1. Introduction

This paper presents a set of planned studies – funded by NSF (IIS00-81550) – as well as relevant background research by the authors. The objective of the research program is to determine how best to design 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., Clark & Schober, 1991; Schober, Conrad & Bloom, 2000), which can lead to poor response accuracy under certain circumstances (see Conrad & Schober, 2000; Schober & Conrad, 1997). 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 (Conrad & Schober, 1999).

The planned studies examine which sorts of interfaces to survey systems promote more uniform interpretations of questions—and thus higher quality data. To do this, they contrast interfaces that require users to recognize when they may have interpreted questions in unintended ways with interfaces that diagnose when users need clarification and provide it without being asked. The studies focus on desktop (windows, keyboard entry, and mouse clicking) and speech survey interviewing systems that government agencies might implement to gather the data for influential statistics like the unemployment rate or the ConsumerPrice Index. These surveys ask questions about verifiable facts and behaviors, rather than about opinions and attitudes: if the accuracy of users' responses is verifiable, we can objectively evaluate how different interface designs affect the users' comprehension and, ultimately, the quality of the data they provide. 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 interfaces promote question interpretations that match the official definitions. The purpose of the project is not to implement actual working systems, but rather, to determine what kinds of interfaces will lead to the most accurate responses, fastest sessions, and highest user satisfaction.

The potential practical value of improving survey interfaces is enormous. The kinds of large-scale government survey we are examining (e.g. the current population survey is administered to 40,000 households each month) is currently carried out by expensive interviewing staffs. Self administration of computerized survey instruments could save the government substantial amounts of money and, if respondents prefer the convenience and privacy of self-administration, could increase response rates.

2. Background studies

In an earlier study (Conrad & Schober, 1999; Schober, Conrad & Bloom, 2000) we demonstrated that a desktop survey interface that offers clarification (an official definition) by allowing users to click on highlighted text or offering unsolicited clarification when users are slow to respond improves response accuracy without reducing satisfaction, but at the cost of increased session duration. We could assess response accuracy because users responded on the basis of fictional scenarios for which we knew the correct answer. The improvements in accuracy only came when users were told explicitly that their own conceptions might differ from the survey designers', and that therefore it was essential that they obtain clarification. Users who were merely told that clarification was available if they needed it almost never requested it; they responded confidently (and relatively inaccurately) before the inactivity thresholds that we had set (median response times when clarification wasn't available) were exceeded.

In a follow-up study (Hahn Lind, Schober & Conrad, 2001), we explored whether it is possible to use more practical techniques than explicit instruction to sensitize users to the possibility that they might conceive of terms in survey questions differently than the survey designers. Our approach was to reword the survey question to include a part of an official definition. The rationale was that by exposing users to excerpts from the definitions, they would be more likely to recognize that the ordinary words in the questions might be used in non-standard ways, thus increasing requests (clicks) for definitions. Question rewording did, in fact, increase requests for clarification but, unlike in the previous experiment, this did not increase response accuracy. Users responded as quickly when they obtained definitions as when they did not, suggesting that they were not reading the definitions carefully. The current research focuses on improving clarification dialogue so that it is easier for users to obtain the part of a definition relevant to their particular uncertainty and on improving the system's ability to diagnose user uncertainty so that the burden is not solely on the user to recognize when to request clarification.

3. Planned Studies

The research consists of 7 laboratory studies organized around three topics: (A) adaptive diagnosis of users' need for clarification, (B) adaptive clarification of survey concepts, and (C) when systems should take the user's perspective on survey concepts. The interfaces include:

- working desktop survey systems
- simulated desktop survey systems, where users can type natural language input and the system seems to understand them (actually, an experimenter presents the typed responses)
- simulated speech interfaces, where users speak into headsets and the system seems to understand them (actually, an experimenter presents text-to-speech responses)

3.1 Adaptive diagnosis of the user's need for clarification

How adaptive should the diagnosis of a user's need for clarification be? Are generic indicators (like user inactivity) sufficiently diagnostic? Or should diagnosis of the need for clarification be adapted to individual users, based on their group membership (age, culture, education, expertise) or on their observed individual baseline behaviors (Kay, 1995)? The following studies are based on the fact that survey system users do not ask for clarification nearly as often as they need to, and thus that they need to be provided with unsolicited clarification if response accuracy is to improve.

