

Seeing through the eyes of the respondent: an eye-tracking study on survey question comprehension

Lenzner, Timo; Kaczmirek, Lars; Galesic, Mirta

Postprint / Postprint

Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:

GESIS - Leibniz-Institut für Sozialwissenschaften

Dieser Beitrag ist mit Zustimmung des Rechteinhabers aufgrund einer (DFG geförderten) Allianz- bzw. Nationallizenz frei zugänglich. / This publication is with permission of the rights owner freely accessible due to an Alliance licence and a national licence (funded by the DFG, German Research Foundation) respectively.

Empfohlene Zitierung / Suggested Citation:

Lenzner, T., Kaczmirek, L., & Galesic, M. (2011). Seeing through the eyes of the respondent: an eye-tracking study on survey question comprehension. *International Journal of Public Opinion Research*, 23(3), 1-22. <https://doi.org/10.1093/ijpor/edq053>

Nutzungsbedingungen:

Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

gesis
Leibniz-Institut
für Sozialwissenschaften

Terms of use:

This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Mitglied der

Leibniz-Gemeinschaft

Seeing Through the Eyes of the Respondent: An Eye-Tracking Study on Survey Question Comprehension

Timo Lenzner¹, Lars Kaczmirek¹ and Mirta Galesic²

¹GESIS – Leibniz Institute for the Social Sciences, e-mail: timo.lenzner@gesis.org

¹GESIS – Leibniz Institute for the Social Sciences, e-mail: lars.kaczmirek@gesis.org

²Max Planck Institute for Human Development, e-mail: galesic@mpib-berlin.mpg.de

Asking questions is the predominant method of gathering information about people's beliefs, values, attitudes, behaviors and states of affairs (e.g. Schuman & Presser, 1981; Foddy, 1993). To ensure that the data obtained through surveys are reliable and lead to valid conclusions, respondents must comprehend the questions as intended by the survey designer and find it easy to answer them accurately. More specifically, they must understand the item, retrieve relevant information, use this information to make a judgment, and select and report an answer (Strack & Martin, 1987; Tourangeau, 1984). Depending on various characteristics of the questionnaire, respondents may find it more or less difficult to perform these steps accurately. For example, question comprehension is impeded by questions containing imprecise terms or complex syntactic structures which make it difficult to identify the question focus or represent the logical form of the question (cf. Tourangeau, Rips, & Rasinski, 2000). A suboptimal wording of survey questions can thus increase respondent burden by requiring more cognitive effort for understanding what the questions are about. If respondents have trouble understanding questions, they are likely to provide inaccurate answers (Schober & Conrad, 1997) and/or apply response strategies that reduce data quality and induce measurement error (e.g. satisficing, Krosnick, 1991; breakoff, Galesic, 2006). Moreover, given that question comprehension is the first step respondents have to perform, it is very likely that cognitive overload occurring at this stage will translate to later stages as well. Consequently, designing questions to minimize the cognitive effort required to process them is an important strategy for reducing comprehension difficulties and thus response error.

Applying a psycholinguistic perspective to survey question design, Lenzner, Kaczmirek and Lenzner (2010) identified seven text features that undermine reading comprehension and thus increase the cognitive burden imposed by survey questions: low-frequency words, vague or imprecise relative terms, vague or ambiguous noun-phrases, complex syntax, complex logical structures, low syntactic redundancy, and bridging

inferences. Using response time as a measure of the cognitive effort required to answer survey questions, they found that six of these seven problematic text features were associated with longer response times in a Web survey, presumably because they induced comprehension difficulties. Moreover, the text features were found to reduce data quality by producing more non-substantial answers (i.e. respondents gave more neutral responses to questions offering a middle category if the questions contained one text feature). While the effects of some of these text features on response bias (e.g. response variability) have received considerable attention by survey researchers (e.g. vague or ambiguous noun-phrases, Fowler, 1992; Smith, 1987), hitherto, their effects on survey question *comprehensibility* and respondent burden have scarcely been discussed in the literature on questionnaire design and have rarely been investigated experimentally (for exceptions see Graesser, Cai, Louwerse, & Daniel, 2006; Saris & Gallhofer, 2007; Tourangeau et al., 2000).

