Interactive Feedback Can Improve the Quality of Responses in Web Surveys

Frederick G. Conrad[1,2]
Mick P. Couper[1,2]
Roger Tourangeau[1,2]
Mirta Galesic[2]
[1]*Institute for Social Research, University of Michigan*
[2]*Joint Program in Survey Methodology, University of Maryland*
Frederick G. Conrad, University of Michigan
426 Thompson Street, Ann Arbor, MI 48104
fconrad@isr.umich.edu

Key Words: web surveys, interactive feedback, data quality, tally items, human-computer interaction

Abstract

The current study explores the benefits of interactive feedback in web surveys using tally items as a test bed. Tally items present respondents with a series of answers that must sum to a constant value. Feedback about the sum of responses led to more answers that were equal to the fixed sum (100%) than no feedback and the benefits were greater when the feedback was displayed while respondents entered each answer (concurrent) than when it followed submission of the whole series of answers (delayed). The advantages of concurrent feedback accrued without increasing response times; in fact, responses were faster with concurrent than delayed feedback. The benefit to response time was observed despite an increase in the number of answers changed with concurrent than delayed feedback. The rate of response change was interpreted as evidence of more careful and, potentially, more accurate responding. Nonetheless, the tendency to revise early items – an indication of least-effort and, probably, less accurate responding – was evident for all respondents, whether or not they received feedback and whether the feedback was concurrent or delayed. We conclude with a discussion of other applications for interactive feedback in web surveys.

1. Introduction

What are the pros and cons of providing feedback to respondents on their performance of survey tasks? There is a tension in the interviewing literature about whether feedback helps or hurts. On the one hand proponents of standardized wording (e.g. Fowler & Mangione, 1990) argue that anything other than the most neutral feedback can be leading and can distort answers. On the other hand, it has been demonstrated that allowing interviewers to help respondents determine whether or not to consider particular events or behaviors when answering a question can improve their response accuracy (e.g. Conrad & Schober, 2000; Schober & Conrad, 1997; Schober, Conrad & Fricker, 2004). Paper questionnaires, in contrast are not interactive, and so can neither hurt nor improve answers through feedback and because no interviewer is present the respondent is control and can complete the questionnaire when convenient. Web questionnaires promise to potentially bridge these approaches by providing reliable feedback about answers without involving an interviewer.

In other domains, feedback is most useful when it is given immediately after the relevant action. For example, in cognitive psychology, immediate feedback on low-level actions has long been recognized as essential to skill acquisition (e.g. Anderson, 1983) and, in human-computer interaction, feedback after each user action has long been recognized as critical to improving user performance and satisfaction (e.g. Shneiderman, 1997). However, computer users seem reluctant to take advantage of immediate feedback if it requires any effort – even an eye movement can be too much (e.g. Gray & Fu, 2004). We have observed a similar reluctance among web survey respondents to use interactive features if they require as much as a click: respondents rarely obtained definitions available for words in the questions when this required a click although when this could be done with a roll-over they were more than four times as likely to obtain definitions (Conrad, Couper, Tourangeau & Peytchev, in press). The amount of effort for a user (or respondent) to obtain feedback is minimal if it is initiated by the system. In this case, the respondent can do what he or she would ordinarily do, i.e. no clicks or roll-overs, and still receive feedback.

When the system generates feedback, the respondent does not need to exert any effort to obtain the feedback. In fact the only effort required by the respondent is to read or otherwise interpret the feedback. Feedback should be especially useful if it makes the task itself easier. Because computers are better than people at arithmetic, feedback that contains the arithmetic results can make the respondent's task easier by eliminating the need to do mental arithmetic. Web designers frequently include arithmetic

functionality in on-line forms, e.g. filling in a user's age based on the value for date of birth and the current date, totaling registration expenses (admission, meals, short courses, tee-shirt, etc.) for an event such as a conference, presenting flight departure and arrival times as well as duration, etc. We discuss a type of interactive feedback that tallies a series of answers for respondents either after the tally has been submitted (delayed) or while the respondent is entering individual answers (concurrent).