STUDY A1: Stereotypic user models in a desktop survey interface

In this study we examine how adjusting a system's thresholds for providing clarification based on users' group membership affects data quality and user satisfaction. We focus on users' *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). In other words, a slow answer by a slow user should be less diagnostic of potential need for clarification than a slow answer by a fast user.

Design. 120 users (60 younger than age 40 and 60 older than age 65) will interact with 6 different interfaces (20 users per interface). The first interface serves as our baseline: (1) *No clarification* available (today's standard procedure); it also allows us to determine the average response latencies per question for those situations where clarification is most likely to help, which will be used in constructing interfaces (3) and (4) below. Three interfaces explicitly compare three alternate views of the utility of user models: (2) *User-initiated clarification* (No user model to diagnose need for clarification—user can click on highlighted text to get clarification, but the system never offers clarification), (3) *Mixed-initiative clarification, generic user model* (user can get clarification, but system also provides unsolicited clarification if user is slow to answer, relative to the average user); (4) *Mixed-initiative clarification, stereotypic user model* (user can obtain clarification, but system also provides unsolicited clarification if user is slow to answer, relative to the average user in his or her age group). Two other interfaces provide necessary control groups that rule out alternative possible explanations for results: (5) *Random clarification* (user is given clarification as often as in conditions 3 and 4, but on a random basis), and (6) *Clarification always* (user is always given clarification whether or not it is needed).

Predictions. If generic user models improve the system's ability to diagnose when to clarify, then task accuracy should be better with a generic user model than without (even if the effects on task duration and user satisfaction do not correlate with the effects on task accuracy). Furthermore, if stereotypic user models really help systems to diagnose when to clarify, then task accuracy should be better with stereotypic models than with generic models (again, the effects on task duration and user satisfaction may go in the same direction or not). Alternatively, when the system diagnoses the user's need for clarification accuracy may be no higher than when users simply request clarification or when the system provides clarification randomly.

STUDY A2: Stereotypic user models in a speech survey system

In this study, we examine the same questions as in Study A1, but in the context of a simulated (Wizard of Oz) speech interface (see Dahlbäck, Jönsson, & Ahrenberg, 1993). The system will present survey questions (using an artificial-sounding text-to-speech computer voice, to enhance believability) and users will listen and respond using a telephone headset. From an earlier study using this setup (see Schober, Conrad & Bloom, 2000), we know that users behave differently with a speech survey interface than with a desktop interface: they ask for clarification less often, as one would expect from Clark & Brennan's (1991) theory on the costs and constraints of grounding understanding in different media. And they benefit more from system-initiated clarification even without strong instructions to ask for clarification.

Users interacting with a speech system also provide more potential cues that could allow a system to diagnose their need for clarification. We have demonstrated that in spoken interviews people are not only more likely to pause longer in their first utterance after the question when they need clarification; they are also more likely to be disfluent—to say *um* or *uh*, to repeat words, or to restart words or phrases (Bloom & Schober, 2000). They are also more likely to describe their circumstances rather than answering questions directly (Bloom & Schober, 2000; see also Schober & Conrad, 1997). Thus, when a speech system uses such cues to diagnose respondents' need for clarification, task accuracy improves, without reduction in user satisfaction (see Schober, Bloom & Conrad, 2000).

Predictions. The logic of Study A2 mirrors that of A1, but we expect the results may not look the same because of the differences between desktop and speech interfaces in the resources for users and systems to ground understanding (Clark & Brennan, 1991). A speech system has more cues for diagnosing a user's need for clarification, but users may be less likely to make the effort to formulate questions, and they may find listening to text-to-speech clarification irritating (see Schober, Bloom & Conrad, 2000).

STUDY A3: Individualized user models in a desktop survey system

Study A3 examines 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 user's baseline performance, rather than basing its diagnoses on the less fine-grained criterion of group membership. This study examines whether the extra programming and design effort that this would entail provides sufficient practical benefits to be worth it.

20 users (10 younger and 10 older) will answer the same questions and scenarios as in Study A1, for comparability. This time the system will provide unsolicited clarification differently for each user, based on information collected from the user during practice questions: during the practice questions, the system will monitor how quickly each user answers them. (Practice questions will be selected from the same surveys as the test questions, and corresponding scenarios will be created; pilot testing will determine which practice questions best predict response times for survey questions). Because we will have data from Study 1 on how user 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 user.

