Can Survey Instructions Relieve Respondent Burden?

Erica C. Yu[1], Scott Fricker[1], Brandon Kopp[1]
[1]Bureau of Labor Statistics, 2 Massachusetts Avenue, NE, Washington, DC 20212

Abstract
Survey designers aiming to reduce respondent burden often choose to reduce the number of questions asked of the respondent. However, such a choice may result in a trade-off between respondent burden and data utility. This research investigates whether survey instructions can affect ratings of burden by manipulating the context in which the survey experience is judged. Participants were recruited from Amazon Mechanical Turk to complete an online survey (n=544). Two factors were manipulated in a between-groups design: actual number of questions asked (24, 42) and burden frame (screened in to an extra survey section, screened out of an extra survey section, no screener instructions). The results show a main effect of survey length whereby the objectively longer survey was associated with higher respondent burden. However, there also was a main effect of burden frame whereby being told one was screened out of a longer survey lowers ratings of burden compared to being told one was screened in. Survey instructions are able to influence respondents' perceptions of the survey experience, regardless of the objective features of the survey, and ultimately affect respondent burden.

Key Words: Respondent burden, context effects, questionnaire design, survey length

1. Introduction

Survey designers aiming to reduce respondent burden often choose to reduce the number of questions asked of the respondent. However, such a choice may result in a trade-off whereby improvements to respondent burden come at the cost of lowered data utility as the quantity and richness of information that can be collected from respondents is reduced. Moreover, survey length is only one dimension of respondent burden, of which there is believed to be many (Bradburn, 1979). Survey designers may be able to reduce respondent burden through other channels.

1.1 Perceived respondent burden
Survey research literature suggests that respondent burden may be defined in two ways: by its objective features, such as the number of minutes taken to complete a survey or the number of questions in a survey, and by a respondent's perceptions of those features, which may manifest in negative feelings about the survey experience due to interview length, required effort, frequency of being interviewed, and stress of difficult to answer questions (Bradburn, 1979, Sharp & Frankel, 1983; Fricker, Yan, & Tsai, 2014). These feelings of burden may arise during the survey while responding to questions or at some time afterwards as a retrospective judgment. In the former case, such burden may lead to poor data quality through lowered respondent effort and higher attrition rates (Rolstad, Adler, & Ryden, 2011). And in both cases, such burden may lead to lowered willingness to participate in future surveys.

Previous studies have shown that announcing a longer length survey lowers response rates (Crawford, Couper, & Lamias, 2001; Marcus, et al. 2007; Galesic & Bosnjak, 2009). However, these studies have not disassociated the announced (perceived) length from the actual experienced length of the survey. For example, Galesic and Bosnjak (2009) use three survey formats of 10, 20, and 30 minutes in length. The introductory page of the survey announced these expected durations and then the subsequent survey instrument was tailored to match those expectations. While these studies can reveal new information on respondents' willingness to participate, it is not possible to distinguish the effects of the objective characteristics of a survey that may cause burden and a respondent's perception of those characteristics.

1.2 Context-dependent judgments

Previous research has shown that judgments often are made relative to a reference set. For example, individuals have been shown to care more about their income level relative to a socially-constructed group of peers rather than in absolute terms, and that this rank judgment is ultimately correlated with ratings of well-being (Boyce, Brown, & Moore, 2010; social comparison, Festinger, 1954). Even ratings of objective physical phenomena such as brightness and loudness are subject to context dependence (Lockhead, 2004). Perceptions need not be tied to their objective reality.

Likewise, a respondent's survey participation is experienced not in a vacuum but in a context relative to other experiences that the respondent considers at the time of judgment. Researchers have shown that responses to attitude questions depend on the context created by earlier items in the questionnaire (Tourangeau, et al., 1989). Building on this idea, we propose that, depending on the context, a respondent may judge the same survey experience to be not at all burdensome (compared to preparing taxes) or extremely burdensome (compared to ordering a pizza). A respondent can perceive that a survey is burdensome but that survey does not need to adhere to any objective measure of what a "burdensome survey" means.

