# Interactive Features in Web Surveys

**Frederick G. Conrad**
**Mick P. Couper**
**Roger Tourangeau**

*University of Michigan, Institute for Social Research and*
*University of Maryland, Joint Program on Survey Methodology*

Joint Meetings of the American Statistical Association
San Francisco, CA
August 6, 2003

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

## 1. Interactivity in web surveys

The web is like many other media but not the same as any of them. It is similar to print in that the content is (still) primarily textual – witness the dominant "page" metaphor. But it is different from print in that hyperlinks make pages dynamic in a way that paper is not. The web has some of the character of conventional desktop software in that the user acts by clicking and typing but it is different in that web pages are largely static, hyperlinks and Java script not withstanding. Like television, the web can display video and animated content and like digital cable with hundreds of channels the web provides niche content. But unlike TV, web content is available on demand rather than when network executives make it available. In essence the web is a diverse and eclectic medium that makes it possible to deliver almost any kind of content through any kind of user interface.

This diversity becomes a double edged sword for content designers because it means they must choose which features to implement and which features to exclude, e.g. not using audio should be the consequence of a thoughtful design decision. Such decisions are essential when content that was previously delivered through another medium is adapted to the web: should the new design emulate the original medium or should it exploit features made available by the web? Designers of questionnaires for web administration are directly confronted with this kind of decision because the principles and guidelines for creating self-administered instruments have been developed for printed questions on paper. Should web questionnaires emulate paper questionnaires or incorporate features made available by the web? Couper, Tourangeau & Conrad (this session) explore the pros and cons of visual features made available by the web, e.g. images. Our focus here will be on interactive features.

A hallmark of interactivity in web surveys is the shift of control from the respondent to the system (questionnaire) and back to the respondent, repeatedly throughout the session. This is analogous to the exchange of the floor in dialogue between people. By this view the respondent takes an action such as clicking a "next page" button, and in response, the system records the submitted answers, presents the next page and returns control to the respondent. In addition, the actions of an interactive questionnaire serve as feedback to respondents that their actions have been recognized and accepted (or not, if, for example, the system's action is to prompt for a missing response or flag one that is out of range). Interactive designs require the system to react to the respondent's actions in real time – however that is perceived – much as in human dialogue. If the system does not react immediately, respondents are sure to break off at high rates in part because they do not feel engaged in an interaction. (More extensive discussions of the interactivity concept along these lines are provided by. Kiousis, 2002 and McMillan & Hwang, 2002)

Of course interactivity comes about only if it is "designed into" the questionnaire. A questionnaire is not interactive if, for example, it designed as a single scrollable form in which the respondent answers all questions before submitting her answers, i.e. 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 initiating respondent action is clicking the "Next Page" button. In response, the system updates the progress information ("percent completed") displayed on the subsequent page. Based in part on this information, the respondent either submits the next page or breaks-off, i.e. terminates his 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 triggering action is clicking a hyperlinked word or phrase in the question; in response, the system displays a definition of the word or phrase; if this proves useful, it could affect 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 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 affect 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.

## 2. Effectiveness of Progress Indicators

*Background.* 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 (i.e. non-interactive) web questionnaires, where the size and location of the scroll bar convey progress information. But in more interactive designs, for example in which one question is presented per page, there is no default progress information, i.e. if there is a scroll bar – and there most likely is not – it only reflects position on the current page. However, the display of progress information can be designed into the questionnaire –typically as either graphical or textual progress indicators – but designers need to consider the consequences of providing such feedback to respondents. 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.

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 time according to the progress indicator. In general, discouraging information, e.g., 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, e.g., that the end is in sight, will not motivate respondents who have already abandoned the task due to discouraging preliminary information.

*Current Experiment.* Conrad, Couper, Tourangeau & Peytchev (2003) explored how the quality (encouraging, discouraging) and timing (early versus late) of information displayed in progress indicators affects the completion of an on-line questionnaire. 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 39,217 email invitations to participate in the current questionnaire were sent, in response to which 3,179 persons (8%) logged into the survey and 2,722 (7%) 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%. Respondents were not given any information in the invitation letter or in the introductory pages of the questionnaire about how much time would likely be required to complete it.

A textual progress indicator (e.g. "17% completed") appeared at the top of each page for half of the respondents selected at random. The other half of the respondents was not given any feedback about their progress. The progress indicator was designed so that download and execution time was the same whether or not any feedback was presented.

