How Effective Are Eye-Tracking Data in Identifying Problematic Questions?

Neuert, Cornelia

Social Science Computer Review
2020, Vol. 38(6) 793-802

DOI: 10.1177/0894439319834289
journals.sagepub.com/home/ssc

How Effective Are Eye-Tracking Data in Identifying Problematic Questions?

Cornelia E. Neuert[1]

Abstract

To collect high-quality data, survey designers aim to develop questions that each respondent can understand as intended. A critical step to this end is designing questions that minimize the respondents' burden by reducing the cognitive effort required to comprehend and answer them. One promising technique for identifying problematic survey questions is eye tracking. This article investigates the potential of eye movements and pupil dilations as indicators for evaluating survey questions. Respondents were randomly assigned to either a problematic or an improved version of six experimental questions. By analyzing fixation times, fixation counts, and pupil diameters, it was examined whether these parameters could be used to distinguish between the two versions. Identifying the improved version worked best by comparing fixation times, whereas in most cases, it was not possible to differentiate between versions on the basis of pupil data. Limitations and practical implications of the findings are discussed.

Keywords

eye tracking, response behavior, pupillometry, data quality

Introduction and Background

Survey questions should produce data that are valid, reliable, and unbiased (Fowler, 2013). A critical step toward collecting high-quality data and reducing measurement error is to design survey questions in such a way that each respondent comprehends the question and understands it as the researcher intended (Conrad & Schober, 2000). When responding to survey questions, respondents are required to perform a series of complex cognitive processes (Tourangeau, Rips, & Rasinski, 2000): Respondents must comprehend the question, recall relevant information, make use of the information to form a judgment, and answer the question by selecting a response. Accurate responses can only be expected when respondents move thoroughly through all four steps of question answering (referred to as "optimizing," in contrast to "satisficing respondent behavior," Krosnick, 1999). How much cognitive effort respondents are willing to invest at each of the four

[1] GESIS—Leibniz Institute for the Social Sciences, Mannheim, Germany

Corresponding Author:
Cornelia E. Neuert, GESIS—Leibniz Institute for the Social Sciences, B2 1, Mannheim 68159, Germany.
Email: cornelia.neuert@gesis.org

stages, and the likelihood of satisficing, depends on the difficulty and complexity of the task involved (e.g., question difficulty), respondents' cognitive ability, and respondents' motivation (Biemer & Lyberg, 2003; Krosnick, 1991). Hence, one way to minimize respondents' contribution to measurement error is to reduce the respondents' burden, thus minimizing the chance of respondents adopting response strategies that might affect data quality adversely. This can be achieved by reducing the cognitive effort required to comprehend and answer a survey question (Biemer & Lyberg, 2003). Therefore, survey researchers have to check for potential cognitive hurdles and their underlying causes by evaluating their draft questions (Fowler, 2013; Miller, 2014). There is a large variety of question testing tools available, such as cognitive interviews, response latency measurement, expert reviews (Presser et al., 2004), or paradata such as mouse movements (Horwitz, Kreuter, & Conrad, 2017). Additionally, the analysis of eye-movement data is an apparently promising technique for identifying problematic survey questions (Kamoen, Holleman, Mak, Sanders, & Van Den Bergh, 2017; Neuert & Lenzner, 2016). There is a strong case for a link between eye movements and cognitive processing (e.g., Just & Carpenter, 1980; Liversedge & Findlay, 2000; Rayner, 1998, 2009). Eye tracking enables the researcher to see where and how long respondents look when reading and answering survey questions. This feature can be used to detect questions that are difficult to understand or otherwise flawed (Galesic & Yan, 2011). Typically, while reading and answering survey questions, respondents go back and refixate on words that are more difficult to comprehend, which means that longer fixation duration and a higher number of fixation counts indicate increased cognitive effort (Rayner, 1998). Hence, when evaluating questions, the question wording that produces shorter and fewer fixations is considered to be less difficult to answer. Graesser, Cai, Louwerse, and Daniel (2006) found that questions containing difficult text features like unfamiliar technical terms had longer total fixation times and more fixation counts than questions containing words that were defined to be nontechnical terms. Lenzner, Kaczmirek, and Galesic (2011) added to this line of research and found that respondents had longer fixation times and more fixation counts when answering questions containing problematic text features (e.g., low-frequency words) compared to control versions of these questions, indicating higher cognitive effort. Eye-tracking methodology was also used by Kamoen, Holleman, Mak, Sanders, and Van Den Bergh (2017) to examine the cognitive processes involved in answering contrastive survey questions. The results revealed that negatively worded questions were reread longer and more frequently than positively formulated questions.