If survey designers could influence the context, or reference set, against which respondents judged their burden, they may be able to affect a respondent's judgment of their burden. This possible context dependence presents a unique opportunity for survey designers in which a survey does not need to sacrifice data quality or quantity for the sake of reducing respondent burden.

2. Methods

2.1 Design and Materials

The goal of this study was to investigate whether survey instructions can affect ratings of burden by manipulating the context in which the survey experience is judged. We used a 2x3 between-groups experimental design that manipulated survey length—an objective characteristic of surveys believed to contribute to burden—and the context in which survey length was perceived. With this design, we are able to separate the effects of objective and perceived survey length on overall ratings of burden.

During the introduction to the survey, participants were told that they would be answering questions about their attitudes toward the Bureau of Labor Statistics. The instructions stated

that the main survey would "take about 5 minutes, on average." This average task duration time was announced to participants in advance of their commitment to participate in the survey.

The questions were part of a web survey and presented in grid format, such that each page presented six questions. One-half of participants were presented with a short survey of 24 questions total while the other half of participants were presented with a long survey of 42 questions total.

The first six questions presented to all participants were presented as screening questions. After submitting their responses to these questions, participants then received instructions that were intended to manipulate their senses of perceived burden. One-third of participants were told they were screened in to a survey that was longer than what others were asked to do, with the following text:

> *Based on your answers, we will now ask you to complete the long version of our survey and go through additional sections and answer extra questions.*
>
> *Unfortunately, you will have to answer more questions that other respondents don't have to. The survey will take more of your time and effort than originally estimated. We appreciate your participation in the survey.*

One-third of participants were told they were screened out of that longer survey and so had to do less than others, with the following text:

> *Based on your answers, we will now ask you to complete the short version of our survey and skip past some survey sections and answer fewer questions.*
>
> *Fortunately, you don't have to answer those questions that other respondents have to answer. The survey will take less of your time and effort than originally estimated. We appreciate your participation in the survey.*

The final third of participants did not receive any special instructions. These participants proceeded from the first question grid directly to the next question grid, without any awareness of the screening protocol used with other participants.

With this factorial design, we are able to assess respondent burden for participants who experience the same objective survey length (e.g., 42 questions) but have different burden frames (screened in or screened out). Likewise, we are able to assess burden for participants who experience the same burden frame (screened out) but experience different survey lengths (24 or 42 questions). Because the manipulation of burden frame occurs during the survey, rather than afterwards as part of the burden judgment question, we believe that the manipulation will affect how participants feel during the survey.

After completing the screener questions and receiving any special instructions, the participants completed the rest of the survey (short: 18 additional questions; long: 36 additional questions). The question grids for the remainder of the survey were presented in a random order; although the participants assigned to complete only the short survey

completed only three additional grids of questions, those three grids were randomly selected from the possible six grids available. Participants assigned to complete the long survey were given all six grids, in a random order. Then, a screen instructed participants that the main survey was over and that participants would next be asked to answer debriefing questions about the experience of completing the main survey. The debriefing questions asked participants for ratings of how burdensome it felt to participate in the survey and perceived: speed of time passing, survey length, ease of completion, importance, interest level, effort required, trustworthiness of the administering institution, and willingness to participate in a similar survey in the future. Open-ended responses describing examples of activities at varying levels of burdensomeness were also collected.

2.2 Participants

Individuals were recruited from Amazon Mechanical Turk (mTurk), an online labor market where researchers can post "jobs" for "workers" to complete for pay. MTurk provides access to a large convenience sample of participants across a range of demographics much broader than is typically available to researchers conducting in-person research with community-based participants. However, mTurk participants are likely to have past experience with similar research studies.

For this study, we recruited 544 participants, of which 488 participants completed the study. Of these participants, 53% were male, 7% identified as Hispanic or Latino, the majority identified as White (85%; 10% identified as Black, 6% as Asian, and fewer than 5% as American Indian or Alaska Native, or Native Hawaiian or Other Pacific Islander). The mean age was 37 years old (SD = 12.3). The median level of education attained was a Bachelor's degree.

