# ASSESSING RESPONDENTS' NEED FOR CLARIFICATION IN WEB SURVEYS USING AGE-BASED USER MODELING[1]

**Tania F. Coiner, New School for Social Research**
**Michael F. Schober, New School for Social Research**
**Frederick G. Conrad, Bureau of Labor Statistics**
**Patrick Ehlen, New School for Social Research**
**Tania Coiner, Dept. of Psychology AL-304, New School University,**
**65 Fifth Ave., New York, NY 10003**

**Key Words: Web Surveys, Data Quality, Question Comprehension, User Interfaces, User Modeling, Question Clarification**

## INTRODUCTION

Respondents in standardized surveys tend to assume that their definitions of everyday terms such as "bedroom" or "job" must match those of the survey designers, even though we know that they often differ substantially. Even when they are offered clarification, they often do not request it because they do not think that it is needed. In our earlier studies of telephone interviews, we found that respondents answer more accurately when they receive clarification about question meaning (Schober & Conrad 1997, Conrad & Schober 2000) This is also true for web survey interfaces (Schober & Conrad, 1998) and applies whether the respondent requests the clarification (by clicking to get official definitions) or the system offers unsolicited clarification.

The distinction between respondents requesting clarification and systems offering it reflects a longstanding debate in the human-computer interaction community between two approaches to interface design: those that emphasize giving users control (e.g. Shneiderman, 1997), where users can adjust the interface as desired, and those that emphasize user modeling, where interfaces automatically adapt to different users (Maes, 1994).

In this study, we contrast typical web survey interfaces (usually standardized for everyone) with interfaces based on user control and also on user modeling (e.g. Kay, 1995). We implemented simple user models that diagnosed respondent uncertainty. If respondents were inactive (no clicks, no typing) for more than a particular duration, this was treated as a signal of uncertainty and triggered the system to clarify the likely source of uncertainty by providing a definition.

We contrasted two variants of this type of user-model. One was a generic model, with thresholds based on how long an average user took to answer a particular question. The second was a group-based model, with thresholds based on how long average users within different groups took to answer a particular question.

For this study, we formed our groups based on age. Survey methods research has shown that age affects responding, largely because working memory declines (e.g., Knäuper, 1999). More germane to our application, the cognitive aging literature documents a more general slowing of behavior with age (e.g., Salthouse, 1976). Therefore one might expect older web survey users' response times to be slower than younger users' times. If that's the case, the same period of inactivity by old and young users may mean different things; a short lag may indicate confusion for a young user but simply ordinary thinking for an older user.

In the current study we contrasted five user interfaces in the laboratory. In the first there was no clarification available to users. The second was user-initiated, where clarification was available if the user requested it by clicking. The third embodied a generic user model, where the respondent could request clarification but the system provided clarification if the respondent's inactivity exceeded a fixed threshold. The fourth was built around group-based user models, identical in approach to generic user models except that the inactivity threshold was differed for different groups of respondents. In the fifth interface, the definition always appeared with the survey question.

## EXPERIMENTAL DESIGN

*Questions.* All respondents answered the same 10 questions about housing and purchases from two ongoing government surveys (used by Conrad & Schober, 2000). Each respondent answered five purchase questions and five housing questions. Half of the respondents answered the housing questions first, and the other half answered the purchase questions first.

*Scenarios.* All respondents answered the questions on the basis of fictional scenarios for which we knew the correct answer, enabling us to measure response accuracy. The questions were presented on a computer

[1] We thank the 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.

and respondents were given a packet of scenarios in the form of floor plans, receipts, and short narratives. The respondents received a total of 10 scenarios, one per question.

Half of the scenarios were designed to be hard to answer correctly without access to the official definition. We call these complicated scenarios. The other half were designed such that, without the use of definitions, respondents would be likely to interpret them as the survey designers intended. We refer to these as straightforward scenarios.

Here is an example of a complicated scenario for the question "How many people live in this house?"

> The Gutierrez family owns the 4-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, one for Linda, and one for Sandy, who is employed by the family as a nanny.

It is complicated because Sandy's status is ambiguous without knowing the definition of living in a house.