Progress was calculated in one of three ways that affected the speed with which it accumulated (see Figure 1). In all cases, progress reflected the percentage of screens, including the current one, displayed up to that point. What differed was how the percent completed was derived. For one type of progress indicator (Constant speed), relevant screens were numbered and progress was based simply on the current screen number divided by 57 (presented as a percent). Thus progress increased as a linear function of screens and, therefore, at a constant rate across the questionnaire. For another type of progress indicator (Faster-to-Slower) the rate of progress decelerated 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 (expressed as a percent). For example, after only 9 screens respondents would pass the 50% mark but would need to complete[1] another 36 screens to reach the 90% mark. Thus, the feedback is more encouraging – progress accumulates faster – in the beginning than the end. Finally, for a third group (Slower-to-Faster), the rate of progress accelerated 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.

We further hypothesized that break-offs would increase on relatively difficult items but the effect would be reduced by early, encouraging information. To test this we included an item designed to be difficult in the middle of the questionnaire. The item was difficult in that it required multiple responses and we created two versions, one intended to be even more difficult than the other in that it required open as opposed to closed responses. The prediction was that break-offs would increase overall for this item and more for the difficult form except for respondents who had received good news initially for whom we expected no difference between the forms. This item concerned automobile ownership so that the two main factors in the experiment were type of progress indicator (4 levels) and form of automobile ownership question (2 levels).

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.

---

[1] By "complete" we mean advance to the next screen, which respondent accomplished by clicking a navigation button. They did not have to enter a response for a given question in order to advance.

Break-off rates varied with the type of progress indicator, $F(3, 3176) = 10.62$, $p < .001$ (see Table 1, Row 1)[2]. Respondents were more likely to break-off when the initial feedback was discouraging (Slower-to-Faster) than when it was encouraging (Faster-to-Slower), neutral (Constant Speed) or there was no feedback at all, comparison of Slower-to-Faster to the other three progress indicator groups, $t(3173) = -5.312$, $p < .001$. 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, i.e., more time than many were willing to invest. This could suggest that constant speed feedback for a longer questionnaire – which would resemble the initial Slower-to-Faster 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.

| Progress Indicator | None | Constant Speed | Slower-to-Faster | Faster-to-Slower |
|---|---|---|---|---|
| % Break-offs | 12.7 | 14.4 | 21.8 | 11.3 |
| How interesting was this survey[3]? | 4.09 | 4.03 | 4.07 | 4.27 |
| How many minutes do you think it took you to complete the survey? | 14.43 | 13.97 | 15.38 | 13.47 |

Table 1. Break-off rates and mean responses to debriefing questions.

If good news up front does indeed encourage respondents to stick to the task, then they should be more likely to continue in the face of a difficult question than those who have not been so encouraged. We can test this by looking at break-offs that occur on the two forms of the automobile ownership item. The number of break-offs for this item overall was larger than for all others (aside from the first screen), though not large overall, so the difficulty of both forms increased break-offs as expected. Moreover, a larger number of respondents (.021 of all respondents) broke off on the screen presenting the more difficult form than on the screen presenting the easier form (.004 of all respondents), $F(1,2907)=11.76$, $p =.001$. Consistent with the idea that good news early on may lead respondents to persevere, the difference between the forms depended on the type of progress indicator, form x progress indicator interaction $F(3,2907)=2.49$, $p =.058$. For the Faster-to-Slower (good news first) group the difference between forms was virtually eliminated (and not reliable): there were no break-offs on the difficult form and only one on the easier form. However, the form difference was reliable for the groups that received no feedback ($z=3.40$, $p < .001$) and uniform speed feedback ($z=2.62$, $p < .01$). The fast early progress apparently continues to help respondents persevere even though the rate at which progress accumulates has slowed down substantially by the time the automobile ownership question is presented. For these respondents, completing the first screen

[2] An alpha level of .05 is used for all statistical tests.

[3] 1=Not at all interesting; 6=Extremely interesting

increased progress by 18% (from 0% to 18%) but completing the 32nd screen (right before the automobile ownership screen) increased progress by only by 1% (from 80% to 81%).

Respondents' self-reports measured in a set of debriefing questions were generally consistent with the break-off data. In particular, the type of progress indicator affected how interesting respondents found the task, $F$ (3, 2709) = 3.95, $p < .01$ (see Table 1, row 2). Those who received good news early on judged the questionnaire to be more interesting than did those in the other progress indicator groups, comparison of Faster-to-Slower to the other three progress indicator groups $t(1,2709)=125.25$, $p < .001$. 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 type of progress indicator affected respondents' judgment of the length of the task, $F(3, 2709) = 4.35$, $p < .01$ (see Table 1, row 3). The same respondents who judged the questionnaire be more interesting, i.e. those who received good news first, estimated that it took fewer minutes to complete than respondents in the other progress indicator groups, comparison of Faster-to-Slower to other three progress indicator groups, $t(1,2709)=3.12$, $p < 001$. In fact there were no differences in actual duration, $F(3,2718)= 0.59$, n.s., so respondents apparently perceived time 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.

