Mid-Western Educational Researcher

Volume 20 | Issue 1 Article 5

2007

Interactive Features of Web Surveys

Frederick G. Conrad
University of Michigan

2006 MWERA Keynote Address

Interactive Features of Web Surveys

Frederick G. Conrad
University of Michigan
University of Maryland

Educational researchers and policy makers have come to rely on data from sample surveys. Survey research on educational issues poses some special challenges. For example, the hierarchical sample designs in which teachers or students or principals may be sampled within schools which may themselves be sampled within counties which may themselves be sampled within states can be difficult and expensive to implement. However, in many respects the survey methodology issues in educational research are the same as those throughout the social and behavioral sciences. By in large these issues concern obtaining the best quality data for the lowest cost, whether quality pertains to representation, such as the degree to which the sample is a microcosm of the population, or measurement, such as the degree to which respondents' answers accurately reflect their circumstances or characteristics. A major factor in survey quality and cost is the mode of data collection, e.g., telephone interviews versus mailed paper questionnaires.

Telephone interviews are almost sure to be less expensive than face-to-face interviews but telephone interviews can only take place with people who have landline telephones. Those without any phones or "mobile-only" users are not included in the frames, or lists of phone numbers, available for selecting samples. In this paper, I discuss recent work on web surveys which is an important, emerging mode of data collection. My focus is on measurement issues in web surveys, in particular how the interactive character of the web can be exploited to promote better and more uniform understanding of survey questions, and can promote completion of questionnaires once respondents begin them.

Interactivity in web surveys

What do I mean by interactivity? Certainly a paper questionnaire is static and does not react to the respondent's actions, beyond revealing additional questions when the respondent turns the page. An interview, in contrast, is highly interactive because two animate people are conversing and each can react to how the other person behaves. For example, if the respondent does not provide an answer from the list of response choices, the interviewer can repeat the choices or repeat the question or, although it is a violation of most interviewing rules, the interviewer can choose the option that seems closest to what the respondent says. A questionnaire administered on the web is usually somewhere in between a self-administered paper form and an interview. For example, it can be a lot like a paper questionnaire if it is just a form into which respondents enter their choices. Alternatively, questionnaires on the web can be designed to react in many ways to what the respondent does. For example, once the respondent answers a question, the questionnaire can "grey out" the question; if a respondent's answers to a multi-part question must sum to a fixed amount (e.g., 24 hours or 100%), the questionnaire software can check that the answers do in fact add up to this total and alert the respondent if they do not; and the questionnaire can determine what question to display based on the respondent's answer to the previous question. This interactivity gives web questionnaires some of the character of an interview, even though they are self-administered, and allows us in principle to combine the best of interviewer- and self-administration. (More extensive discussions of the interactivity concept are provided by Kiousis, 2002, and McMillan & Hwang, 2002)

Of course interactivity comes about only if it is "designed into" the questionnaire. A web-based questionnaire is not interactive if, for example, it is designed as a single scrollable form in which the respondent answers all questions before submitting her answers. In other words, questionnaires are not interactive if there is no "back and forth" between respondent and system until the questionnaire is completed. One influential text (Dillman, 2000) has advocated designing web questionnaires so that they emulate their paper precursors: "Present each question in a conventional format similar to that normally used on paper self-administered questionnaires" (p. 379). Dillman's (2000) recommendation comes largely from his concern that web-specific features require more bandwidth and computational resources than are available to many users. While designers should certainly be sensitive to this, the kind of interactive features we are concerned with typically involve standard HTML code or small Java scripts that download and execute quickly. Moreover, by treating the web as if it were paper, one fails to capitalize on features that may potentially improve data quality.

We explore three types of interactivity here. In the first, the system displays progress information ("percent completed") on each page and updates this as each additional question is completed. Respondents seem to use this feedback in deciding for each question whether to press the "next page" button or instead to break-off or terminate participation in the survey. We vary the way progress is calculated and examine the effect on break-off rates and the respondents' experience. In the second, the respondent's can request a definition to clarify a concept in the question

We thank the National Science Foundation (Grants SES0106222 and IIS-0081550), and the National Institutes of Health (Grant R01 HD041386-01A1) for their support.

that he or she may not understand by clicking a hyperlinked word or phrase; in response, the system displays a definition of the word or phrase. If this proves useful, it could have an effect on the likelihood of requesting subsequent definitions in subsequent questions. We vary the usefulness of the definitions and examine their effect on future requests. In the third type of interactive sequence, it is actually the respondent's *inaction* (no typing or clicking) that triggers a system (questionnaire) action. The system interprets the lack of respondent actions as an indication that the respondent is confused or uncertain about the meaning of the question and provides a definition; this should in turn have an effect on the respondent's understanding of the question and the accuracy of her responses. We programmed the questionnaire to offer clarification after different periods of inactivity for different groups of respondents and examined the effect on response accuracy and satisfaction with the experience.