Although we focus on a particular type of interactive feedback in web surveys, we wish to make two general points about such feedback in web surveys: feedback can improve respondents' performance and immediate (concurrent) feedback produces even greater improvement than delayed feedback.

1.1 Tally (Constant Sum) Questions

Web surveys frequently pose response tasks in which a series of answers must add to a constant value. The value may be imposed by the designer (e.g. 24 hours) or based on a previous answer. Some tally questions are programmed with code to sum the component answers. Older browsers may not support client-side computation but we don't think that is sufficient reason to withhold arithmetic feedback from all respondents. Rather, the decision about whether to provide or withhold such feedback should depend on whether it helps or hurts performance – which is what the current study is intended to do.

In determining whether such feedback helps or hurts we would ideally be able to measure its impact on response accuracy. However, we could not do this in the current study because we did not have access to verification data. Instead our primary measure of quality was whether the component answers sum to the target value. So we really measure whether the tallies are well-formed, not whether they are accurate. Of course a tally that does not equal the target sum cannot be entirely accurate but it is possible for a tally to match the target and be inaccurate because some of the component answers are incorrect.

2. Current Study

2.1 Experimental Design

All respondents were randomly assigned to one of three feedback conditions: (1) None, (2) Delayed, and (3) Concurrent + Delayed (which we refer to simply as Concurrent). The server presented feedback about the sum of the individual answers after the respondent had submitted them (Delayed) if the sum did not equal 100%; if the sum was equal to 100%, then no delayed feedback was presented. The concurrent feedback took the form of a running tally, computed on the respondent's browser, incremented as the respondent entered each component answer.

The tally item asked about the respondent's Internet usage in nine categories: email, news, searching for and retrieving information, instant messaging and chatting, commerce, travel planning, video and music downloads, playing games, taking a course and other (see Figure 1). The order in which the categories appeared was randomized for each respondent with the constraint that the "other" category always appeared in the final position. The delayed message, if it appeared, appeared just once. The system accepted all subsequently submitted tallies without feedback, whether or not they were well formed. In fact respondents could have submitted an unrevised tally or, for that matter, any response at all.

The tally item with delayed and concurrent feedback appears in Figure 1. The concurrent feedback appeared in the bottom field labeled "Total" (the tally was "85" when this screen was captured) and the delayed feedback appeared in red font at the top of the screen ("Your answers do not add up to 100%. Please revise your answers so that they add to 100%").The delayed feedback appeared after the respondent has submitted the tally by pressing "Next Screen."

2.2 Respondents

3195 respondents were randomly assigned to the three conditions. The respondents were recruited from two commercial opt-in panels, 1438 (52.5%) from SSI Survey Spot and 1301 (47.5%) from AOL Opinion Place. SSI Survey Spot maintains a list of email addresses and many of the panel members are experienced web survey respondents. AOL Opinion Place routes AOL users to web surveys through a banner advertisement and they tend to be less experienced respondents. As an incentive SSI offers respondents who complete a survey inclusion in a sweepstakes with cash prizes; AOL Opinion Place offers them American Airlines miles.

2.3 Results

We turn first to accuracy which we assess in two ways: Overall accuracy is the proportion of tallies that were ultimately equal to 100%, i.e. including those that were revised after initial submission; initial accuracy is the proportion of tallies equal to 100% prior to submission, i.e. correct without revision. The clear result is that for both measures, respondents' tallies are more accurate with than without feedback. Overall accuracy was 84.8% without feedback, 93.1% with delayed feedback and 96.5% with concurrent

feedback (χ^2 (2) = 84.45, p <.0001 overall, and χ^2 (1) = 10.67, p<.01 for delayed versus concurrent feedback).

Looking only at the percent of tallies that were equal to 100% when initially submitted, there was again an advantage for concurrent over delayed feedback. Those respondents who received only delayed feedback submitted initial well-formed tallies on 84.8% of occasions (equivalent to the rate for respondents who received no feedback) while those who also received concurrent feedback submitted well-formed tallies on the first try 93.1% of the time (χ^2 (1) = 30.27, p <.0001).