For completing the survey, participants received a payment of $0.75. This amount was fixed and did not depend on their responses, only on the completion of the survey.

2.3 Procedure

The online study was administered April 23 and 24, 2015. Upon starting the survey, participants began by reading instructions and answering the first six survey questions. Then, participants saw the randomly assigned screener outcome, and went on to complete the rest of the survey and the debriefing questions.

3. Results

3.1 Ratings of Burden

After completing the main survey, participants were asked to rate how burdensome it felt to participate in the survey, on a five-point scale ranging from "Extremely Burdensome" to "Not at all Burdensome". The responses to this question are reverse-coded so that low burden ratings correspond to low levels of burden and summarized in Figure 1. Overall, the mean burden rating was 1.68, suggesting that this task was just less than "Slightly burdensome" to participants.

As expected, there was a main effect of manipulated survey length whereby participants who experienced the longer survey rated the survey as more burdensome ($F(1, 480) = 8.73$, $p = 0.003$. There also was a main effect of screener whereby participants who were told they were screened in rated the survey as more burdensome than participants who were

told they were screened out ($F(2, 480) = 4.38$, $p = 0.013$; Bonferroni post-hoc comparison: $p = 0.014$). Participants in the control condition who did not receive any special screener instructions fell in the middle and were not significantly different from either screener group. Importantly, this main effect demonstrates that, regardless of the length of the survey experienced, the burden frame instructions pushed ratings of burden up or down as intended. The interaction between survey length and burden frame instructions was not significant ($F(2, 480) = 1.10$, ns).

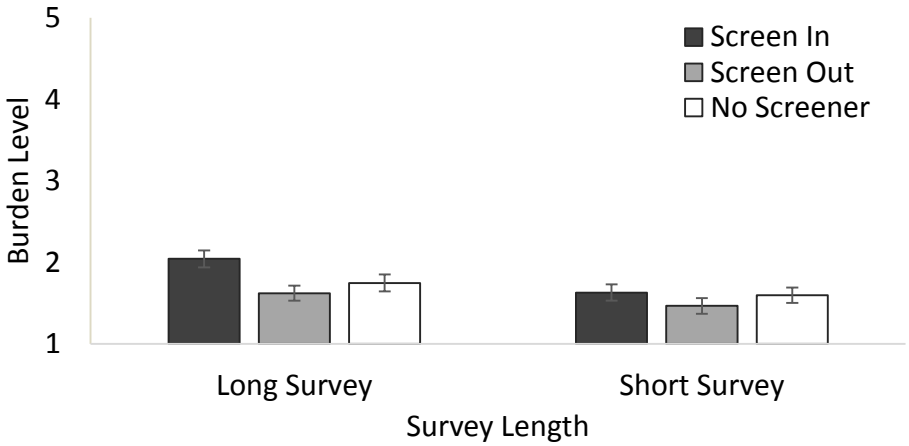

Figure 1: Ratings of overall burden from participating in the survey.

A model including task duration (sum of time spent on each question grid excluding time spent reading about the outcome of the screener) as a covariate finds that task duration has no effect on burden ratings ($F(1,469) = 0.009$, ns) while the effects of burden frame and actual survey length remain unchanged. In other words, participants' ratings of burden were related to the number of questions asked and their perceptions of the length of the survey, but not related to the time spent on the survey.

3.2 Ratings of Survey Length

During debriefing, participants also completed ratings of perceived survey length. The question asked participants to rate how short or long they felt the survey was, on a seven-point scale ranging from "Very short" (1) to "Very long" (7). These responses are summarized in Figure 2. Overall, participants rated the length at 2.12, or between "Short" and "Somewhat short" on the response scale.

The survey instrument collected the amount of time spent on each grid of six questions from the time that the page was presented until the time that the page was submitted. Summing the time spent on each question grid results in a measure of overall task duration excluding any time spent on reading the outcome of the screener (this was excluded due to anticipate variance due to reading speed and the absence of any screener outcome text in the control condition). The mean task duration across all participants was 187.65 seconds, and the median 158.16 seconds.