Besides the eye-tracking metrics typically used to measure cognitive processing, such as fixation times and number of fixations, there is the measure of pupil dilation (Beatty & Lucero-Wagoner, 2000; Kruger, Hefer, & Matthew, 2013). Pupils not only narrow in response to light and dilate in response to darkness, they also dilate as a function of cognitive processing and task difficulty (Beatty & Lucero-Wagoner, 2000; Iqbal, Zheng, & Bailey, 2004; Laeng, Sirois, & Gredebäck, 2012). Hence, pupil dilation can be used as a measure of the cognitive effort needed when performing a task (Beatty, 1982). The increase in pupil size is involuntary and rather small but large enough to determine a "task-evoked pupillary response" (Beatty, 1982; Beatty & Lucero-Wagoner, 2000).

Two studies in the 1960s were among the first that established the link between pupillary dilation and cognitive effort (Beatty & Kahnemann, 1966; Hess & Polt, 1964). Hess and Polt (1964) asked participants to solve multiplication tasks, while their eyes were recorded with a camera. The evaluation of the videos showed that the pupils widened as the participants pondered over their problem. As soon as the solution was found, the pupils immediately returned to their normal size. Since then, it has been shown in several tasks that pupils dilate depending on the difficulty of the task, for instance, in digit sorting (Siegle, Steinhauer, Stenger, Konecky, & Carter, 2003), sentence comprehension (Just & Carpenter, 1993), and visual search (Porter, Troscianko, & Gilchrist, 2007). This characteristic could make pupil dilation a valuable tool for measuring the intensity of cognitive processing while responding to web surveys. Yan, Williams, Maitland and Tourangeau (2016) used pupil

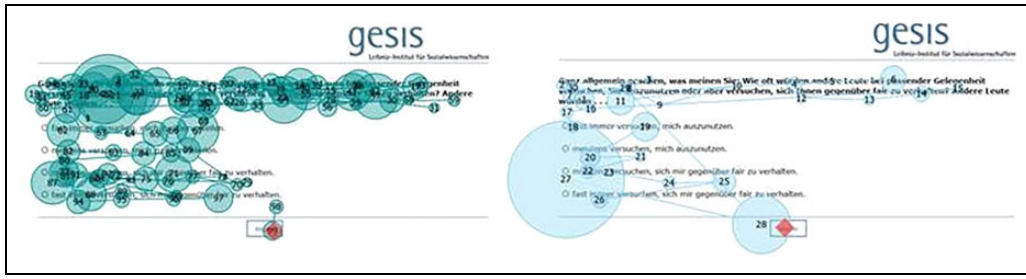

Figure 1. Sample gaze plots of different respondents answering the problematic (left) and the improved version (right) of Experiment 2. Each gaze plot displays the eye movements of one respondent. The circles indicate fixations and the saccades are plotted as connecting lines in-between. The numbers within the circles indicate the order of the fixations and the circle radius relates to the fixation time.

dilation as an indicator for response burden in surveys and compared it to respondents' self-reports. The authors found more pupil dilation for questions which were rated as harder by respondents.

Pupil diameters vary between 1.5 mm in bright light and 9 mm in total darkness. By recording respondents' eyes while controlling for other factors that might affect pupil size, like brightness or color, the task-evoked pupil response can be determined; it is usually smaller than 0.5 mm (Sirois & Brisson, 2014). This article examines the potential of eye movements and pupil dilation as indicators for evaluating survey question difficulty and examines whether one of the measures is a better predictor.

Research Design and Hypotheses

In a laboratory experiment ($N = 131$), eye tracking was used to investigate whether eye movements and pupil dilation can be implemented to evaluate questions while respondents complete a web survey. Each of the six experiments consisted of question pairs, each comprising one problematic and one improved version. Respondents were randomly assigned to one of the experimental comparisons (experimental group size is shown in Online Appendix A). Following the argumentation that difficult questions produce longer and more fixations, it is hypothesized that respondents would fixate longer on the problematic versions compared to the improved versions (Hypothesis 1a) and that the problematic versions would require more fixations (Hypothesis 1b; see Figure 1 for a sample illustration). As pupil size increases with cognitive effort, it is hypothesized that respondents' pupils would dilate more when answering the problematic questions compared to the improved versions (Hypothesis 3).