In this study, we extend the earlier findings by Lenzner et al. (2010) in two ways. First, we use eye tracking as a more direct method to examine whether comprehension is indeed impeded by these text features. While response time is a valuable indicator of the overall cognitive effort required to answer a survey question, it does not enable us to distinguish between the time required to read and understand a question (comprehension stage) and the time it takes to provide an answer (including retrieval, judgment, and response selection). By contrast, recording respondents' eye movements while answering a Web survey allows us to identify the specific parts of the question they struggle with during the comprehension stage. Of course, this does not imply that respondents always perform these cognitive tasks in a sequential order. Sometimes respondents may start to retrieve relevant information while reading and comprehending the question, for example. Nevertheless, given that eye tracking allows us to examine respondents' fixation times and counts on specific parts of the question, this technique enables us to identify comprehension difficulties with

much greater precision than does the collection of response times. Second, we examine whether these text features have different effects for different types of questions. While the earlier findings were almost exclusively limited to attitudinal questions, we include several behavioral and factual questions in this study.

DETERMINANTS OF QUESTION COMPREHENSIBILITY

Theoretical and empirical evidence from psycholinguistics suggests that survey designers can enhance the comprehensibility of their questions by paying attention to the seven problematic text features mentioned above (e.g. Duffy, Morris, & Rayner, 1988; Haviland & Clark, 1974; Horning, 1979; Inhoff & Rayner, 1986; Kimball, 1973; Kintsch & Keenan, 1973; Mosier, 1941). In general, these text features undermine reading comprehension by placing high demands on people's limited working memory capacity (cf. *capacity theory of comprehension*, Just & Carpenter, 1992). A first attempt to systematically link these text features to survey question comprehension has been made by Graesser et al. (2006) who developed the computer tool *Question Understanding Aid* (QUAID; University of Memphis, 2010). QUAID evaluates survey questions with regard to the first five text features listed above, and labels those questions as problematic that include one or more of these features. Arguing for an extension of QUAID's five components, Lenzner et al. (2010) proposed that two further problematic text features may undermine question comprehension processes to a similar degree, namely low syntactic redundancy and bridging inferences. In the remainder of this section we will briefly discuss these seven text features to provide the theoretical basis of our study. A more detailed account of these text features can be found in Lenzner et al. (2010).

Low-frequency words. Word frequency is a well-known indicator of text comprehensibility: people are slower at accessing the meaning of low-frequency words and must work harder to comprehend sentences in which they occur, compared to higher-frequency words.

Vague or imprecise relative terms. Vague or imprecise relative terms (e.g. *many, often, rarely, substantially*) can be interpreted in various ways, making it potentially difficult for respondents to determine the meaning intended by the survey designer.

Vague or ambiguous noun-phrase. Noun-phrases with unclear (e.g. *cultural events*) or ambiguous (e.g. *bank*) referents are difficult to comprehend, because respondents may not immediately know what the noun-phrase refers to or which sense of the word is relevant in the question.

Complex syntax. Complex syntactic structures (e.g. left-embedded syntax, propositionally dense sentences) quickly overload the processing capabilities of readers and require re-readings of unclear parts of the question.

Complex logical structures. Questions with complex logical structures (e.g. numerous logical operators such as *or*) require respondents to keep a large amount of information in mind while simultaneously processing other information. Thus, they quickly overload respondents' working memory capacity.

Low syntactic redundancy. Low syntactic redundancy reduces the predictability of the grammatical structure of a question, and thus makes it harder for readers to comprehend the course of action (e.g. passives, nominalizations).

Bridging inferences. Drawing bridging inferences is a time-consuming process that is required if the actual survey question is preceded by an introductory sentence and information from both sources needs to be connected.

METHOD

Design and Hypotheses

We conducted an eye-tracking experiment to examine the effects of these problematic text features on survey question comprehension during Web survey completion. If the text features do indeed undermine the survey response process at the comprehension stage, then this should show up in the eye-tracking record in the form of longer fixations and larger numbers of fixations. Our reasoning is based on two common assumptions about eye movements. The first assumption is that the eye remains fixated on a word as long as it is being processed (*eye-mind assumption*, Just & Carpenter, 1980). Thus, there is a direct link between the time spent fixating on a word and its comprehensibility: difficult words require longer fixations. Second, when larger regions of text such as phrases, clauses, or sentences are difficult to understand, readers are likely to re-fixate earlier words in order to re-read unclear parts of the text, resulting in more fixations on the text (*selective reanalysis hypothesis*, Frazier & Rayner, 1982). Adopting these two assumptions, we examined whether people fixate longer on questions that include a text feature and require more fixations to process the questions.