*Interface.* Questions were presented to laboratory respondents on a computer using a web-browser interface (see Figure 1). Respondents answered questions by selecting radio buttons with a mouse for 'yes'/ 'no' questions or by typing with the keyboard for questions requiring a numerical answer. In the conditions where they were able to request clarification, respondents clicked on a hyperlinked term or phrase (see Figure 2) and the system displayed the definition (see Figure 3). When the system initiated the clarification, the definition simply appeared after appropriate threshold (Figure 3).

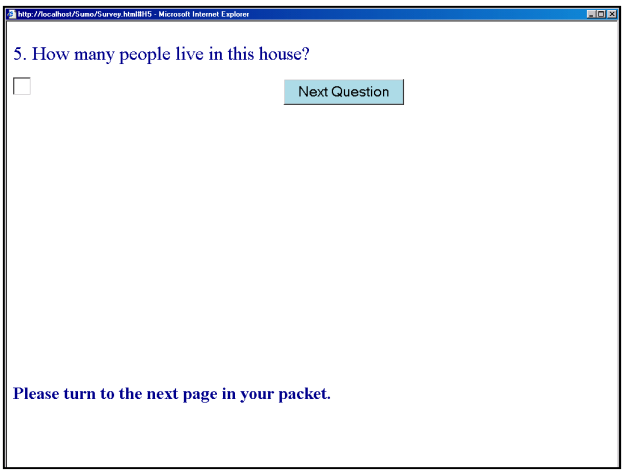


**Figure 1. Survey question with no clarification available**

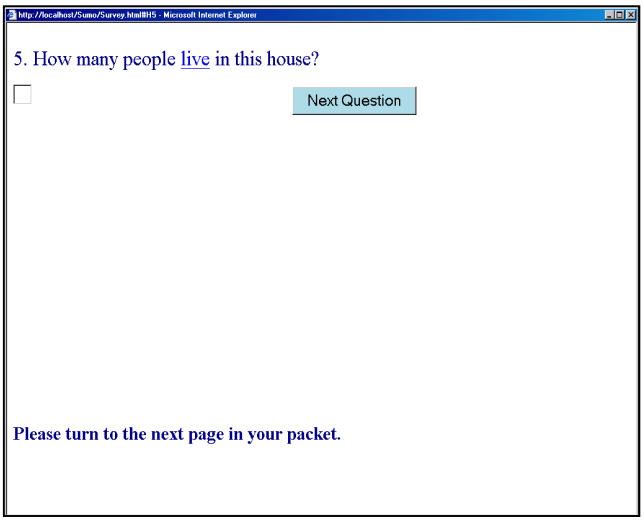


**Figure 2. Survey question with hyperlink**

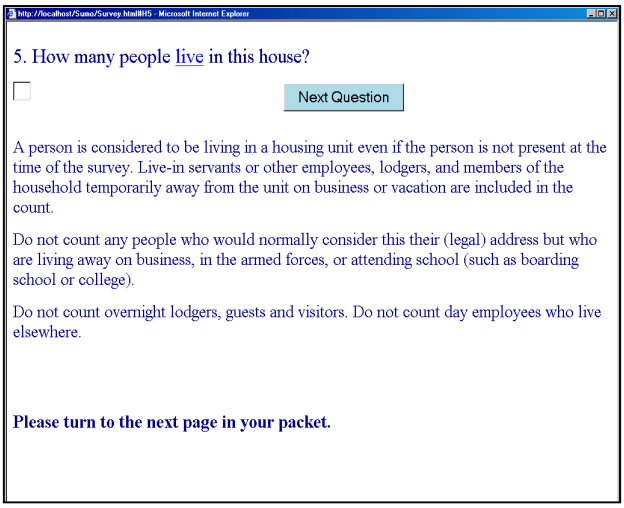


**Figure 3: Survey question with definition displayed**

*Thresholds.* To establish the inactivity thresholds, we examined response times for the first 20 respondents in the no-clarification condition as well as the response time for the 12 respondents in the user-initiated condition who did not request clarification. Across the questions, response times for straightforward and complicated items were most different at the 40$^{th}$ percentile, so we used this time as the inactivity threshold in the generic user model. The group-based user models were also based on the 40$^{th}$ percentile response time for complicated mappings but computed separately for old and young users.