Our focus on measurement issues in web surveys is not meant to imply that all is well with respect to representation. In fact, web surveys have been criticized because the degree to which results can be generalized to a general population is uncertain (see for example Couper, 2000). Web survey frames—typically lists of voluntarily provided email addresses—include only those with internet access and who wish to be contacted about participating in surveys. This leads to sample characteristics that are quite different than the general population. Nonetheless, for methodological studies such as those presented here, the main point is that participants are randomly assigned to experimental conditions, whatever population they ultimately represent.

Effectiveness of Progress Indicators

Paper questionnaires inherently communicate information about respondents' progress: the thickness of the yet-to-be-completed part of the booklet provides immediate and tangible feedback to the respondent about how much work remains. This is also the case in long, one-page or non-interactive web questionnaires, where the size and location of the scroll bar convey progress information. In more interactive designs, for example in which one question is presented per page, there is no default progress information. However, the display of progress information can be designed into the questionnaire—typically as either graphical or textual progress indicators. If progress feedback does not reduce break-offs relative to no such feedback, the investment of resources to make it available is almost certainly not worthwhile. And if respondents are discouraged by the rate of their progress, then communicating progress information might actually increase break-offs relative to no progress information. This is surely not worth the expenditure of resources! But if fewer respondents break off when they know how much more of the questionnaire remains, progress indicators may be a valuable addition to the design of web questionnaires.

Background. The evidence about the effectiveness of progress indicators in web surveys is limited and mixed. In one study (Couper, Traugott & Lamias, 2001), there was no difference in response rates when progress indicators were used and when they were not used. Couper et al. (2001) proposed that because their progress indicator was a graphical image (similar to a pie chart indicating percent completed), the questionnaire on which it was displayed took longer, page-by-page, to transfer to respondents' computers than did a questionnaire with no progress indicator. This extra download time, they propose, was a deterrent to completing the questionnaire, thus mitigating any advantage from the feedback. Crawford, Couper and Lamias (2001), controlled transfer time and actually found a lower response rate when progress indicators were used than when they were not. They observed that much of the abandonment occurred on questions requiring open-ended responses, presumably a more difficult response task than selecting from fixed choices. They report results from a follow-up study in which the problematic questions had been excised from an otherwise identical questionnaire. The respondents who were given information about their progress completed the questionnaire at a four percent higher rate than those who were not given progress information.

Part of the explanation for these mixed results may have to do with what information is actually conveyed by the progress indicator. Crawford, et al. (2001) suggest that the progress indicator may have understated actual progress thus discouraging respondents who (correctly) believed they were further along than indicated. In particular, respondents completed almost 40 percent of the questionnaire in the first 20% of the elapsed time spent on the questionnaire response task. In general, discouraging information, for example that the task will take a long time or more time than expected, may well deter respondents from completing the questionnaire. And the timing of the information may matter as well. Encouraging information, for example that the end is in sight, will not motivate respondents who have already abandoned the task due to discouraging preliminary information.

Current Study. Conrad, Couper, Tourangeau and Peytchev (2005) explored whether the character of progress feedback affects the impact of progress indicators. In particular, we asked whether encouraging progress feedback might reduce break-offs while discouraging feedback might increase them. Half the respondents were presented with textual progress information (e.g., "17% completed") at the top of each page and half were given no feedback. For those who were given feedback, progress was calculated in one of three ways (see Figure 1).