Respondents were randomly assigned to one of two versions of a Web survey: One group (n = 22) received questions which contained one text feature (text feature condition) and the other group (n = 22) received control questions which did not contain the text feature (control condition). Dependent variables were word/phrase fixation time, question fixation count, and question fixation time as indicators of question comprehensibility. Following our argumentation that the problematic text features presented above induce comprehension

difficulties, we hypothesized that respondents would fixate longer on the specific text feature words/phrases, require more fixations to process the question stems, and fixate longer on the question stems in the text feature condition compared to the control condition (Hypothesis 1). Whereas previous research has focused only on attitudinal questions, we expected to identify these effects for factual and behavioral questions as well, and thus independent of question type (Hypothesis 2).

Participants

The study was conducted in June and July 2009 at the Max Planck Institute for Human Development in Berlin, Germany. 49 participants were recruited from the respondent pool maintained by the institute. Technical difficulties made it impossible to accurately record the eye movements of one respondent wearing very thick glasses and another respondent dropped out from the study because of illness. In addition, the recordings of three respondents were of unsatisfactory quality displaying a systematic shift to the line below the one that was fixated. These three recordings were excluded from the analyses, leaving 44 respondents (22 in each condition) in the experiment. Sixty-one percent (27) of the participants were female and all were between 19 and 34 years of age with a mean age of 26 (SD = 3.7). All participants had at least 12 years of schooling and 68 percent (30) were currently enrolled as university students. The native language of all participants was German (the language in which the questionnaires were designed).

Apparatus

Participants' eye movements were recorded by a Tobii T120 Eye Tracker and analyzed with the Tobii Studio 2.0.3 software. In the T120 system the eye-tracking cameras are integrated into a 17" TFT monitor allowing for unobtrusive recording of respondents' eye movements. The documentation of the T120 describes its accuracy to be within 0.5° with less than 0.3°

drift over time and less than 1° due to head motion. It allows for head movement within a 30 x 22 x 30 cm volume centered up to 70 cm from the camera. The sampling rate is 120 Hz, meaning that 120 gaze data points per second are collected for each eye. The accuracy of the T120 was found to be generally sufficient to determine on which words respondents fixate. However, to make sure that all fixations were unequivocally allocated to the words respondents had actually read, we used a larger font size of 18 pixels and double-spaced text with a line height of 50 pixels. Screen resolution was set to 1280 by 1024. In our analyses, we included all fixations that lasted at least 100 milliseconds and encompassed 20 pixels (about four characters of text) in the Web surveys (see Galesic, Tourangeau, Couper, & Conrad, 2008, for similar methodology).

Questions

The questionnaires in both conditions included 28 experimental questions on a variety of topics such as social inequality, environment, health, leisure time, and citizenship. With the exception of one question that was designed by the first author (Q10), the questions were adapted from various existing surveys, such as the International Social Survey Programme (ISSP), the German General Social Survey (ALLBUS), and the German Socio-Economic Panel (GSOEP). Each of the seven text features was operationalized by a set of four questions (two attitudinal, one factual, and one behavioral question). We created two versions of each question by manipulating the characteristic of one text feature according to the rewriting rules reported in Lenzner et al. (2010). The language of the questionnaire was German. Both the German questionnaires as well as a translation of the questionnaires in English are available from the authors on request.

Procedure

The randomized experiment was part of a larger study with several unrelated experiments. The whole study took about two hours of which one hour was devoted to eye tracking. Respondents in this experiment completed the Web survey in about 10 minutes during the first hour of the study. As calibration could decrease in accuracy over time, respondents were recalibrated every 10 to 15 minutes. This was done by a technical assistant who was present in the same room as the respondent during data collection and ensured adherence to the procedure. The technical assistant was seated at a table next to the respondent and was monitoring his or her eye movements on a separate computer monitor. Respondents were seated in front of the eye tracker so that their eyes were approximately 60 cm from the screen. They were instructed to read at a normal pace while trying to understand the questions as well as they could. After participants had successfully completed a standardized calibration procedure, they were presented with the welcome page of the Web survey.