Analysis of these data shows a significant main effect of the actual survey length manipulation on perceived survey length, such that participants who were asked relatively more questions judged the survey to be longer than those in the short survey condition ($F(1, 480) = 39.26$, $p < 0.001$). Although this is an expected finding, this result does indicate that

the scale could distinguish between long and short surveys, even though both surveys were relatively short.

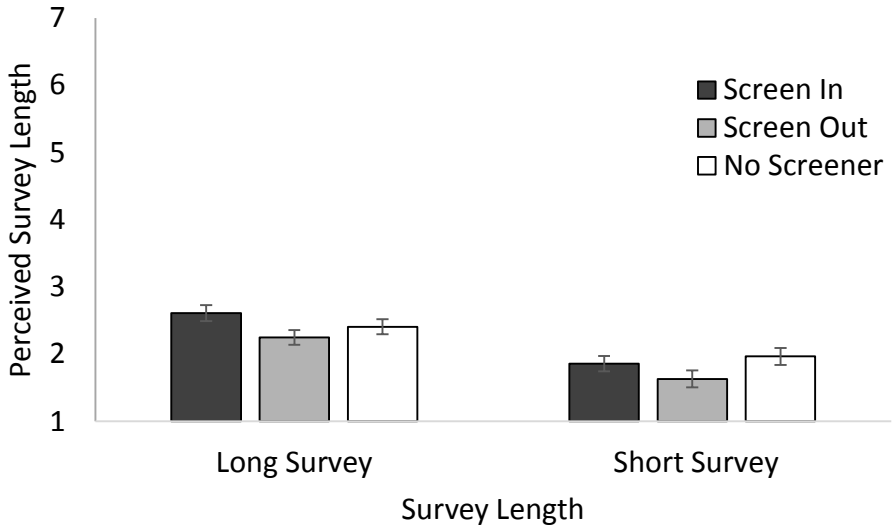

Figure 2: Ratings of perceived survey length.

There also was a main effect of screener on the ratings of perceived survey length ($F(2, 480) = 3.72$, $p = 0.025$), showing that the participants who were told they would experience a longer survey rated the survey as marginally longer than participants who were told they would experience a shorter survey, regardless of the actual survey length experienced (screened in: M = 2.24; screened out: M = 1.94; Bonferroni post-hoc comparison: $p = 0.073$). A model including task duration finds that task duration was not related to ratings of perceived length ($F(1, 479) = 1.99$, ns). This result indicates that perceptions of survey length were unrelated to the amount of time each participant spent on the task.

3.3 Willingness to Participate in the Future and Other Debriefing Ratings

Participants answered several other debriefing questions, including a question on their willingness to participate in another similar survey in the future. Although only a hypothetical question, this question provides the only indication as to whether the burden experience would affect future survey participation. Analysis finds that objective survey length does significantly affect willingness to participate in the future ($F(1, 480) = 4.41$, $p = 0.036$) but no effect of screener instructions. Analysis of participants' ratings of how quickly or slowly time passed during the survey, and how easy or difficult the survey was to complete had a similar pattern of results, whereby participants in the longer survey thought the time passed more slowly ($F(1, 480) = 13.40$, $p < 0.001$) and that the survey was more difficult to complete ($F(1, 480) = 8.33$, $p = 0.004$). Other debriefing questions that were not related to ease or time showed a different patterns of results: importance, interest, and effort were unaffected by survey length or screener instructions.

3.4 Time Spent on the Survey

Given that the order of the grids presented was randomized, we can compare how long participants took to complete blocks over time to judge whether participants spent less time on later questions. These data have been re-coded to reflect the chronological point at which the questions were presented; participants may have seen different questions at

"Block 2" but these data show the time spent for all participants on their second block of questions. The data are summarized in Figure 3 for only those participants who completed the long version of the survey.

We found that participants spent less time on later blocks, due to general learning effects, less consideration given to the questions, or some other reason ($F(2.76, 1392) = 4.21$, $p = 0.007$; Greenhouse-Geisser correction). For example, the mean time taken to answer the first six questions was 44.52 seconds, while the mean time taken to answer the final six questions was 22.64 seconds. In the same model (including factors for both block within subjects and burden frame between subjects), we found that the screener instructions did not have a significant effect on timing ($F(2, 232) = 2.04$, ns), suggesting that those participants who persisted in completing the survey despite being told they were being given more questions than others did not reduce their considerations of later questions.