For one group of respondents (Constant speed), progress increased as a linear function of screens and, therefore, at a constant rate across the questionnaire. For another group (Fast-to-Slow) the rate of progress decreased across the questionnaire, accumulating quickly at first but more slowly toward the end. We produced this pattern of feedback by dividing the log of the current screen by the log of the final screen. For example, after only 9 screens respondents would pass the 50% mark but would need to complete another 36 screens to reach the 90% mark. By 'complete', I mean ad-

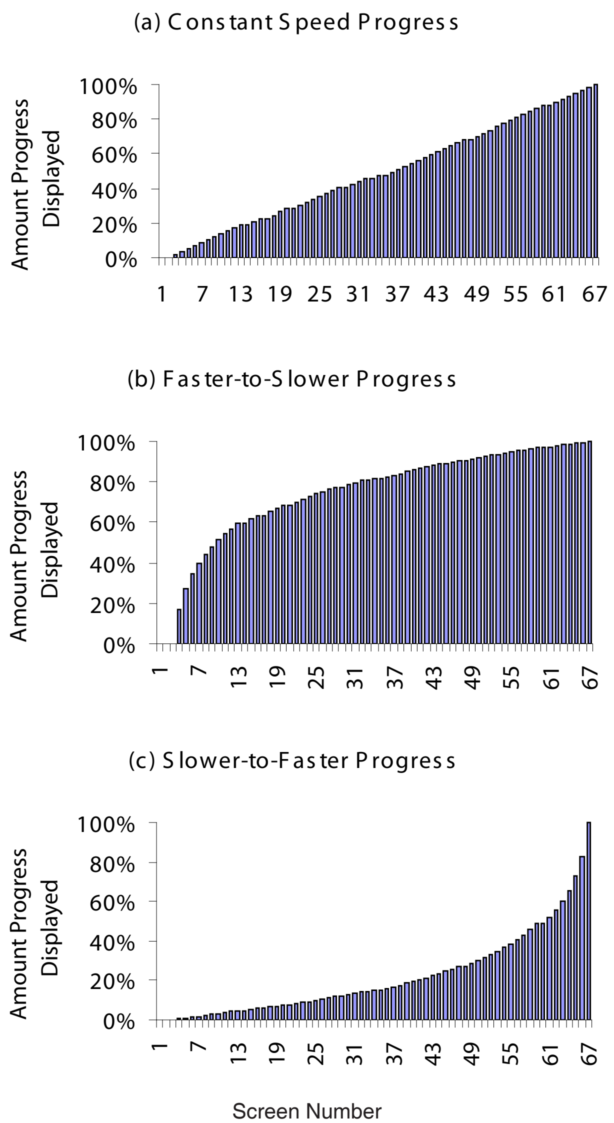

Figure 1. Rates of progress displayed in three progress indicator.

vancing to the next screen, which respondents accomplished by clicking a navigation button. They did not have to enter a response for a given question in order to advance. Thus, the feedback is more encouraging—progress accumulates faster—in the beginning than the end. Finally, for a third group (Slow-to-Fast), the rate of progress increased across the questionnaire, accumulating slowly at first and more quickly toward the end of the questionnaire. We produced this pattern of feedback by dividing the inverse log of the current screen by that of the final screen. For example, to reach the 50% mark, these respondents would need to complete 60 screens but only another 7 screens to surpass the 90% mark. Thus this feedback is discouraging early on—moves slowly—and gets more encouraging toward the end of the questionnaire. We hypothesized that the speed of progress early in the questionnaire would affect overall break-off rates so that when it is slow, break-off rates would be higher than when it is fast.

The questionnaire was comprised of 67 screens, 57 of which presented at least one question. On ten screens no question was presented and these were not considered in the calculation of progress. Respondents moved between all screens, both backward and forward, by clicking a navigation button. The progress indicator was designed so that download and execution time was the same whether or not any feedback was presented.

Respondents from two commercial panels were invited by email to answer a questionnaire administered on the web concerning a variety of "lifestyle" topics. As an incentive to complete the questionnaire, panel members qualified for entry into a sweepstakes in which they could win up to $10,000 by reaching the final screen. A total of 3,179 panel members (8% of all those invited) connected to the survey page into the survey and 2,722 (7% of all those invited) completed it. Thus a total of 457 persons started the survey but did not complete it, representing an overall break-off rate of 14.4%. It is the distribution of these break-offs across the different progress indicator conditions that we are interested in.

As it turned out, respondents were more likely to break-off when the initial feedback was discouraging (Slow-to-Fast) than when it was encouraging (Fast-to-Slow), neutral (Constant Speed), or there was no feedback at all. Apparently, respondents receiving discouraging news at the outset reasonably assumed progress would continue to accrue slowly and inferred that the questionnaire would take more time than it actually did or more time than many were willing to invest. This could suggest that constant speed feedback for a longer questionnaire—which would resemble the initial Slow-to-Fast information for the current questionnaire—is a disincentive to continue. Even for the current, relatively short questionnaire, constant speed feedback did not motivate respondents to complete the questionnaire relative to no progress information. In fact, the proportion of respondents who abandoned the questionnaire with constant speed feedback was higher (though not significantly) than for those receiving no feedback.