Only one question at a time was displayed on the screen. First, participants answered three questions of different length which were identical in both conditions. These were used to compute the individual reading rate and the fixation rate for every respondent, which were later used as covariates in the analyses to control for interindividual differences. Second, they received the 28 text feature questions or control questions in a random sequence to control for context effects and effects of the position of the questions in the questionnaire. Finally, they answered a series of background questions on sex, age, education, and their native language. After they had completed the survey, the technical assistant recalibrated the eye tracker and started the next experiment. For their participation in the eye tracking part of the study, respondents received a compensation of 10 Euros. The study was approved by the Ethics Committee at the Max Planck Institute for Human Development, where the study was conducted.

RESULTS

The results of this experiment will be reported in terms of word/phrase fixation time, question fixation count, and question fixation time. *Word/phrase fixation time* refers to the total duration of fixations on a specific text feature (e.g. a low-frequency word or an ambiguous noun-phrase), including re-readings of these features. *Question fixation count* refers to the sum of fixations respondents made on the question stem (excluding the answer options), again including re-readings. *Question fixation time* corresponds to the total duration respondents fixated on the question stem (excluding the answer options). These three measures are commonly used to investigate processing difficulty in both word recognition (word/phrase fixation time) and higher-order comprehension processes (question fixation count and question fixation time; cf. Rayner & Pollatsek, 2006). Question fixation counts and question fixation times only included fixations on the question stem (excluding the answer options), because we were primarily interested in examining the comprehension stage of the response process. During the comprehension stage, respondents usually fixate on the question stem to find out what the question is about, and hence any comprehension difficulties should show up in the form of longer and higher numbers of fixations in this region. In contrast, while carrying out the remaining tasks of the response process (information retrieval, judgment, formatting, and editing), respondents are more likely to fixate on the answer options. Given that longer fixation times on the answer options can either reflect difficulties in performing these tasks or an optimizing response style, we excluded all fixations on the answers from our analyses.

Ideally, it would have been the case that the specific text feature words or phrases as well as the questions consisted of the same number of characters and the same number of words, respectively. However, in our experiment this was not possible without constructing very artificial questions that respondents would not normally encounter in the real world.

Following the recommendations of Ferreira and Clifton (1986), we corrected for differences of word/phrase length and question length between the two question versions by dividing all three eye-tracking parameters by the number of characters in the words/phrases and questions (including character spaces and punctuation marks). Hence, word/phrase fixation times, question fixation counts, and question fixation times *per character* are reported in our results.

Text Features

The effect for each text feature was analyzed in separate general linear models with repeated measures on the four questions per text feature and reading rate or fixation rate as a covariate, respectively. Reading rate and fixation rate were computed from respondents' fixations on three introductory questions. Reading rate refers to the average question fixation time for these three questions; fixation rate refers to the average question fixation count for the three questions. We controlled for reading rate and fixation rate because both account for most of the differences between respondents' fixation times and numbers of fixations. The correlation between reading rate and the total fixation time for all 28 questions was $r = .80$. The correlation between fixation rate and total number of fixations for all 28 questions was $r = .72$. Reading rate was used as a covariate in analyses of word/phrase fixation times and question fixation times; fixation rate was used as a covariate in analyses of question fixation counts.

Supporting our Hypothesis 1, six out of seven text features were found to undermine survey question comprehension as indicated by the three eye-tracking measures (see table 1). First, *word/phrase fixation times* were longer in the text feature condition than in the control condition, indicating that these words were difficult for respondents to comprehend. Statistically significant effects were found for low-frequency words [$F(1, 41) = 21.25, p = .0001, \text{partial } \eta^2 = .34$], vague or imprecise relative terms [$F(1, 41) = 14.19, p = .001, \text{partial}$

$\eta^2 = .26$], vague or ambiguous noun-phrases [$F(1, 41) = 8.60, p = .005, \text{partial } \eta^2 = .17$], complex syntax [$F(1, 41) = 8.42, p = .006, \text{partial } \eta^2 = .17$], complex logical structures [$F(1, 41) = 14.90, p = .0001, \text{partial } \eta^2 = .27$], and low-syntactic redundancy [$F(1, 41) = 8.40, p = .006, \text{partial } \eta^2 = .17$]. No significant effects were found for bridging inferences [$F(1, 41) = 0.07, p = .787, \text{partial } \eta^2 = .00$].