In general, the effect of the progress indicator might have been even stronger with respondents who were chosen at random and generally less motivated than were ours who were self-selected and had previously agreed to participate in surveys on the Internet. It is also possible the effect of the progress indicator would have had a different form if respondents had access to other relevant information such as the likely duration of the questionnaire. Crawford, et al. (2001) report an interaction between the likely duration indicated in an invitation letter and the presence of a progress indicator. The point is that people's perception of duration and effort, their sense of boredom, and their mood, are sensitive to the information available during a task

One implication of the current work is that, if the questionnaire is very long, a regular progress indicator may not be very effective in reducing break-offs. One could therefore make the case for presenting no progress information. But what about variable speed progress indicators? While we do not necessarily advocating their use because they could be viewed as misleading– in this study, the faster-to-slower 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.

**3. Use and Non-Use of Definitions**

*Background.* 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 & 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, i.e. that 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, i.e. 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, e.g. deciding that a definition might actually help achieve the goal or evaluating the results of getting a definition, i.e. "did it 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 & Newell, 1983; see for example Gray, John & Atwood, 1993). Alternatives to clicking designed to involve less effort, e.g. "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 & 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 no 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 & Baker (2003) tested these ideas in an experiment that the ease of obtaining definitions, respondents' likely awareness that definitions might be helpful, and their usefulness. Respondents from two commercial panel were invited by email to answer a questionnaire administered on the web concerning a variety of "lifestyle" topics[4]. 2871 respondents completed the questionnaire for a response rate of 18%. (Our goal was randomization rather than representativeness.) 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?"

[4] While the panels and the general topic of the current survey were the same as in the progress indicator study, the current experiment was carried out in a separate survey separated in time by almost a year.

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). 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). Finally, in the "click-and-scroll" interface, clicking displayed the complete list of definitions 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.

Definitions were actually available for 16 highlighted terms, though only one group of four questions was presented to a given respondent. The assignment of respondent to group of questions was determined randomly. The group of four questions concerned either technical (e.g. "saturated fatty acid") or ordinary (e.g. "vegetables") concepts and the definitions were either not 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).

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, i.e. to do more than the minimum necessary to complete the task.

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 (see Figure 3). 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, $F(2,452)=9.71$, $p < .001$. 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. However, requesting a definition for a technical term (2.32 request per respondent on average) was no more likely to lead to follow-up requests than was a request for the definition of an ordinary term (1.54 request per respondent on average), $F(1,452)= 2.51$, , n.s.). Presumably this is because definitions for technical and ordinary terms were useful equally often in this experiment.

For ordinary terms– the ones people are more likely to assume they understand without a definition– getting a useful definition leads to more follow-up requests than does getting a non-useful definitions. The pattern is reversed for technical definitions – where people were more likely to believe they might need a definition in the first place. It could be that for a complex technical concept, "useful" information, i.e. information that is surprising or counterintuitive, is more than people can assimilate so they do not request as many subsequent definitions as when the definitions are not useful, i.e. intuitive or not surprising, interaction of term x usefulness $F(1,452)= 3.79$, $p = .052$. This pattern is actually moderated by the number of clicks needed to obtain a definition, interaction of interface x term x usefulness interaction, $F(2, 42) = 3.49$, $p =.033$. The effect is seen primarily on requests for definitions of ordinary terms (Figure 4). In particular, when only one click is required, useful definitions are requested 3.67 out of 4 times on average and only 1.39 times out of 4 on average for non-useful definitions. However, when more than one click is required, respondents rarely request more than one definition whether they useful or not. The main point is that if more than 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, e.g. pressing a mouse button. For example, Hert and Marchionini (1997, section 4.3.2) observed that a substantial proportion of visits to a Federal web site involved just one page (the proportion of one-page sessions ranged from 22%-52% depending on the content area of the site) suggesting that if the information users sought was not immediately available they quickly went elsewhere. It could be that this is an artifact of who adopts new technologies earlier and later, i.e. that the current population of web users are just impatient compared to non-users, but we suspect this is not the case and that as web access becomes increasingly universal, new users will exhibit similar impatience. This introduces yet another reason why the web should not be treated as if it were paper.