Method

Participants

In total, 131 respondents, 50% female ($n = 65$), were recruited. Of these participants, 38% were between 18 and 24 years old, 38% were between 25 and 44 years old, 19% were between 45 and 64 years old, and 5% were 65 years or older; 7% had graduated from a lower secondary school, 22% from an intermediate secondary school, and 71% from a college preparatory secondary school or university. All of the participants came between April and May 2017 to the pretest lab at GESIS—Leibniz Institute for the Social Sciences (Mannheim, Germany) to take part in the study. Each respondent received a compensation of €20.

Survey Questions

Six experiments containing nine question pairs in total were selected (four single questions and two grid questions with 2 and 3 items each). The questions for this study were chosen mainly from textbooks on questionnaire design and the GESIS pretest database (http://pretest.gesis.org/). Prerequisites for inclusion in the questionnaire were that an improved version was reported besides the problematic wording (and not just a description of the question's problems) and that the questions could be administered to the general population. From textbooks, three question pairs (Experiment 3 and Experiment 4 with 2 items) were selected (Fowler, 2001; Porst, 2011). Five question pairs (Experiment 1 with 3 items, Experiment 5, Experiment 6) were chosen from pretest reports (Lenzner, Neuert, & Otto, 2014; Lenzner et al., 2015; Lenzner et al., 2016). One additional question (Experiment 2) was adapted from the International Social Survey Programme (ISSP). This question was selected because it already has been asked in several rounds (ISSP, 2012, 2016) in two different wordings which differed in complexity by using either a low frequency or a more common verb. The language of the questionnaire was German. The English translations of the questions can be found in Online Appendix C. Respondents were randomly assigned to one or the other experimental question for each experiment resulting in varying treatment and control group compositions and sizes across the six experiments (see Online Appendix A). The order of questions in the survey was fixed. After the experimental questions, the questionnaire contained two questions that measured questionnaire difficulty (see Online Appendix B).

Eye-Tracking Equipment

Eye movements were recorded with a Senso Motoric Instruments (SMI) RED250 mobile Eye Tracker, which is a mobile device that was mounted on the bottom frame of a 17″ TFT monitor (resolution 1,280 × 1,024). The system is typically accurate within 0.4° and has a resolution of 0.3°. It permits head movements within a range of 32 × 21 cm at 60 cm distance. Eye movements were recorded at a sampling rate of 250 Hz (every 4 ms). For gaze analysis, the SMI BeGaze Version 3.6.57 was utilized. Pupil diameter is measured automatically by the SMI RED250 eye-tracking system and provided in millimeters. Data were sampled binocularly, but for the analysis of pupil data, only data from the right eye were used. The online questionnaire was programmed with a font size of 16 pixels and double-spaced text with a line height of 40 and 32 pixels for the question text and response categories, respectively. The questions appeared as black text on a white background. This ensured that any changes in pupil diameter could not be attributed to luminance.

Procedure

All respondents were tested individually in a moderately lit room. After they had completed a calibration exercise, the respondents answered the web survey. At the beginning of the questionnaire, all respondents answered two introductory questions, which were used to calculate their individual fixation rate, reading rate, and baseline pupil diameter.[1] Individual fixation rate and reading rate were later used as covariates in the statistical analyses of fixation count and fixation time to control for interindividual differences. The questionnaire, which included several other experiments, took about 30 min to complete. During this time, the experimenter stayed in the room next door to be able to intervene in case of technical difficulties and to monitor the eye movements on a second screen.

Analytical Strategies

To examine how much cognitive effort and attention is invested in answering the questions, respondents' fixation time and fixation count were considered within "areas of interest" (AOIs), which

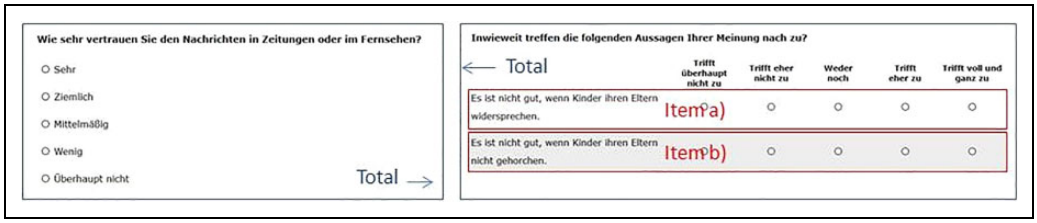

Figure 2. Areas of interest for the analysis of fixations in single (left) and grid questions (right).

fully covered the question and the response options. In case of a grid question with multiple items, each item was covered by a separate AOI (see Figure 2). Fixation time is the total duration of fixations on an AOI. Fixation count is the total number of fixations on an AOI. To analyze the differences in fixation time and fixation count across question versions, separate analyses of covariance were employed with reading rate and fixation rate as covariates, respectively.