*Participants.* 114 paid participants were recruited from the New York City area by means of an advertisement in the *Village Voice* and recruitment by

fliers at Senior Centers. There were 56 females and 58 males. Half of the participants were young (defined here as less than 35 years old) with a mean age of 26.8, and half were old (defined as over 65 years old) with a mean age of 72.4. Ethnicities, educational backgrounds and experience with computers were roughly balanced across age groups.

## RESULTS

*Response accuracy*. As can be seen in Figure 4, all respondents were quite accurate when answering on the basis of straightforward scenarios (95% of questions answered correctly); for complicated mappings, accuracy varied depending on how and when respondents received clarification, interaction $F(4, 104) = 16.58$, $p<.001$. Accuracy increased linearly across the five groups, linear trend $F(1,104) = 8.16$, $p<.001$. When respondents could not obtain clarification at all for complicated mappings, accuracy was quite poor (24% of questions answered correctly). When the system didn't provide clarification, but respondents could obtain definitions by clicking on hyperlinks, accuracy was better, but still poorer than when the system also clarified concepts (35% of questions answered correctly). Presumably this difference reflects the occasions on which respondents did not realize their interpretation differed from the designers' and the additional system-initiated clarification improved accuracy. Accuracy was better when the system took respondent's age into account (group-based user modeling) than when thresholds were set for the average user (generic user modeling) (48% of questions answered correctly for generic user modeling and 58% correct for group based). Accuracy was best of all when respondents received definitions along with the questions (70% of questions answered correctly).

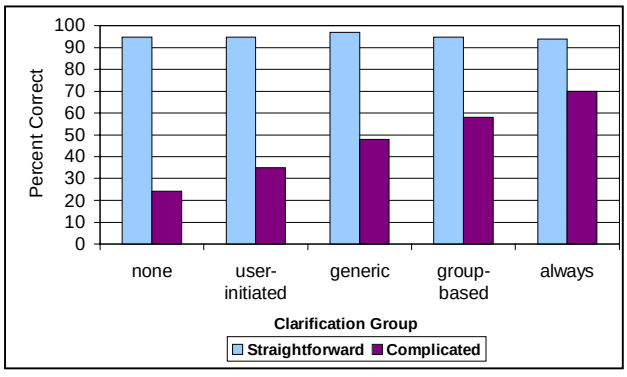


**Figure 4. Response accuracy for all ages**

Although group-based user modeling boosted accuracy above generic user-modeling, this was based mostly on the accuracy of younger users, interaction $F(4,104) = 3.22$, $p = .016$ (see Figure 5). Older respondents performed equally well with generic modeling (50% of questions answered correctly) and with group-based modeling (46% of questions answered correctly). In contrast, younger respondents performed better with group-based modeling (70% of questions answered correctly) than with generic modeling (46% of questions answered correctly). In fact, younger respondents performed as well with group-based user modeling as when they always received definitions (72% correct).

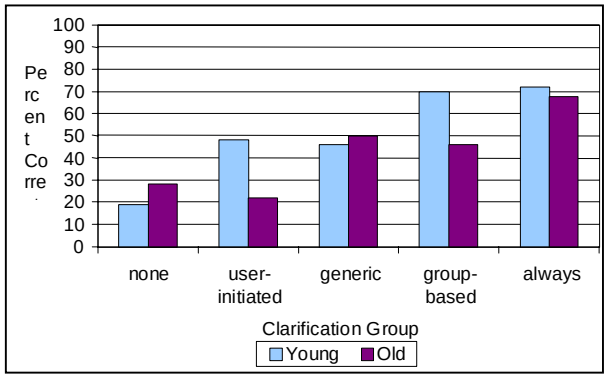


**Figure 5. Accuracy by age for complicated mappings**

*Rates of clarification seeking*. Figure 6 indicates that younger users received clarification more often with group-based modeling (94% of the questions) than with generic modeling (72% of the questions). Older users received clarification slightly more often with generic than group-based modeling (for 78% of the questions with generic and 68% of the questions with group-based), interaction $F(1,36) = 4.82$, $p = .035$. Apparently, accuracy corresponds with how often respondents receive clarification, accounting for the younger respondents' greater accuracy in the group-based condition.