Respondents' self-reports measured in a set of debriefing questions at the end of the questionnaire were generally consistent with the break-off data. In particular those who received good news early (and completed the questionnaire) judged the questionnaire to be more interesting than did those in the other progress indicator groups. Apparently people evaluate the content of the questionnaire more favorably when things initially appear to be going well than when they do not. In addition, the same respondents who judged the questionnaire to be more interesting, that is those who received good news first, estimated that it took fewer minutes to complete than respondents in the other progress indicator groups. In fact it took them longer to complete the questionnaire than those in the other groups. Apparently perceived time seems to move more quickly when progress accumulated quickly at the outset than when it accumulated slowly at the outset.

Overall, the debriefing results are striking given that, by the time respondents completed these questions, the rate of progress had largely reversed for the variable speed indicators yet did not seem to reverse respondents' perceptions. It appears, from these data, that respondents form opinions about the task early on and these first impressions are not substantially modified by later evidence.

One implication of the current work is that, if the questionnaire is very long, a garden-variety progress indicator (like our Constant speed progress indicator) might not be very effective in reducing break-offs. As respondents come to realize just how slowly they are making progress they may be at increased risk for breaking off. One could therefore make the case for presenting no progress information. But what about variable speed progress indicators? While we do not necessarily advocate their use because they could be viewed as misleading, in this study, the Fast-to-Slow indicator reduced break-offs and left respondents feeling better about the experience. However, it could be that the subjective experience of progress is not a linear percentage of completed screens but one in which the completion of early screens is weighted more heavily than the completion of later ones. If this is so, then larger increments per screen at the outset may not distort progress at all. Moreover, it may be that respondents seek encouragement most actively at the start of the task when they are least certain about their ability to complete it. This would argue for further exploration of this type of technique.

Use and Non-Use of Definitions

It has long been recognized that many survey concepts are not understood as intended (e.g., Belson, 1981) and it has been demonstrated that when interviewers can define concepts for respondents—despite inevitably different wording for respondents who are given definitions and those who are not—they answer more accurately (Conrad & Schober, 2000; Schober & Conrad, 1997). Rather than giving definitions to those respondents who do not need them, interviewers can provide them when respondents request them or when they believe respondents might otherwise misunderstand (see Schober, Conrad and Fricker, in press). It is a simple matter to make definitions available on the web by linking them to the corresponding words in questions. Respondents need only click on a link to obtain a definition. But making definitions available in this way does not guarantee respondents will use them.

There are at least three obstacles to respondents' use of hyperlinked definitions. First, clicking for a definition may require more effort than respondents are willing to expend. Second, respondents may not realize that definitions might be useful because they might not understand as intended without obtaining a definition. Third, respondents may request a definition and discover that in fact it is not useful, thus inhibiting subsequent requests.

Turning first to effort, one reason respondents might find even a click to involve more effort than they're willing to expend is because it is not necessary to obtain a definition in order to answer the question. For example, getting a definition is not on the "critical path" (Gray, John, & Atwood, 1993). Given that respondents consider their goal to be answering the question—a goal for which they do not consider definitions to be essential—then any action that defers the goal, including a click, is effortful. Of course, getting a definition may be on the critical path if the respondents view their task as answering a question that they have understood as its author intended but it seems unlikely most respondents take this perspective.

In addition, by many analyses of human-computer interaction, a click entails more than just a click. In particular, each overt user action of which clicks are an example, is immediately preceded and followed by mental actions that take time thought. These mental actions include deciding that a definition might actually help achieve the goal or evaluating the results of getting a definition such as the thought "Did that click move me closer to the goal?" The reality of such invisible decision making along side overt user actions has been demonstrated numerous times with the GOMS family techniques developed by Card, Moran and Newell (1983). Examples are offered by Gray, John and Atwood (1993). Alternatives to clicking designed to involve less effort, such as "mouseovers" or "hovering text," in which text appears if the cursor falls within a designated area on the screen, may also be perceived as effortful if their use is not on the critical path because they involve moving the cursor and, in many cases, waiting until the text appears, both of which defer the goal.