**4. Diagnosing Respondent Uncertainty**

Respondents in the previous study requested definitions relatively rarely. While they requested some definitions more frequently when they were easy to obtain, the overall rates were still low. To the extent that this reflects a lack of awareness that definitions might be helpful, an alternative approach is to design the web questionnaire to voluntarily provide definitions when respondents seem uncertain or confused. We (Coiner, Schober, Conrad & Ehlen, 2002) have explored this approach in a laboratory study in which respondents answered questions on the basis of fictional scenarios allowing us to determine the accuracy of their answers.

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. One approach emphasizes giving users control (e.g. Shneiderman, 1997), where users can adjust the interface as desired; the other emphasizes user modeling, where interfaces automatically adapt to different users (Maes, 1994). In this study, we contrasted typical web survey interfaces (usually standardized for everyone) with interfaces that provide increased user control and interfaces built around user models (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. We formed our groups based on age. 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 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.

We contrasted five user interfaces in the laboratory. In the first there was no clarification available to users. The clarification was available if the user requested it by clicking – we refer to this as "user-initiated" clarification. 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 different for different groups of respondents. 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). All respondents answered the questions on the basis of fictional scenarios for which we knew the correct

answer, enabling us to measure response accuracy. 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. Questions were presented to laboratory respondents on a computer using a web-browser interface. 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 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.

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 40th 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 40th percentile response time for complicated mappings but computed separately for old and young users.

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. Ethnicities, educational backgrounds and experience with computers were roughly balanced across age groups.

As can be seen in Figure 5, 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 of scenario type x clarification interface, $F(4, 104) = 16.58$, $p<.001$. Accuracy increased linearly across the five clarification 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 (35% of questions answered correctly) but

still poorer than when the system also clarified concepts. 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 still[5] 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). (The advantage for group-based modeling was largely due to the advantage for young respondents [see Coiner, et al., 2002, for details].) Accuracy was best of all when respondents received definitions along with the questions (70% of questions answered correctly).

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.

As another measure of user satisfaction, 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 those in the no clarification group (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. The preference by some residents for human interviewers over computers was summarized in the following comment collected in the debriefing questionnaire: "Even though the questions were simple, it is nice to have some human contact and not this loud machine."

While not a universal sentiment, this does indicate that 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. Whether the model is group- or individual-based, system-initiated clarification blurs the distinction between a self-administered questionnaire and one administered by an agent other than the respondent—more reason to treat the web as a unique medium that bears only a superficial similarity to paper.

---

5

## Conclusions

The three types of interactivity that we have examined can be arrayed along a dimension of user control. With hyperlinked definitions, it is up to respondents (users) whether or not to request clarification – and as we saw, they do not often take advantage of the option. In contrast, the kind of progress indicators and system-initiated clarification we examined are under the system's control. Progress is displayed whether users want it or not and system-initiated clarification appears even if some users find it disruptive. It may well be that web users prefer to control the features of the interaction, though system control of some features may make them more effective. This is a trade-off designers will need to confront.

We have considered three interactive features of web surveys that can be implemented with available technology require 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, e.g. 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. Finally animated agents – so-called avatars – may help establish a social connection for example, in providing encouraging messages to respondents in order to keep them engaged. While not all of these or future technologies will necessarily be useful in surveys on the web, they will be available to designers who should 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, Inc. Publishers.

Clark, H. H. & Schober, M. F. (1992). Asking questions and influencing answers. In Tanur, J. M. (Ed.), *Questions about Questions: Inquiries into the Cognitive Bases of Surveys.* New York: Russell Sage Foundation, pp.15-48.

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. & Baker, R. (2003). Use and non-use of clarification features in web surveys. Paper presented at 58th Annual Conference of the American Association of Public Opinion Research, Nashville, TN.

Conrad, F., Couper, M., Tourangeau, R. & Peytchev, A. (2003). Effectiveness of Progress Indicators in Web Surveys: It's What's Up Front that Counts. *Proceedings of the Fifth International ASC conference*. Chesham, UK: Association for Survey Computing.

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. http://www.ils.unc.edu/~march/blsreport/mainbls.html

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.

Kiousis, S. (2002). Interactivity: A concept explication. *New Media & 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.* Taylor & Francis, Philadelphia.

Landauer, T. K. (1995). *The Trouble with Computers: Usefulness, Usability and Productivity.* Cambridge, MA: MIT Press.