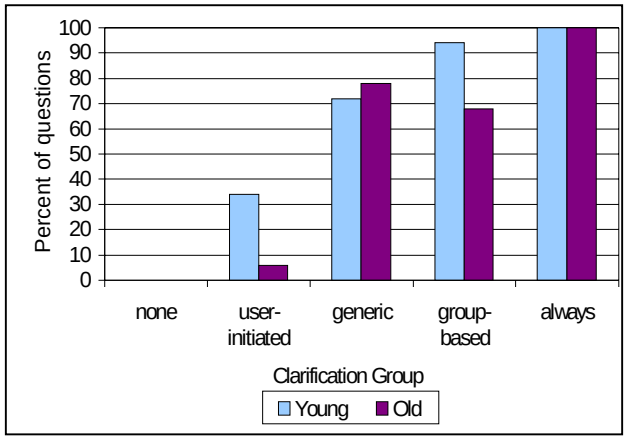


**Figure 6. How often respondents received definitions for complicated mappings**

The generally low rates of user-initiated clarification (see Figure 7) indicate that respondents were not good at recognizing when they needed clarification. This is especially true for older respondents who initiated clarification far less often (8% of the time) than younger respondents did (35.3%) $F$ (1, 54) = 14.45, $p$<.001. The difference between the bars in Figure 6 and Figure 7 is due to system-initiated clarification, indicating that much of the clarification was provided by the system.

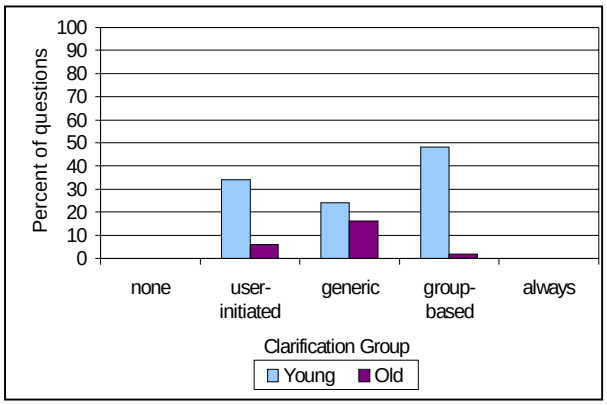


**Figure 7. Respondent initiated requests for clarification (complicated mappings)**

*Response time.* As one might expect, clarification takes time. In Figure 8, we see that respondents took longer to answer questions when they received definitions than when no clarification was available. Respondents were fastest (and least accurate) when the only clarification they received was respondent-initiated (18.6 seconds for younger users and 32.5 seconds for older users). Respondents took longest (and were most accurate) when definitions were displayed all the time (37.2 seconds for younger users and 58.2 seconds for older users), simple contrast $F$(1, 104) = 4.82, $p$<.001. We also see that in all the groups, older respondents took substantially longer than younger respondents did, consistent with the Salthouse (1982) finding (30.3 vs. 44.1 secs) $F$(1, 104) = 25.03, $p$<.001.

*User satisfaction.* Respondents in both age groups were relatively satisfied with respondent-initiated clarification (3.36 out of 4 points) more so than with clarification that was also initiated by the system, always present or not available. This preference for respondent-initiated clarification was, apparently, not related to accuracy: recall that respondents were least accurate when the system never initiated clarification. The older respondents were least happy with group-based user modeling (rating of 2.2), perhaps because the definitions were initiated by the system more often and came after they had already formulated an answer.