Similarly, the *question fixation count* was higher when respondents answered text feature questions, indicating that understanding the question text required re-reading some parts of the question. Again, statistically significant effects were found for low-frequency words [$F(1, 41) = 14.14, p = .001, \text{partial } \eta^2 = .26$], vague or imprecise relative terms [$F(1, 41) = 14.58, p = .000, \text{partial } \eta^2 = .26$], vague or ambiguous noun-phrases [$F(1, 41) = 8.96, p = .005, \text{partial } \eta^2 = .18$], complex syntax [$F(1, 41) = 9.91, p = .005, \text{partial } \eta^2 = .18$], complex logical structures [$F(1, 41) = 12.01, p = .001, \text{partial } \eta^2 = .23$], and low syntactic redundancy [$F(1, 41) = 5.74, p = .021, \text{partial } \eta^2 = .12$]. There was no significant effect of bridging inferences [$F(1, 41) = 0.08, p = .783, \text{partial } \eta^2 = .00$] on the number of fixations respondents made on the question text.

Finally, *question fixation times* were longer in the text feature condition compared to the control. Similar to the other two eye-tracking parameters the text feature effects were significant for low-frequency words [$F(1, 41) = 17.66, p = .0001, \text{partial } \eta^2 = .30$], vague or imprecise relative terms [$F(1, 41) = 15.77, p = .0001, \text{partial } \eta^2 = .28$], vague or ambiguous noun-phrases [$F(1, 41) = 8.49, p = .006, \text{partial } \eta^2 = .17$], complex syntax [$F(1, 41) = 13.21, p = .001, \text{partial } \eta^2 = .24$], complex logical structures [$F(1, 41) = 12.87, p = .001, \text{partial } \eta^2 = .24$], and low syntactic redundancy [$F(1, 41) = 4.94, p = .032, \text{partial } \eta^2 = .11$]. No significant effects were found for bridging inferences [$F(1, 41) = 0.00, p = .956, \text{partial } \eta^2 = .00$].

-----Table 1-----

Question Types

After having analyzed the effects for each text feature we examined whether these effects were different for different question types. Each of the seven text features was operationalized with two attitudinal, one factual, and one behavioral question. For two questions, the distinction of question type was a little bit fuzzy. As a result of the text feature manipulation, the hypothetical questions Q17 and Q19 were attitudinal questions in the control condition but could have been conceived as either behavioral or attitudinal questions in the text feature condition. We treated Q17 as an attitudinal question and Q19 as a behavioral question. However, we also analyzed these two questions as if they were other question types but all of our conclusions remained unchanged.

For all three question types we observed similar patterns (see table 2). First, respondents had longer *word/phrase fixation times* when answering text feature questions compared to control questions. In analyses of variance with repeated measures on the individual questions per question type and reading rate as a covariate, the between-subjects effects were significant for attitudinal [$F(1, 41) = 24.01, p = .0001, \text{partial } \eta^2 = .37$], factual [$F(1, 41) = 10.83, p = .002, \text{partial } \eta^2 = .21$], and behavioral questions [$F(1, 41) = 17.57, p = .0001, \text{partial } \eta^2 = .30$]. Second, the *question fixation count* was significantly higher for the three question types when respondents answered text feature questions [attitudinal: $F(1, 41) = 9.14, p = .004, \text{partial } \eta^2 = .18$; factual: $F(1, 41) = 9.20, p = .004, \text{partial } \eta^2 = .18$; behavioral: $F(1, 41) = 21.98, p = .0001, \text{partial } \eta^2 = .35$]. And finally, *question fixation times* were significantly longer in the text feature condition for all three question types [attitudinal: $F(1, 41) = 9.54, p = .004, \text{partial } \eta^2 = .19$; factual: $F(1, 41) = 10.56, p = .002, \text{partial } \eta^2 = .21$; behavioral: $F(1, 41) = 26.92, p = .0001, \text{partial } \eta^2 = .40$].