The second deterrent to requesting definitions may be that respondents simply do not realize their understanding differs from the surveyors. This is particularly likely when ordinary words are used with non-standard or technical meaning. For example, in the Current Population Survey question, "How many hours per week do you usually work at your job?" the word "usually" is defined as "50% of the time or more, or the most frequent schedule during the past 4 or 5 months" (U.S. Department of Commerce, 1994). "Usually" is such a common term that there is little reason for respondents to expect it has a technical meaning and thus request a definition. A respondent might reasonably assume that the question authors have chosen this word because they believe the respondent will understand it as intended (Clark and Schober, 1992, refer to this as the "presumption of interpretability"). For more technical terms, they might make a similar assumption: the author must believe I am familiar with the word so the meaning that comes to mind must what is intended. (Of course this presumes that something comes to mind.) And in a question of the form "Have you ever …?" they might reason that because the word is very unfamiliar, the answer must be "no": I would know what a "myocardial infarction" is if I had had one.Finally, after obtaining a definition, respondents may realize they would have answered the same way with or without a definition either because they had already understood the term as intended or because the definition contains material irrelevant to their circumstances.

For example the Census definition of "residence" goes into detail about borders and children in the armed forces, when it is possible these will not apply to a particular respondent. Having concluded that the available definitions aren't helpful, it is unlikely that respondents will request more of them. Landauer (1995) used the phrase "creeping featurism" to describe the phenomenon of including features in software because designers believe they will make the product more competitive but not because they are helpful to users. He describes a survey of one software company's user base which found that fewer than one third of the available features were ever used; presumably many of those used were used only once as we would expect to be the case for uninformative definitions.

Current experiment. Conrad, Couper, Tourangeau and Peytchev (2006) asked whether respondents' requests for definitions is affected by the ease of obtaining definitions, respondents' likely awareness that definitions might be helpful, and the apparent usefulness of definitions. Respondents answered four questions arrayed in a grid with concepts as the rows and response options as the columns (see Figure 2a): "The following questions concern the amount of food and nutrients that you typically consume. If you are uncertain about the meaning of a particular food or nutrient, please click on the word to obtain a definition. How much of the following items do you typically consume?"

A given respondent was able to obtain definitions with one of three user interfaces, designed to vary the required number of clicks and therefore effort. The particular interface presented to a respondent was determined at random. In the "one-click" interface, respondents clicked on a highlighted word and a definition appeared (Figure 2b). This should not be confused with double clicking. In the "two-click" interface, clicking on the definition produced a list of all terms for which definitions were available and respondents needed to then click on the relevant term (Figure 2c). With this interface, each click produced a distinct system action beginning with a list of terms and then a definition for one term. Finally, in the "click-and-scroll" interface, clicking displayed the complete list of definitions (essentially a

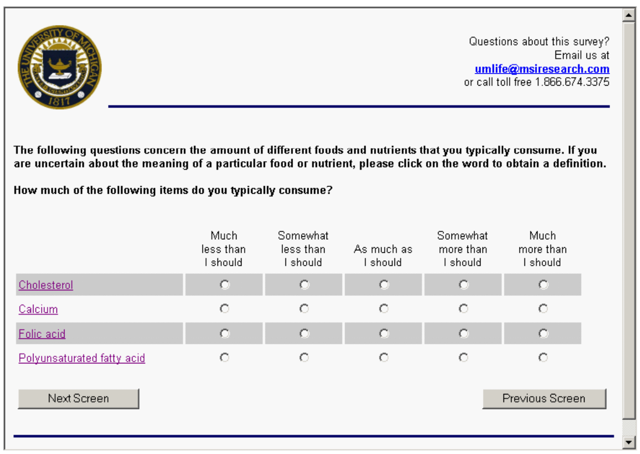

Figure 2a. Item for definitions available.

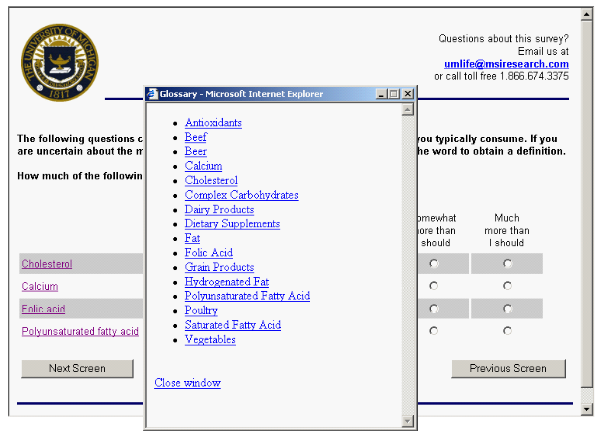

Figure 2c. List of terms for which definitions available made available by clicking on term in grid (Figure 2a) for two-click interface.