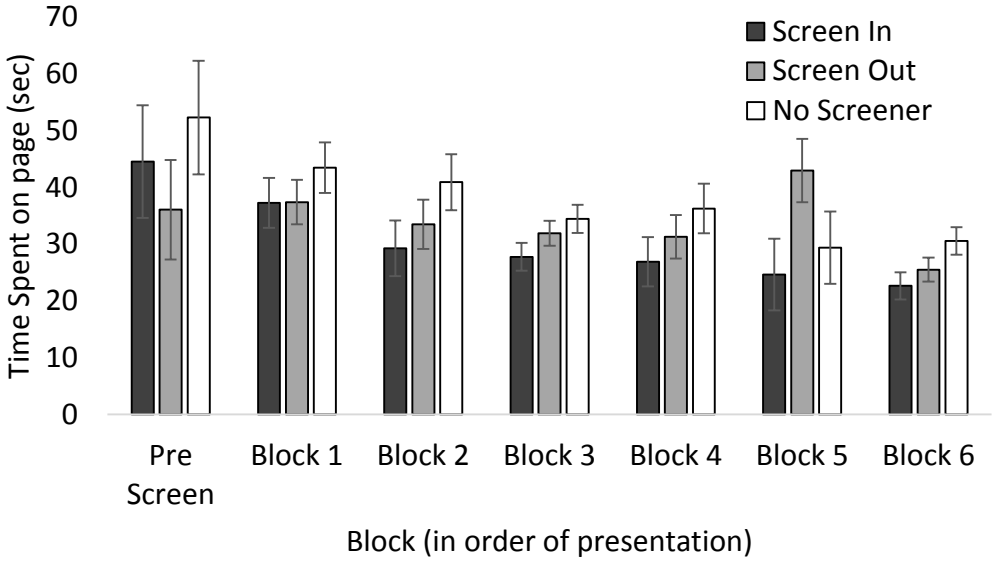

Figure 3: Time spent on each page, in order of presentation.

3.5 Break-offs from the survey

The screener outcome manipulation was presented to participants after answering the first six questions. We can evaluate the number of people who broke-off from the survey immediately after presentation of the screener outcome. A total of 56 participants quit the survey. These data are summarized in Table 1.

Table 1: Break-offs from the survey immediately after screener instructions

Condition	Break-offs	Rate
Screened in	18	10.29%
Screened out	12	6.70%
No screener	5	2.63%

A chi-square analysis finds a significant difference in break-offs due to screener outcome (X^2 (2, n = 544) = 8.90, $p = 0.012$). There was no significant difference due to survey length, as expected given that the participants had not yet experienced any differences in survey length (X^2 (1, n = 544) = 0.46, ns). Examining the rates of break-off from each condition

shows that the highest rates of break-off are from those participants who were told they were screened in to take a longer survey. Immediately after learning they would be taking a longer survey than everyone else, 10.29% of participants decided to quit. Within the domain of mTurk, this finding is especially meaningful because these participants forgo all payment despite already having invested time and effort into the task. Interestingly, participants who were told they were screened out still broke-off at a higher rate than did the control condition participants. These data suggest that any information about screening pushes people away from the survey, compared to simply carrying on with survey questions.

3.6 Qualitative analysis of burden definitions

In addition to ratings, we collected open-ended responses on how participants think of burden. Participants were asked to name activities that they would give ratings of "extremely burdensome", "somewhat burdensome", and "not at all burdensome". Text analysis of these data are still ongoing but we can highlight two interesting findings.

Several activities were universally understood to be not at all burdensome (e.g., eating) or extremely burdensome (manual labor). However, different participants named the same activity for all three levels of burden, indicating that burden is not a stable concept across people. For example, cleaning and chores, taking care of kids, and cooking appeared repeatedly at each of the three levels of burden.