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

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. and Fricker, S.S. (in press). Misunderstanding standardized language. *Applied Cognitive Psychology.*

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

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

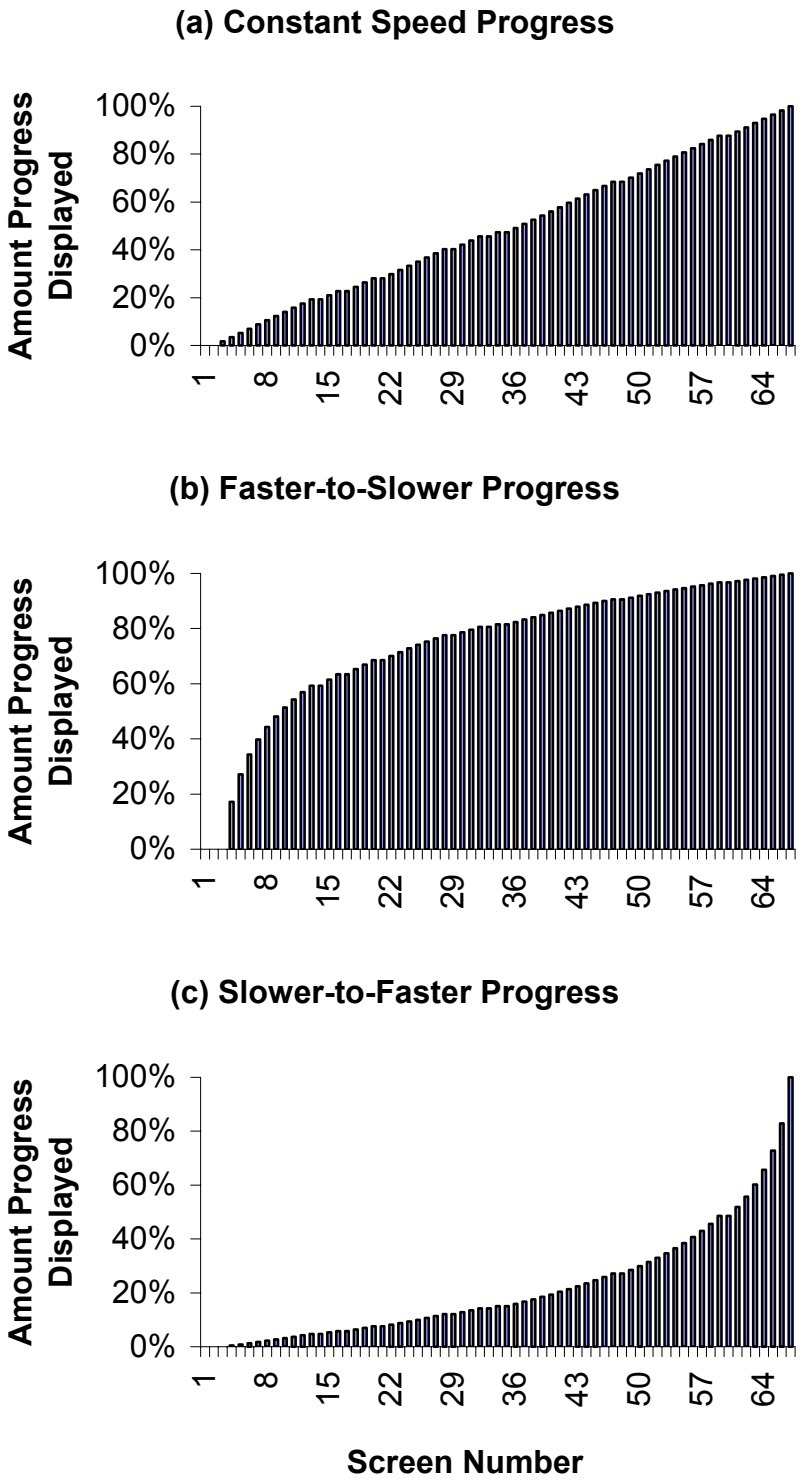

Figure 1. Rates of progress displayed in three progress indicators.

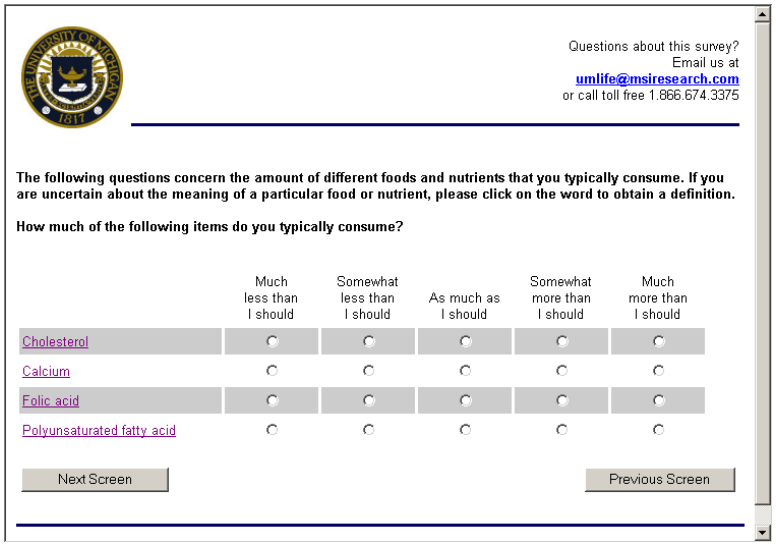


Figure 2a. Item for definitions available.