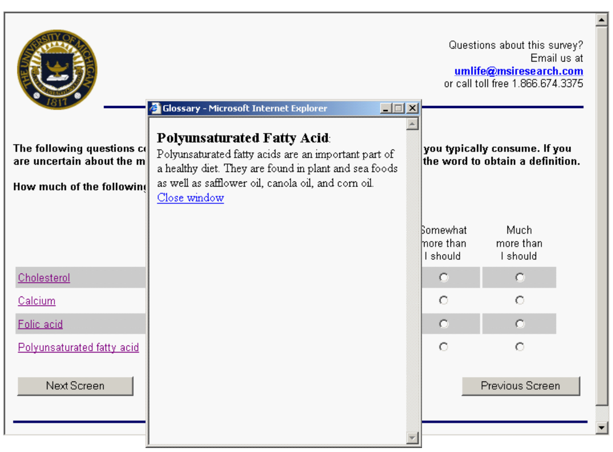

Figure 2b. Definition made available by clicking on term in grid (Figure 2a) for one-click interface or on term in list (Figure 2c) for two-click interface.

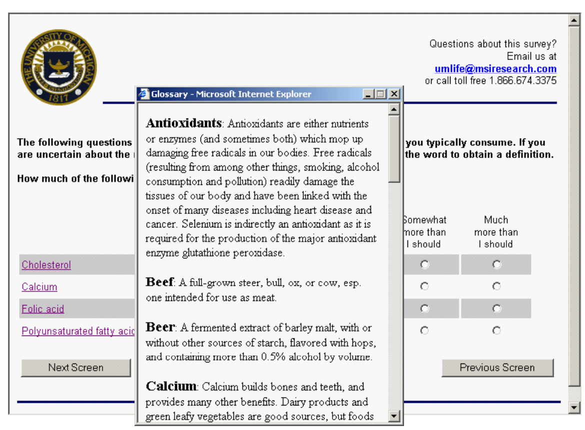

Figure 2d. Glossary (all definitions for all terms) made available by clicking on term in grid (Figure 2a) in click-and-scroll interface. If definition is not visible, respondent must scroll to it by using scroll bar at right.

glossary) in a text window so that if the definition of interest was not visible, the respondent needed navigate to it by clicking in the scroll bar (Figure 2d). Note that the number of clicks required under the three interfaces was something of a surrogate for the total amount of effort: when more than one click was required, more reading and decision making was required as well—much as is assumed from the GOMS perspective mentioned earlier.

The group of four questions presented to any one respondent concerned either technical (e.g., "saturated fatty acid") or ordinary (e.g., "vegetables") terms and the definitions were either useful or not useful. Definitions that were not useful lacked any information that would be likely to affect respondents' answers (e.g., "In saturated fatty acid, the carbon atoms are bonded with single bonds; they share one set of electrons. Saturated fatty acids are mostly found in animal products.") whereas definitions that were useful contained counterintuitive or surprising information (e.g., "In general, vegetables include the edible stems, leaves, and roots of a plant. Potatoes, including French fries, mashed potatoes, and potato chips are vegetables"). We expected respondents to recognize the need for definitions of technical terms and request them more often than for ordinary terms and we expected an initial request for a useful definition to lead to more subsequent requests than if the initial definition was not useful. For a given respondent all definitions were either helpful or not helpful. Thus the design crossed three levels of difficulty (one-click, two-clicks, click-and-scroll) with two types of concepts (technical or ordinary) and two types of definitions (useful or not useful).

Respondents from two commercial panels were invited by email to answer a questionnaire administered on the web concerning a variety of "lifestyle" topics. The panels and the general topic of the current survey were the same as the progress indicator. 2871 respondents completed the questionnaire for a response rate of 18%. Again, our goal was random assignment rather than representativeness.

Requests for definitions were rare overall: only 17.4% of respondents who finished the questionnaires (13.8% of those who answered the questions with definitions) ever clicked. This suggests that many misconceptions may go uncorrected despite the availability of clarification features. It could be that something as simple as a stronger instruction to use definitions could increase the number of requests, but it may also be the case that many respondents are unwilling to stray from the critical path or do more than the minimum necessary to complete the task.