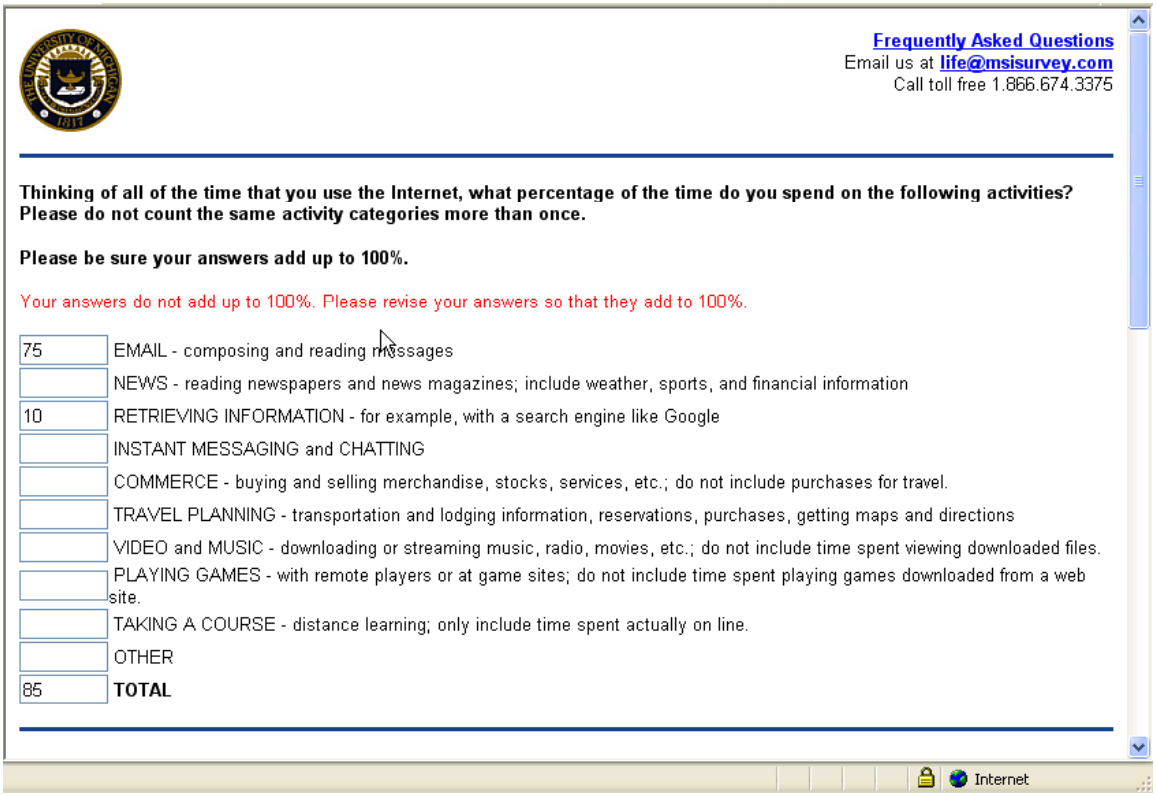

Figure 1: Tally item with delayed and concurrent feedback.

It is possible that concurrent feedback increases the number of tallies that equal 100% but also increases the time to submit a tally: feedback could promote more thoughtful responding (e.g. more thorough recall) or increased tinkering with the individual responses until they add to 100%, both of which would increase the overall response time. Alternatively, concurrent feedback could reduce overall response time by reducing the amount of mental arithmetic or by simplifying revision of tallies "bounced back" by the server. In fact, concurrent feedback led to reliably quicker responses (88.8 seconds) than delayed feedback (98.0 seconds), (F[1,2707]=11.17, p<.0001) and responses that were equivalent in duration to those for the group that did not receive feedback (86.9 seconds), (F(1,2707) < 1, n.s.). Concurrent feedback seems to have had its effect in speeding up the production of accurate (well-formed) tallies prior to submission rather than in reducing the time required to revise tallies returned by the server. Time to submit an initially well-formed tally was reliably faster if concurrent feedback was provided (85.1 seconds) than if the feedback was delayed (94.1 seconds), t (1565) = 2.28, p=.02. In contrast there was no difference between overall response times under the two feedback conditions