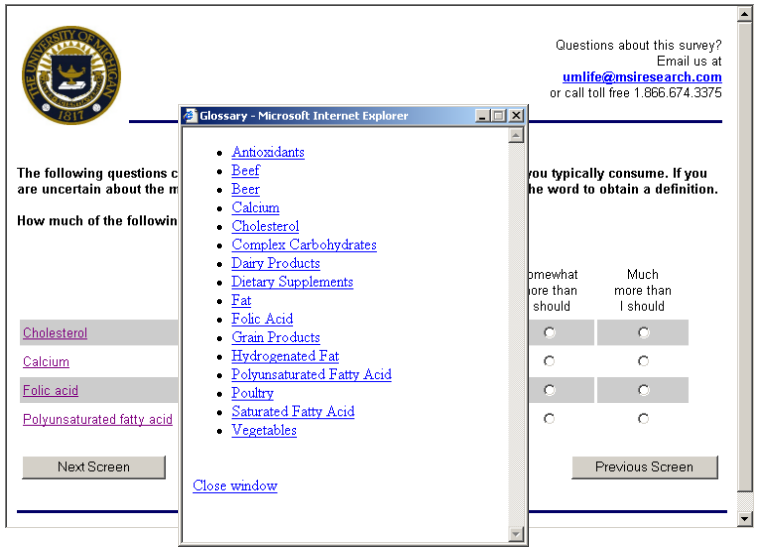


Figure 2b. List of terms for which definitions available made available by clicking on term in grid (Figure 2a) for two-click interface

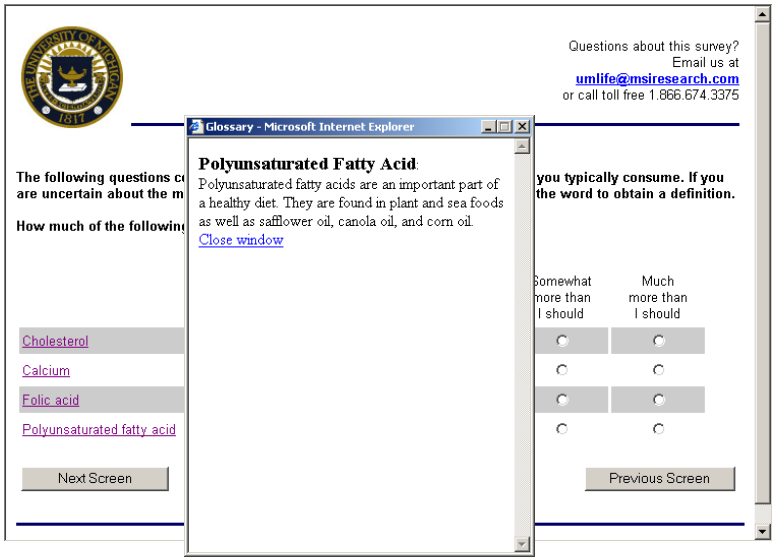


Figure 2b. Definition made available by clicking on term in grid (Figure 2a) for one-click interface or on term in list (Figure 2c) for two-click interface

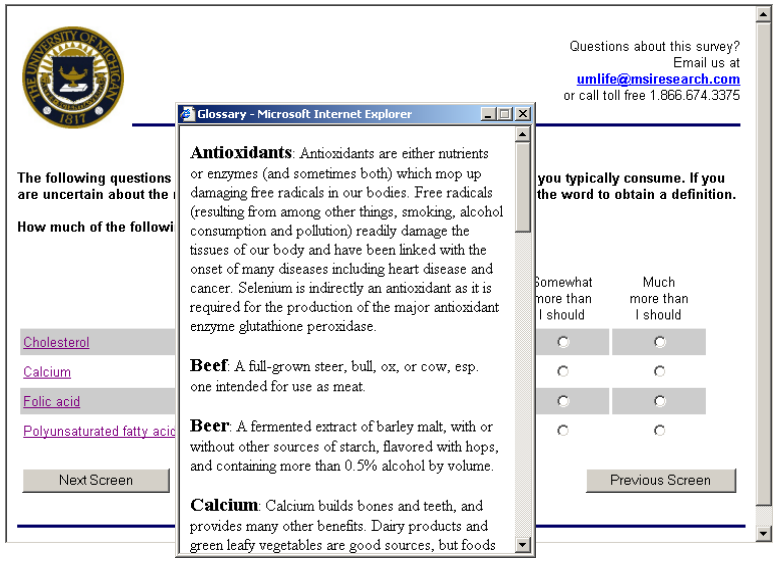


Figure 2d. Glossary (all definitions for all terms) made available by clicking on term in grid (Figure 2a) in click-and-scroll interface. If definition is not visible, respondent must scroll to it by using scroll bar at right.