Many activities named were physical, such as mowing the lawn or building a brick wall fence, but an overwhelming number of activities were mental or emotional, such as juggling work and life obligations, dealing with customers, or being unemployed. Themes that emerged included unfair and stressful activities.

4. Discussion

This study investigated the effect of burden frame on ratings of respondent burden to disassociate respondents' perceptions of survey length from the objective survey length. We found that each has its own effect on overall burden, and that perceptions of burden as influenced through survey instructions can change respondents' perceptions of the survey experience. Moreover, measures of actual time spent on the survey were unrelated to either ratings of survey length or burden, further supporting the hypothesis that perceptions are not grounded by objective measures. We also gathered evidence that "burden" is a complicated concept that can vary from person to person and represent a range of activities, including mental and emotional factors that may be impossible to quantify on an objective scale.

This study illustrates that the "burden" concept measured by survey designers is a psychological phenomenon, and as such it can be affected by cognitive biases including context effects. By recognizing this pathway from context to burden, survey designers can take responsibility for the entire survey experience from the advance letter to the end of an interview as potentially affecting a respondent's feelings of burden. Beyond this, survey designers can also pro-actively manage a survey's context to reduce a respondents' feelings of burden. Survey instructions are another tool in the survey designer toolbox.

The finding that task duration was not related to ratings of burden may indicate that time spent on surveys (e.g. "burden hours") may not be an appropriate characterization of the overall burden we impose on survey respondents. However, whether this result would remain with longer surveys is unknown. It is also possible that factors such as interest, effort, and perceived importance play moderating roles. Future research should explore alternative measures of burden.

The manipulation used to alter participants' perceptions of burden relied on two comparisons – a comparison to participants' original expectations ("the survey will take more of your time and effort than originally estimated") and a comparison to a social peer group ("unfortunately, you will have to answer more questions that other respondents don't have to"). These two factors were used together to increase the likelihood that the manipulations would be successful. However, given that these two factors have been tied together in this study, we cannot say which is the primary driver of the effect on perceptions. Future research should isolate these factors to test their individual effects.

One important difference between the present study and typical surveys is that our participants initiated their survey participation themselves rather than by receiving an invitation. In regards to feelings of burden, we expect that burden would be lower among our volunteers and that any effects of perceived burden or inferior social comparisons would be amplified among a typical survey sample.

Acknowledgements

The authors thank Brian Harris-Kojetin for his comments on the experimental design.

References

Boyce, C. J., Brown, G. D., & Moore, S. C. (2010). Money and happiness: Rank of income, not income, affects life satisfaction. *Psychological Science, 21*(4), 471-475.

Bradburn, N. (1979). "Respondent burden." In L. Reeder (ed.), Health Survey Research methods: Second Biennial Conference, Williamsburg, VA. Washignton, DC: U.S. Government Printing Office.

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

Festinger, L. (1954). A theory of social comparison processes. *Human relations, 7*(2), 117-140.

Fricker, S., Yan, T., & Tsai, S. (2014). Response burden: What predicts it and who is burdened out? In *JSM Proceedings*, Alexandria, VA: American Statistical Association. 4568-4577.

Galesic, M. & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349-360.

Lockhead, G. R. (2004). Absolute judgments are relative: A reinterpretation of some psychophysical ideas. *Review of General Psychology, 8*(4), 265-272.

Marcus, B., Bosnjak, M., Lindner, S., Pilischenko, S., & Schuetz, A. (2007). Compensating for low topic interest and long surveys: A field experiment on non-response in web surveys. *Social Science Computer Review, 25*, 372-83.

Rolstad, S., Adler, J., & Ryden, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health*, 14(8), 1101-1108.

Sharp, L. M. & Frankel, J. (1983). Respondent burden: A test of some common assumptions. *Public Opinion Quarterly, 47*(1), 36-53.

Tourangeau, R., Rasinski, K.A., Bradburn, N., & D'Andrade, R. (1989). Belief accessibility and context effects in attitude measurement. *Journal of Experimental Social Psychology, 25*(5), 401-421.