STUDY A4: Individualized user models in a speech survey system

Study A4 examines whether individualized user models improve response accuracy for our speech survey system above and beyond any improvements shown in A2 for generic or stereotypic user models. The design is the same as for Study A3, except that during the (to-be-determined) practice questions the system (with the wizard's help!) will monitor not only response latencies but also rates and types of disfluencies for each user. That is, if a user constantly says *um*, its presence in the first utterance following a survey question is less likely to be diagnostic of the user's need for clarification than if the user rarely says it. The same should be true of other potential indicators of the need for clarification like restarts and repeats.

3.2 Adaptive clarification in survey interfaces

In this set of studies we examine how survey systems should present clarification of question concepts, which are often complex, typically composed of several subdefinitions. Should systems reason to determine exactly the information the user needs to know? Developing such systems would require a major research and development effort. Or is it sufficient to present more information than necessary (for example, the entire official definition of "a job," as in the planned studies in section 3.1 and require users to extract what they need? The latter approach would certainly be simpler technologically.

The common wisdom in HCI is that interface design should be "user-centered" (e.g. Norman & Draper, 1986): users should be in control of the interaction, and interfaces should be designed to fit users' ways of thinking, rather than requiring users to learn counterintuitive commands or concepts (e.g. Nielsen, 1993). Systems should do what it takes to ease the user's interpretive burden. Although these are sensible principles, it isn't clear exactly how they should be implemented in survey systems. Obviously users don't control the initiative in a computerized survey; even if a user voluntarily agrees to participate in a survey, it is still up to the system to present the questions in the order determined to be appropriate. One might propose that users should be in control of any clarification subdialogs but we already know that if it is left entirely up to users to initiate clarification they are unlikely to do so as much as they need to, because it often won't occur to them that their conceptions of ordinary words like "job" and "furniture" may not correspond with other people's.

STUDY B1: How should clarification be tailored to the user's specific informational needs?

In Study B1, we examine whether tailoring clarification to particular users' needs improves task accuracy or user satisfaction sufficiently to justify developing the natural language understanding software that would be needed to implement it. 60 users will interact with four different versions of a desktop survey interface (15 users per interface) with mixed-initiative clarification: users can request clarification, or the system will initiate a clarification subdialog if they are slow to respond. In the two control conditions, users will be presented with full definitions or with no definitions at all. The other two interfaces will implement two alternatives for tailoring clarification to users; in both cases users will interact with a Wizard rather than a working system. In one, users are empowered to formulate precise questions (by typing into a dialog window), and the system (Wizard) will provide specific answers, that is, the appropriate parts of the relevant definitions. If users type vague questions, the system will prompt them to be more specific. If users are slow to answer, the system will present a dialog box with the question “What do you need to know?”, prompting users to type specific questions. In the other interface, the system constrains the user’s queries. Users are presented with a menu of topics that the system knows the answer to (e.g., “I know the answer to: (A) Does a person who works for multiple employers have one job or more than one job? (B) Does a person who receives pay in kind have a job?” etc.), and they click on a topic to get the answer.

These two conditions thus pit user-formulated querying against system-constrained querying. In the first case, users may appreciate the value of clarification and ask for it more often because they believe they will truly be understood. In the second case, may be appreciate and ask for clarification more often because it is clear to them what the system can offer, but they may be put off or even bewildered by the array of options.

3.3 User vs. system perspectives in survey interfaces

Although user-centered design demands that interfaces should fit users’ ways of thinking, it is unclear how best to do this in surveys. One possibility is to require the agent who does *not* initiate a clarification dialog to adopt the perspective of the agent who does. Users of survey systems might believe that because the system asks the questions, it is responsible for what it means (following the ordinary “principle of speaker’s meaning,” Clark & Schober, 1991), and it should be responsible for making sure the user has interpreted questions as it intends. By this logic, information retrieval systems should take the user’s perspective, because the user is the one who initiates the commands, and information collection systems should require the user to take the system’s perspective, because the system initiates the questions.

Under a second approach, one could argue that users should always take the system’s perspective because computer systems are limited partners, without human comprehension abilities.A third approach would be to design systems to always take the user’s perspective, because users should generally be in control.

STUDY C1. Should systems force users to understand the survey designers’ concepts?

Although in B1 the clarification is more tailored to users’ needs, the interfaces still require users to answer from the system’s perspective—to be educated about the survey designers’ conceptualizations. One could imagine a still more user-centered approach to survey interfaces, in which systems don’t challenge users’ intuitions about what survey questions mean. Rather, the system could reason or engage in dialog to figure out what the correct answer should be without requiring users to understand survey designers’ (possibly idiosyncratic) definitions.