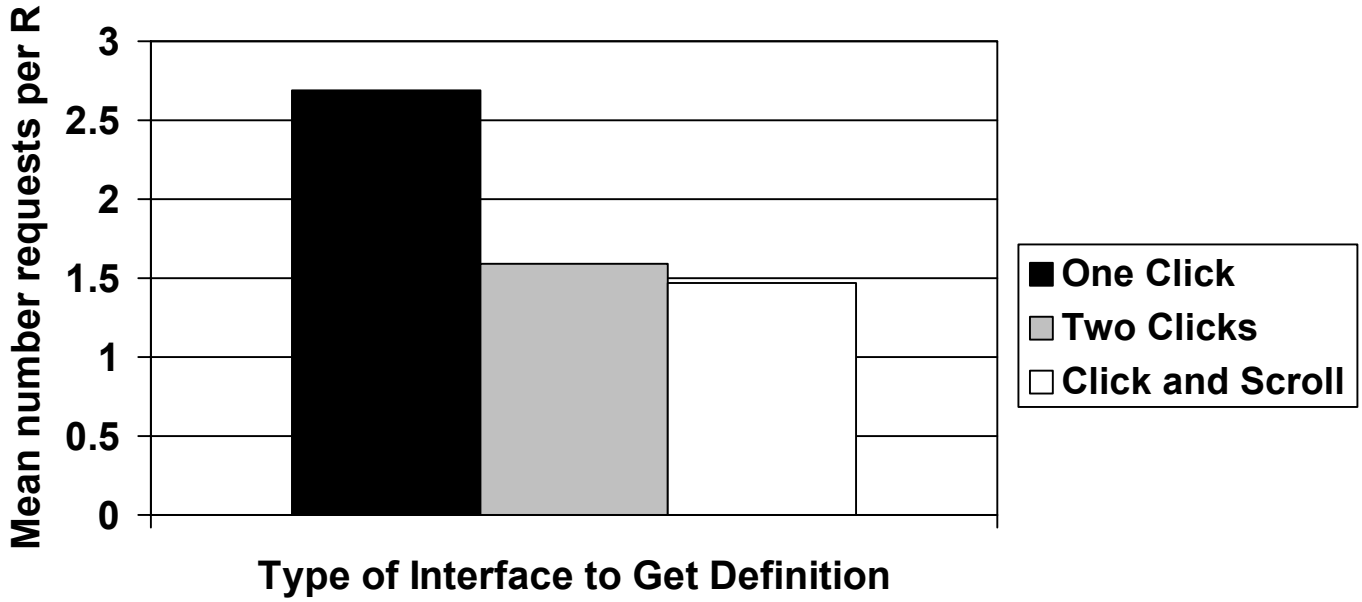


Figure 3. Mean number of requests for clarification (for respondents who requested any definitions) for three user interfaces