-----Table 2-----

DISCUSSION AND CONCLUSION

Extending earlier research by Graesser et al. (2006) and Lenzner et al. (2010), this study examined whether survey question comprehension is impeded by seven psycholinguistic text features and whether these text features have different effects for different question types (attitudinal, factual, and behavioral questions). Using eye-tracking methodology, we examined word/phrase fixation times, question fixation counts, and question fixation times while respondents answered two versions of similar questions (text feature version vs. control) in a Web survey.

We found strong evidence that at least six of these text features reduce question comprehensibility and undermine the survey response process at the comprehension stage. Respondents had longer fixation times and needed more fixations in the text feature questions than in the control questions, indicating that processing of these questions required additional cognitive effort. Significant effects were found for low-frequency words, vague or imprecise relative terms, vague or ambiguous noun-phrases, complex syntax, complex logical structures, and low syntactic redundancy. Only bridging inferences were not found to have a detrimental effect on question comprehensibility. In general, bridging inferences are drawn in order to establish coherence between implicit information from an introductory sentence and explicit information from the actual question. The purpose of the introductory sentences is usually to provide a context for the questions, however, understanding (or even reading) these sentences is not a prerequisite for answering the questions (i.e. introductory sentences do not necessarily determine the question focus). Hence, establishing coherence between introductory sentence and actual question is mostly optional rather than mandatory. Our results indicate that

bridging inferences may only undermine question comprehension if the introductory sentence contains implicit information which is crucial for understanding and answering the question.

We also found support for our second hypothesis that those text features that negatively affect question comprehension do so independent of question type. Similar effects were found for attitudinal, factual, and behavioral questions, namely that respondents required longer fixation times and more fixations when these questions contained a text feature. Hence, the text feature effects can be generalized to all three types of questions.

There are two limitations to this study. First, our experiment does not examine whether these text features reduce the quality of responses. While we know that answering questions including the text features requires more time, it is still unclear whether this additional cognitive effort leads to an increase in measurement error. Earlier studies found some negative effects of the text features on response quality (e.g. that they produce more midpoint responses; Lenzner et al., 2010), however, further research is needed to systematically assess their influence on data quality. Second, our sample overrepresents higher educated individuals and therefore is by no means representative of the general population. However, assuming that our participants were comparatively good readers, the text feature effects may even be larger among poorer readers. Hence, we would argue that we can very likely generalize our findings to the broader population.

With regard to the practical implications of using eye-tracking methodology for evaluating survey questions, it is important to note that the interpretation of fixation times and counts is by no means definite. Long fixation times and high numbers of fixations are not problematic per se, but may also indicate an increasing interest in the question or a more conscientious response style. For example, optimizing respondents may require considerable time to select the “optimal” response among the answer options offered. While retrieving relevant information, making a judgment, and formatting and editing the answer, these

respondents would fixate longer on the answer options, resulting in a relatively large fixation time on the question as a whole. Hence, the interpretation of fixation times on the answer options and on the question as a whole is very complicated and rather speculative.

By contrast, it is much easier to interpret fixation times and counts on the question stem, excluding the answer options. Respondents fixate on the question stem while trying to understand what the question is about and usually they turn to the answer options as soon as they have retrieved the question's meaning. If the question is incomprehensible, respondents require more time to interpret it and require more fixations to re-read parts of the question to resolve uncertainties (see Rayner, 1998, for a general overview of eye-tracking measures and their interpretation). Thus, comprehension difficulties occurring in survey questions should become apparent in longer and higher numbers of fixations on the question stem. After all, there is no reason why respondents should fixate on this region after having retrieved its meaning (unless something remains unclear). In sum, eye-tracking methodology currently allows us to detect problems occurring at the comprehension stage of the response process only.

One conclusion that we can derive from our results is that the text feature *low syntactic redundancy* should be included into QUAID, given that it was found to reduce question comprehensibility. An extension of QUAID's five components would increase the validity of this tool in identifying questions that are difficult for respondents to comprehend. Second, we advise survey designers to avoid these text features when designing survey questions. Given that previous work on questionnaire design mostly neglected the importance of psycholinguistic text features for question comprehensibility, it may be fruitful to develop manuals that describe these text features in detail. These manuals may supplement the existing guidelines on survey question design, lend further precision to these rules, and help practitioners to improve the comprehensibility of their questions.