When examining data from those respondents who requested at least one definition, it is apparent that the number of requests is quite sensitive to the amount of effort (number of clicks) involved. When only one click was required, respondents obtained more than 2.5 out of 4 definitions but when two or more clicks were required, they obtained closer to 1.5 out of 4 definitions, which is a reliable difference. Those respondents who had to click twice to get a definition abandoned the request after the first click 36% of the time (383 first but only 246 second clicks) providing additional evidence that effort (2 clicks versus 1) matters.

Respondents seemed to recognize the potential value of a definition more often for technical than ordinary terms: 89% of definitions requested concerned technical terms. But it is really for ordinary terms that may be used in non-standard ways that clarification is especially important. As it turns out, people request more definitions of ordinary terms when the definitions are useful, presumably because they come to realize that despite being familiar these words may mean something other than the respondent initially assumes. However, the impact of useful definitions is only observed when respondents can obtain a definition with a single click. If more than one click is required, respondents request definitions infrequently and equally often whether definitions are useful or not. What this tells us is that for an "off-path" activity like requesting definitions, effort must be extremely low. If more than one click is required, there is little that will convince respondents to request a definition.

These results almost certainly extend beyond on-line definitions and even beyond web surveys to web use in general. People seem to be impatient when they use the web, perhaps because of the vast amount of information that is available through very minor actions such as pressing a mouse button. This introduces yet another reason why the web in general and web surveys in particular, should not be treated as if they are paper.

Diagnosing Respondent Uncertainty

Respondents in the previous study requested definitions relatively rarely. While they requested some definitions more frequently when easy to obtain, the overall rates were still low. Infrequent requests for definitions could reflect respondents' lack of awareness that they misunderstand a term or their reluctance to request definitions because it involves additional clicks and reading. If so, an alternative approach to the design of web questionnaires could involve programming the questionnaire so that it can volunteer definitions when respondents seem uncertain or confused.

Current experiment. We (Conrad, Schober and Coiner, in press) have explored this approach in a laboratory study in which the survey system could sometimes offer respondents definitions if they were inactive for more than a certain amount of time. Inactivity was treated as an indication that respondents were confused or uncertain or at some kind of impasse. The basic idea was to see if (1) providing definitions when respondents seemed to need them but did not ask for them improved their understanding and response accuracy above the levels observed when they were required to request definitions by clicking, and (2) whether the benefits of this approach are greater if inactivity is modeled differently for different groups of respondents. Kay (1995) argues for the benefits of group-based or stereotypic user models.

Our groups were based on the respondents' age—one group was young and one older. Respondents' age has been shown to affect the size of question and response order effects, largely because working memory declines with age (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 respondents' response times to be longer than younger respondents' 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.

We contrasted five user interfaces in the laboratory. In the first there was no clarification available to users. In the second, the clarification was available if the user requested it by clicking—we refer to this as "respondent-initiated" clarification. In the third and fourth, clarification was "mixed-initiative," it could be either respondent- or system-initiated by which we mean the system could volunteer a definition when the respondent was inactive for more than a certain amount of time. That "certain amount of time" was modeled differently in the third and fourth user interfaces. In the third interface, the system-initiated clarification was based on the same inactivity threshold for all respondents or a generic respondent model; in the fourth interface the threshold was set differently for old and young respondents, or a group-based respondent model. In the fifth interface, the definition always appeared with the survey question.

All respondents answered the same 10 questions about housing and purchases from two ongoing government surveys (used by Conrad & Schober, 2000) based on fictional scenarios for which we knew the correct answer, enabling us to measure response accuracy. Half of the scenarios 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. The other half were designed to be hard to answer correctly without access to the official definition. We call these complicated 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.

In the conditions where they were able to request clarification, respondents clicked on a hyperlinked term or phrase and the system displayed the corresponding definition. When the system initiated the clarification, the definition simply appeared after the appropriate threshold accompanied by a brief, computer-generated tone to attract the respondent's attention.

Through a newspaper advertisement and fliers at senior centers, we recruited 114 paid participants. 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

The results support the idea of programming web-based questionnaires to volunteer clarification when respondents seem to need it, and to interpret evidence of needing clarification differently depending on respondents' age. All respondents were quite accurate when answering on the basis of straightforward scenarios (95% of questions answered correctly); for complicated scenarios, accuracy increased linearly across the five clarification groups: no clarification (24%), respondent-initiated clarification (35%), mixed initiative clarification, generic model (48%), mixed initiative clarification, group-based model (58%), definitions always (70%).