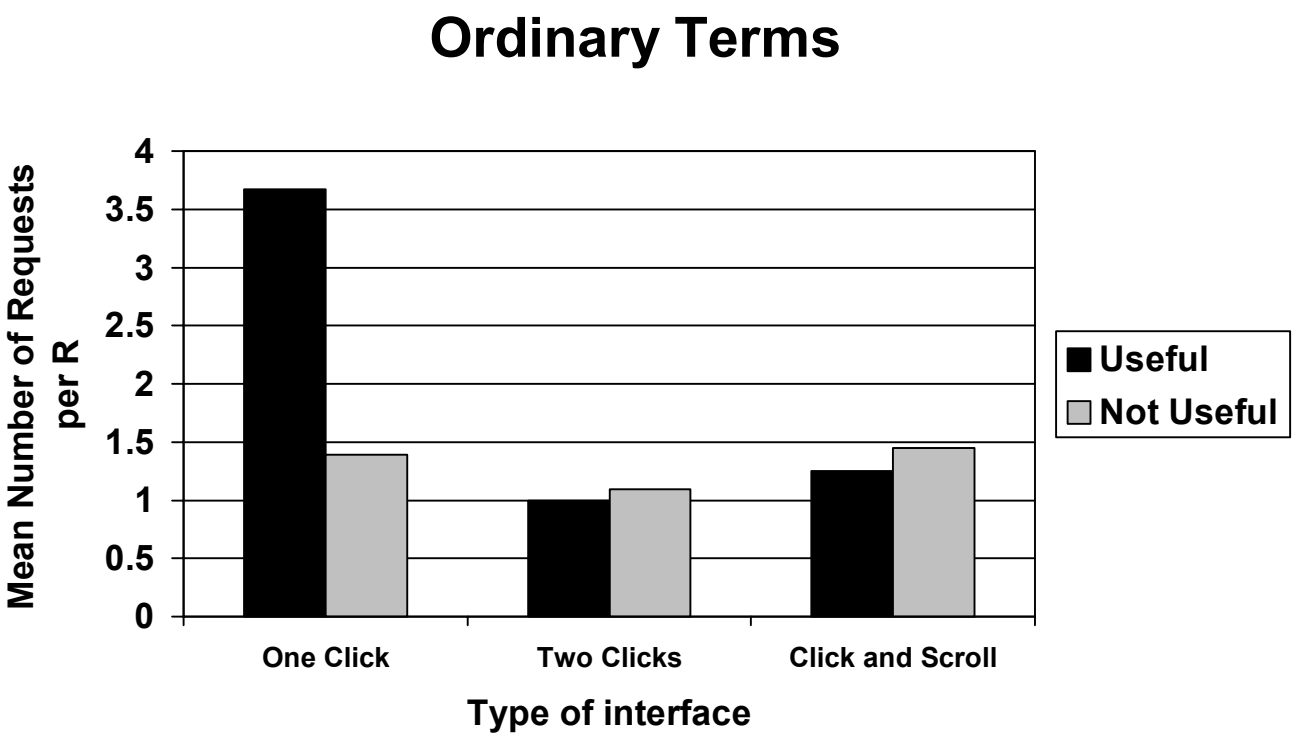


Figure 4. Mean number of definitions requested (by respondents who requested at least on definition) for ordinary terms that were useful and not useful using three user interfaces.

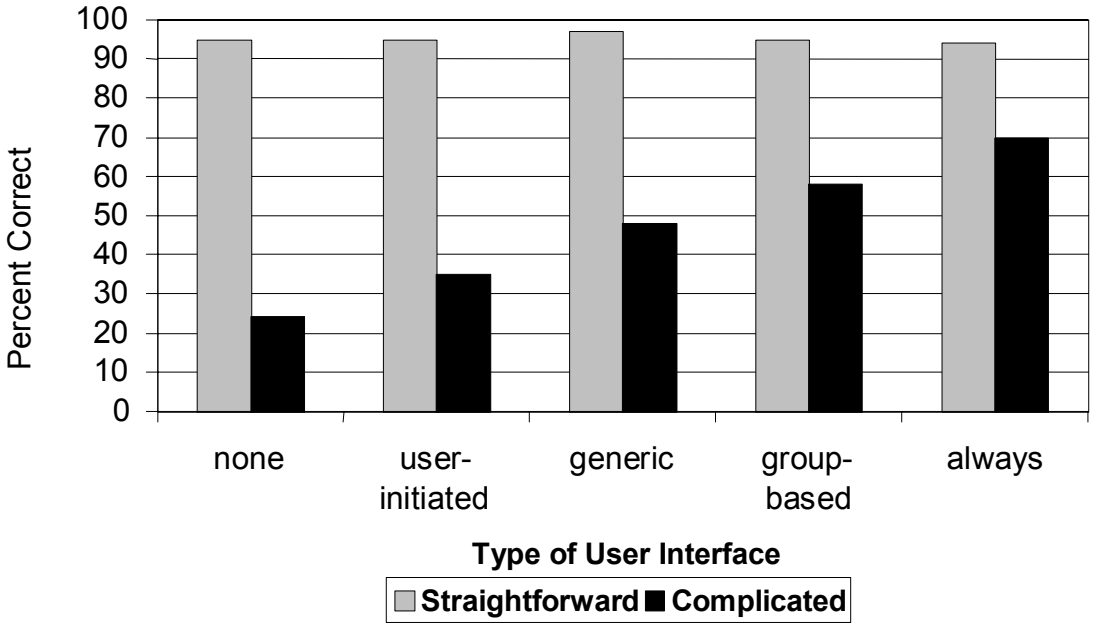


Figure 5. Response accuracy for straightforward and complicated scenarios when respondents used five types of user interface.