when the initial tally did not equal 100% and revision was undertaken at least some of the time, concurrent (139.2 seconds) versus delayed (134.5 seconds) feedback $t(196)= -0.41$, n.s. Perhaps those respondents who receive concurrent feedback but submit ill-formed tallies do not use the concurrent feedback extensively either before or after submitting their answers.

Feedback in general and concurrent feedback in particular, seemed to help respondents enter answers that add to the constant sum (100%). However, this does not assure that these answers are necessarily more accurate than are answers that do not sum to 100%. While it is generally true that an ill-formed tally includes at least one error (e.g., one answer is too large and so pushes the total over the target value), a well-formed tally is not necessarily based only on accurate answers. Any individual answer can suffer from measurement error just as an isolated response might suffer. And there could be additional measurement error as a result of the interdependencies among answers: it is possible that respondents might "coerce" certain answers so that the set of answers totals appropriately rather than because the adjusted values are more accurate[2].

We cannot directly measure response accuracy in the current study because we do not have a gold standard against which to compare responses but we can look at certain measures that may suggest more or less accurate responses due to feedback. First, it is possible that more changes to initial answers on the basis of feedback could indicate more thought about the particular values and thus greater validity. For example, respondents might change certain answers because the process of answering other questions brings information to mind that affects their thinking about the earlier ones. If this is the case then we would expect more changes for concurrent than delayed feedback.

Second, if changes tend to be heaped in certain serial positions – for example a predominance of changes to the first category – this would seem to reflect lower accuracy, probably due to the kind of coercion mentioned above. The categories in this tally item were not ordered in a meaningful way, so unless respondents believed they were, changes made to increase accuracy should have been distributed uniformly across the list. However, if respondents' goal was to produce well-formed tallies with the least possible effort then modifying responses primarily from the earlier part of the list would be a sensible strategy. This would be close in spirit to the satisficing explanation proposed by Krosnick and Alwin (1987) for primacy effects observed when respondents choose answers from unordered, visually presented lists.

Concurrent feedback tended to produce more response change than delayed feedback prior to submitting tally. There was difference for tallies that were well formed (2.41 versus 2.48 changes per respondent for concurrent and delayed feedback respectively, $F[1,2734] < 1$, n.s.) but there was significant effect for tallies that were returned by the server because they did not equal 100% (2.88 versus 1.65 changes per respondent for concurrent and delayed feedback respectively, $F[1, 2734]=4.28$, $p=.04$). Clearly those who do not receive feedback prior to submitting the tally are relatively unlikely to adjust their answers (although in this analysis concurrent feedback did not actually lead to well-formed tallies). While this pattern generally speaks to improved quality of estimates due to feedback, we cannot conclude this on the basis of the current data.

The patterns of change by serial position of the categories are suggestive of least-effort revision irrespective of the type of feedback and so bode less well for improved accuracy than do the overall number of changes. The numbers of changes at each position before the delayed feedback or any changes for the No Feedback respondents are displayed in Figure 2. These data test the impact of concurrent feedback against no feedback because the server has not presented any delayed feedback at this point. Note that there is a strong primacy effect across the three groups ($F[9, 24624] = 172.24$, $p<.001$ and the patterns of change are virtually identical across the groups (interaction of position and group, $F[18, 24624)= 1.02$, n.s.). The respondents made the largest number of changes to whatever items happened to appear in the first three positions and the number of changes drops off more or less monotonically over the ten serial positions. This seems consistent with a satisficing approach in which respondents revised the first items they encountered while reading from top to bottom until they judged their tally ready to submit.