ACKNOWLEDGEMENTS

We are grateful to Gregor Caregnato, Christian Elsner, and Mona Merkert, for their help in conducting this study.

REFERENCES

- Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, 27, 429–446.
- Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, 25, 348–368.
- Foddy, W. (1993). *Constructing questions for interviews and questionnaires: Theory and practice in social research*. Cambridge: Cambridge University Press.
- Fowler, F. J. (1992). How unclear terms affect survey data. *Public Opinion Quarterly*, 56, 218–231.
- Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14, 178–210.
- Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics*, 22, 313–328.
- Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, 72, 892–913.
- Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question understanding aid (QUAID). A web facility that tests question comprehensibility. *Public Opinion Quarterly*, 70, 3–22.
- Haviland, S. E., & Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, 13, 512–521.
- Horning, A. S. (1979). On defining redundancy in language: Case notes. *Journal of Reading*, 22, 312–322.

- Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, 40, 431–439.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329–354.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99, 122–149.
- Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition*, 2, 15–47.
- Kintsch, W., & Keenan, J. M. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 5, 257–279.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213–236.
- Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology*, 24, 1003–1020.
- Mosier, C. I. (1941). A psychometric study of meaning. *Journal of Social Psychology*, 13, 123–140.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372–422.
- Rayner, K., & Pollatsek, A. (2006). Eye-movement control in reading. In M. J. Traxler, & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (Vol. 2, pp. 613–658). Amsterdam: Elsevier.
- Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, NJ: Wiley.

- Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61, 576–602.
- Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. San Diego, CA: Academic Press.
- Smith, T. W. (1987). That which we call welfare by any other name would smell sweeter. *Public Opinion Quarterly*, 51, 75–83.
- Strack, F., & Martin, L. L. (1987). Thinking, judging and communicating: A process account of context effects in attitude surveys. In H. J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 123–148). New York: Springer.
- Tourangeau, R. (1984). Cognitive science and survey methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73–100). Washington, DC: National Academy Press.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.
- University of Memphis (2010). *Question Understanding Aid*. Retrieved 18 October, 2010 from <http://mnemosyne.csl.psyc.memphis.edu/QUAID/quaidindex.html>.

Table 1. Mean word/phrase fixation time, question fixation count, and question fixation time for text feature versions and controls

	Word/phrase fixation time	Question fixation count	Question fixation time
Low-frequency words	654	0.92	212
Control	155	0.59	121
Vague or imprecise relative terms	240	0.86	188
Control	140	0.58	119
Vague or ambiguous noun-phrases	413	0.70	151
Control	242	0.54	112
Complex syntax	308	0.82	183
Control	202	0.61	124
Complex logical structures	154	0.81	175
Control	99	0.61	125
Low syntactic redundancy	203	0.70	152
Control	134	0.55	116
Bridging inferences	191	0.69	146
Control	168	0.65	136

Note: Fixation times are reported in milliseconds. To control for differences of word/phrase or question length between the two question versions, we divided all three eye-tracking parameters by the number of characters in the question. Hence, word/phrase fixation times, question fixation counts, and question fixation times *per character* are reported here. Question fixation counts and question fixation times only refer to fixations on the question text, excluding fixations on answer options.

Table 2. Mean word/phrase fixation time, question fixation count, and question fixation time for text feature versions and controls by question type

	Word/phrase fixation time	Question fixation count	Question fixation time
<u>Attitudinal (n=14)</u>			
Text feature questions	1091	2.69	582
Control	604	2.14	446
<u>Factual (n=7)</u>			
Text feature questions	588	1.42	315
Control	261	1.02	208
<u>Behavioral (n=7)</u>			
Text feature questions	480	1.40	308
Control	268	0.96	199

Note: Fixation times are reported in milliseconds. To control for differences of word/phrase or question length between the two question versions, we divided all three eye-tracking parameters by the number of characters in the question. Hence, word/phrase fixation times, question fixation counts, and question fixation times *per character* are reported here. Question fixation counts and question fixation times only refer to fixations on the question text, excluding fixations on answer options.