Respondents' preferences were not directly related to their response accuracy: 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. Recall that respondents were least accurate when the system never initiated clarification so rather than most preferring the interfaces that promoted accuracy, the respondents seemed to prefer the interfaces that allowed them to think about their answers without interruption by an unsolicited definition. The distaste for system-initiated clarification was most pronounced among older respondents. The respondents also did like having definitions always present, even though these led to the highest levels of accuracy.

This suggests that some aspects of designing system-initiated clarification still need to be worked out. It may be a matter of fine-tuning the inactivity thresholds so that, for example, system-initiated clarification does not interrupt respondents but still offers clarification before they respond. But it may also be that there is no single threshold that is appropriate for an entire group. In this case, individualized thresholds, possibly based on response times to a small battery of calibration questions, would lead to accuracy on the level of providing clarification all the time but with higher satisfaction. In any case, the accuracy results suggest that enabling web-based questionnaires to offer clarification can improve respondent understanding of questions beyond the levels of ordinary self-administration.

Conclusions

We have considered three interactive features of web surveys that can be implemented with available technology requiring relatively simple programming. One can imagine other features to help improve the interaction that are based on more experimental technology. For example, the questionnaire could make use of natural language dialogue allowing the respondent to type open-ended questions into the interface and responding by generating informative text. For example, instead of presenting multi-paragraph definitions, the system could tailor its output—probably text—to the respondent's query about a specific situation. Another technology that could be useful is speech recognition. The respondent could speak to the system, for example requesting progress information, while thinking about the answer to the current question. Speaking is a highly practiced skill and one that people can use while performing other tasks. This could make it easier for a respondent to invoke features that might otherwise require too much effort. And speech contains many more cues about the speakers' mental and emotional state than does textual or mouse input, thus allowing the system to better diagnose respondents' uncertainty and take appropriate actions.

A technology that very much blurs the distinction between self- and interviewer-administration is animated agents or avatars. Introducing a virtual interviewer into the web questionnaire may help establish a social connection for example, in providing encouraging messages to respondents in order to keep them engaged or reminding respondents that the system has the human-like ability to provide clarification thus promoting requests. But one can imagine than a virtual interviewer may hurt when respondents are asked sensitive questions because the agent might trigger self-presentation concerns in the way human interviewers do, undermining the clearest benefit of self-administration. While not all of these or future technologies will necessarily be useful in surveys on the web, they will be available to designers who will need to weigh and consider their use.

References

Belson, W. A. (1981). *The design and understanding of survey questions*. Aldershot, UK: Gower.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum.

Clark, H. H., & Schober, M. F. (1992). 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.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*, 461-494.

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

Conrad, F., Couper, M., Tourangeau, R., & Peytchev, A. (2006). Use and non-use of clarification features in web surveys. *Journal of Official Statistics*, *22*, 245-269.

Conrad, F. G., Couper, M. P., Tourangeau, R., & Peytchev, A. (2005). *Effectiveness of progress indicators in Web surveys: First impressions matter*. Proceedings of SIGCHI 2005: Human Factors in Computing Systems, Portland, OR.

Couper, M., Traugott, M., & Lamias, M. (2001). Web survey design and administration. *Public Opinion Quarterly*, *65*, 230-253.

Crawford, S. D., Couper, M. P., & Lamias, M. J. (2001) Web surveys: Perception of burden. *Social Science Computer Review*, *19*, 146-162.

Gray, W. D., John, B. E., & Atwood, M. E. (1993). Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world performance. *Human-Computer Interaction*, *8*(3), 237-309.

Hert, C. A., & Marchionini, G. (1997). Seeking statistical information in Federal websites: Users, tasks, strategies, and design recommendations. Retrieved from http://www.ils.unc.edu/~march/blsreport/mainbls.html

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

Kiousis, S. (2002). Interactivity: A concept explication. *New Media and Society*, *4*, 355-383.

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*, Philadelphia: Taylor and Francis.

Landauer, T. K. (1995). *The trouble with computers: Usefulness, usability and productivity*. Cambridge, MA: MIT Press.

McMillan, S. J., & Hwang, J. (2002). Measures of perceived interactivity: An exploration of the role of direction of communication, user control, and time in shaping perceptions of interactivity. *Journal of Advertising*, *31*, 29-42.

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

Schober, M. F., Conrad, F. G., & Fricker, S. S. (In Press). Misunderstanding standardized language. *Applied Cognitive Psychology*.

U.S. Department of Commerce (1994). *Current Population Survey Interviewing Manual* (CPS-250). Washington, DC: Bureau of the Census.