In this study, we contrast these two approaches in a simulated (Wizard of Oz) adaptive speech interface. In the first approach, the system presents official definitions when users ask for or seem to need clarification, but then requires users to determine, given the definition, what the correct answer should be. In the second approach, the system queries the user who gives evidence of needing clarification about the component parts of the survey concept (e.g., “Do you have any rooms originally designed as bedrooms that are being used for something else?”—and then decides for the user what the correct answer should be. In doing so, it can either inform the user about its ultimate decision (“From our perspective, you have three bedrooms”) or not.

C.2. What allows users to maintain the system’s perspective?

Although users may prefer not to learn any complex survey definitions, they be more inclined to do so if they learn the definitions the first time the concepts appear but don't have to relearn them in later questions. In some surveys the concepts build upon each other, and entire series of questions can refer back to the same definitions. In such surveys users may prefer for concepts to be clarified once early on. Alternatively, users may prefer clarification subdialogs question by question only when they seem relevant, and they may not want to learn about components of official definitions that aren't relevant to the current question.

125 users will interact with 5 interfaces (25 per interface). In the baseline interface (1), users won't have the option of getting clarification, and their baseline response times will be monitored. In the *Mixed initiative clarification* interface (2) users can click on highlighted terms to get definitions; if they are slow to respond the system will provide all the definitions relevant to the current question—even if they have appeared before. In the *Definition once, user-initiated clarification* interface (3), users are presented with the official definitions only the first time survey concepts appear in a question. After that, the definition will appear again only if the user clicks on highlighted text in a subsequent question. In the *Definition once, mixed-initiative clarification* interface (4), users are presented with the official definitions only the first time survey concepts appear in a question. After that, the definition will appear again if the user clicks on highlighted text in a subsequent question, or if the user is slow to respond. In the *Constant reminder* interface (5), the definitions relevant for each question appear on the screen when the question is presented, no matter how many times the definitions have been presented before.

Predictions. If users can maintain the system's perspective throughout an entire session, they should perform as well with and prefer Interface 3 to Interfaces 4 and 5. If users don't learn definitions well when they are first presented, they should produce more accurate data with Interface 5, and they might even appreciate the extra support.

4. Implications and Future Directions

We chose these interfaces primarily because they are easy to implement in the laboratory. Other user actions or user characteristics could be modeled, and other dialog interfaces could be implemented. An improved survey system of the future might be one that allows users to choose whether to speak or click, and to read or listen to tailored or untailored clarification.

More broadly, we plan to compare survey systems with other information collection systems, like electronic voting, where system clarification might influence opinions in unknown ways. For example, in voting systems, system-initiated information about a candidate might better inform the voter, or it might bias the results, or both. As we see it, enhancing interfaces with dialog capabilities may not always lead to improved outcomes. The current set of studies begins to map out this territory.

References

Bloom, J.E., & Schober, M.F. (2000). 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: American Statistical Association.

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: APA.

Clark, H.H., & Schober, M.F. (1991). 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.

Conrad, F.G., & Schober, M.F. (1999). A conversational approach to text-based computer-administered questionnaires. In *Proceedings of the 3rd International Conference on Survey and Statistical Computing* (pp. 91-101). Chesham, UK: Association for Survey Computing.

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

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies—why and how. *Knowledge-Based Systems, 6,* 258-266.

Hahn Lind, L., Schober, M.F. and Conrad. F.G. (2001). Clarifying question meaning in a web-based survey. Paper presented at annual conference of the American Association of Public Opinion Research. Montreal, Canada.

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.

Nielsen, J. (1993). *Usability engineering*. Cambridge, MA: AP Professional.

Norman, D.A, & Draper, S.W. (eds.) (1986). *User-centered system design: New perspectives on human-computer interaction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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, 60,* 576-602.

Schober, M.F., Conrad, F.G., & Bloom, J.E. (2000). 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* (pp. 108-115). Menlo Park, CA: AAAI Press.

Schober, M.F., Conrad, F.G. & Bloom, J.E. (2000). Clarifying word meaning in computer-administered survey interviews. In *Proceedings of the Twenty-second Annual Conference of the Cognitive Science* Society. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, pp. 447-452.