There is a curious peak at the third position for all groups. Perhaps many respondents anchored their

[2] It is also possible that in some tasks (e.g. measuring time use) categories may legitimately be double counted (watching TV while cooking) and add to more than the constant sum (24 hours); requiring the answers to reach a fixed value could lead to respondents to change answers to less accurate values.

answers on the basis of the first two items and then revised the items that immediately followed. We can only speculate what this indicates; however, because the items are randomly ordered, any position effects seem to reflect suboptimal responses.

Figure 2: Number of changed answers by serial position.

2.3 Summary

Feedback about the sum of responses led to more answers that add to the constant value relative to no feedback. The advantage was even greater with feedback that was displayed while respondents entered each answer (concurrent) than feedback which followed until after the set of answers was submitted. The advantages of concurrent feedback accrue without the cost of longer response times; in fact, responses were faster with concurrent than delayed feedback. Yet respondents changed more answers with concurrent than delayed feedback which could reflect more careful and, potentially, more accurate responding. Nonetheless, the tendency to revise early items – an indication of least-effort responding – was evident for all respondents, whether or not they received feedback and whether the feedback was concurrent or delayed.

3. Implications

The interactive feedback in the current study certainly produced cleaner if not more accurate answers. This basic idea seems applicable far beyond tally questions. DeRouvray and Couper (2002) found that in a web TV questionnaire, a prompt that stressed the importance of answering – but did not require it – led to a reduction in the selection of the "decline" option and advancing without answering at all. One can imagine a more interactive approach in which, for example, the prompt can be tailored to the number of skipped items.

Interactive feedback can be used to improve data quality in other ways. A respondent who answers a grid format item (in which the columns contain response options – usually radio buttons – and the rows present related questions) by selecting the same response for all questions ("straight-lining") could be prompted to be more discriminating. A respondent who answers a question very quickly – perhaps faster than other respondents in the same age group – could be reminded to give the question more thought or a respondent has not answered after a fixed interval– perhaps the median response time for others in the same age group – could be offered help (see Ehlen, Schober & Conrad, 2005, for an example of this approach).

The web makes it possible to provide useful information to respondents at little relatively little development cost and so makes it possible to combine the best of interviewer- and self-administration. And

the more interactive the presentation of this information the more it may help. But interactive feedback may not be a panacea. Until a general theory of interactivity and web survey response has been developed and until there is a body of definitive studies, each case will need to be evaluated so that it is clear whether it helps or hurts.

Acknowledgments

We thank Reg Baker and Duston Pope for invaluable advice on the design of the tally item under study and its implementation. We also thank the National Science Foundation, Grant No. Grant SES0106222 and National Institutes of Health: Grant R01 HD041386-01A1.

References

Anderson, J.R. (1983) *The architecture of cognition.* Cambridge, MA: Harvard University Press 1983

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

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

DeRouvray, C. & Couper, M.P. (2002). Designing a strategy for reducing "no opinion" responses in web-based surveys. *Social Science Computer Review,* 20, 3 - 9.

Ehlen, P., Schober, M.F., & Conrad, F.G. (July, 2005). Modeling speech disfluency to predict conceptual misalignment in speech survey interfaces. *Proceedings of the Symposium on Dialogue Modelling and Generation, 15th Annual meeting of the Society for Text & Discourse,* Vrije Universiteit, Amsterdam, 2005

Fowler, F. J. & Mangione, T.W. (1990). S*tandardized survey interviewing: minimizing interviewer-related error*. Newbury Park, CA: SAGE Publications, Inc.

Gray, W.D. & Fu, W.-T. (2004) Soft constraints in interactive behavior: the case of ignioring perfect knowedge in-the-world for imperfect knowledge in-the-head. *Cognitive Science*, 28, 359-382.

Krosnick, J.A. & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly,* 51, 201-219.

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. (2004). Misunderstanding standardized language in research interviews. *Applied Cognitive Psychology, 18*, 169-188.

Shneiderman, B. (1997). *Designing the user interface: strategies for effective human-computer interaction.* Reading, Massachusetts: Addison-Wesley Publishing Company.