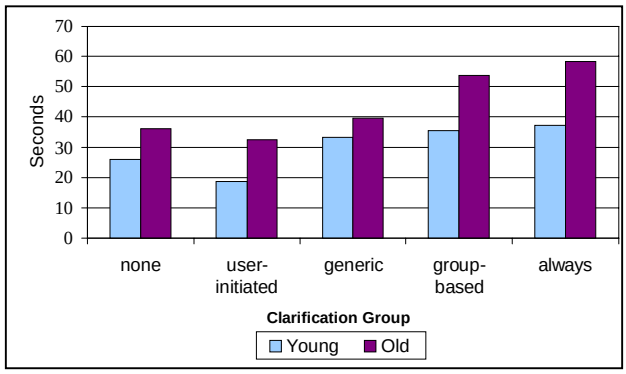


**Figure 8. Overall response times**

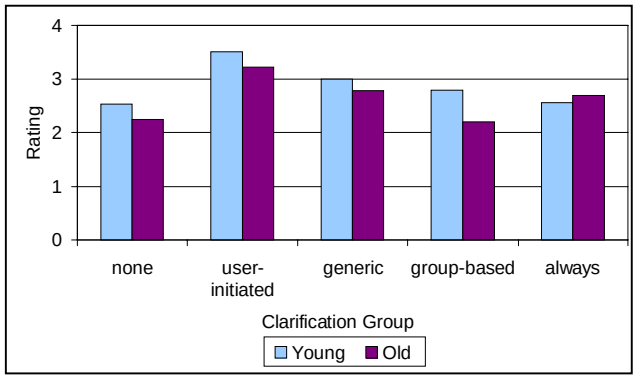


**Figure 9. Satisfaction ratings**

As another measure of user satisfaction, shown in Figure 10, we asked respondents whether they would prefer future surveys like this with an actual interviewer or with a computer. More of the older respondents said they would prefer a human interviewer, especially when they couldn't get clarification (80% preferring an interviewer), or with group-based user modeling (70% preferring an interviewer). Younger users tended to prefer a computer, only preferring a live interviewer when they received clarification all the time (56% preferring an interviewer), perhaps because they expected an interviewer would provide clarification only when they needed it.

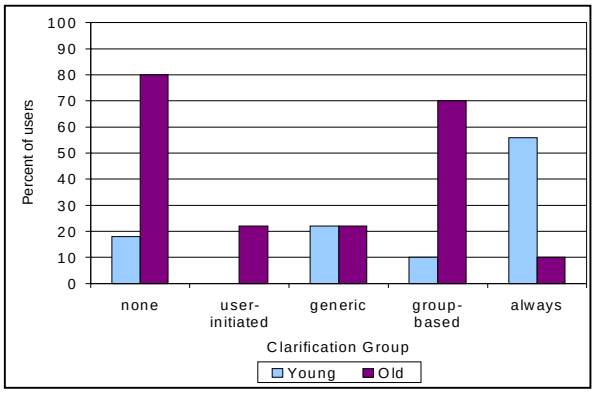


**Figure 10. Respondents preferring interviewers over computers**

## DISCUSSION

Our data are encouraging about the prospect of modeling group differences to improve comprehension and thus increase response accuracy in computer-administered surveys. In this set of interfaces, older respondents were indeed slower than younger respondents were. Accuracy is better when the system also provides clarification than when it only relies on respondents to determine when they need it. Providing definitions along with all questions may be a simple, effective way to improve accuracy, but it is unknown if people will read definitions all the time in an actual web survey. It is also unknown whether in an actual web survey the extra time required for clarification would decrease user satisfaction and increase break-offs. This suggests to us that user modeling is a good idea to the extent that users reliably display uncertainty when they actually need clarification.

In addition overall speed of responding, it might be helpful to model other age-related characteristics of respondents, e.g. working memory capacity (Knäuper, 1999). Other characteristics beside those that are age-related may also be candidates for modeling, e.g. computer experience or education. Finally, instead of group-based characteristics, it might be possible to construct individual user models. Individual respondents' uncertainty could be assessed with inactivity – just like the groups in the current study. However, individual thresholds would be set on the basis of earlier behavior in a web sessions – e.g. response times on a small number of questions requiring clarification. It is in this direction that we turn our attention next.

## REFERENCES

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

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

Kay, J. (1995). Vive la difference! Individualized interaction with users. In CS. Mellish (Ed.) *Proceedings of the 14th Joint Conference on Artificial Intelligence*, pp. 978-984. San Mateo, CA: Morgan Kauffman Publishers.

Knäuper, B. (1999). Age differences in question and response order effects. In N. Schwarz, D. Park, B. Knäuper, & S. Sudman (eds.), *Cognition, aging, and self-reports.* Taylor & Francis, Philadelphia.

Maes. P. (1994) Agents that reduce work and information overload. *Communications of the ACM*, 37, 31-40.

Salthouse, T.A. (1976). Speed and age: Multiple rates of age decline. *Experimental Aging Research, 2*(4), 349-359.

Salthouse, T.A. (1982). *Adult Cognition: An experimental psychology of human aging.* Springer-Verlag: New York.

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.

Shneiderman, B. (1997). Direct manipulation for comprehensible, predictable, and controllable user interfaces. *Proceedings of IUI97*, 1997. International Conference on Intelligent User Interfaces, Orlando, FL, January 6-9, 1997, 33-39.