To analyze pupil dilation, pupil data were cleaned by identifying and removing dilation values reflecting measurement errors or short gaps of missing data. Baseline diameter was then subtracted from peak dilation[2] for each experimental question (Beatty & Lucero-Wagoner, 2000; Yan, Williams, Maitland, & Tourangeau, 2016). The baseline pupil diameter for each respondent was established while reading the instruction screen and answering the first two introductory questions of the questionnaire and calculated as the average pupil size. The higher the peak dilation after subtracting the baseline diameter, the more cognitive effort is likely to have been required. To analyze differences in peak dilation across question versions, separate analyses of covariance were employed.

In some instances, the problematic and improved question versions differed with regard to the number of words. To account for the differences in length between question versions, fixation time and number of fixations are divided by the number of characters (Höhne & Lenzner, 2018). Hence, fixation time and number of fixations per character are reported in the results, thus ensuring comparability regardless of the length of question.

Data and Manipulation Checks

To evaluate possible differences in the sample composition between the two question versions of all six experiments, χ^2 tests were conducted. The results showed no statistically significant differences for age and education. However, the distribution of women and men was not equal in Experiment 1 and Experiment 2, with more women than men in the "problematic" condition in Experiment 1 and fewer women in this condition in Experiment 2 (Information on sociodemographic characteristics and results of χ^2 tests can be found in Online Appendix D). Therefore, the variable sex was included in the analyses as a covariate.

The quality of the eye-tracking data was also checked before analyzing the data. Due to technical difficulties or shifts in the recordings, the eye movements between 7 and 14 respondents per experiment were excluded from the analysis.

Results

Fixation Time and Fixation Count

When considering *fixation time*, significant differences were found for Items a and c in Experiment 1, for Experiment 2, for Item a in Experiment 4, and for Experiment 5, supporting Hypothesis 1a. For all other question pairs, except Item b in Experiment 1 and Item b in Experiment 4, fixation times were longer in the problematic than in the improved question version, but these differences were not

Table 1. Mean Fixation Time and Number of Fixations Between Problematic Versus Improved Versions for Each Experimental Question Pair.

	Fixation Time				Number of Fixations (*n*)			
Question Difficulty Indicators	Problematic	Improved	*F* Value ($df_1 = 3/2$)	*p*	Problematic	Improved	*F* Value ($df_1 = 3/2$)	*p*
Experiment 1a	85.99	70.04	3.913†	.050	.31	.27	2.252	.136
Experiment 1b	64.37	68.20	.396	.531	.26	.27	.082	.775
Experiment 1c	88.83	71.14	6.771*	.010	.33	.27	6.016*	.016
Experiment 2	46.68	40.20	2.813†	.096	.18	.17	1.552	.215
Experiment 3	70.39	62.57	2.103	.150	.27	.21	5.889**	.000
Experiment 4a	119.30	89.24	10.172**	.002	.42	.31	13.762**	.000
Experiment 4b	92.29	120.21	4.846*	.030	.41	.42	.066	.798
Experiment 5	155.94	133.90	3.149†	.079	.62	.54	2.663	.105
Experiment 6	54.65	53.96	.020	.889	.19	.21	.472	.493

Note. Fixation times are in milliseconds. Reported are estimated marginal means after controlling for the covariates respondents' reading rate and fixation rate, respectively. Due to the randomly unbalanced distribution of women and men in Experiment 1, Experiment 2, and Experiment 3, gender is included as covariate ($df_1 = 3$; for Experiment 4–6: $df_1 = 2$). To control for differences in the number of word length between question versions, fixation times and fixation counts are divided by the number of characters per question and reported as corrected fixation times and corrected fixation counts per question.
†$p < .1$. *$p < .05$, **$p < .01$.

statistically significant. The results are shown in Table 1. Contrary to Hypothesis 1a, fixation times were significantly longer when answering the improved compared to the problematic question in Experiment 4 Item b.

Similarly, question *fixation count* was significantly higher when respondents answered the problematic versions of Item c in Experiment 1, of Experiment 3, and of Item a in Experiment 4. For Item a in Experiment1, and for Experiment 2 and Experiment 5, respondents fixated more often on the problematic than on the improved question version, but these differences were not statistically significant. Hypothesis 1b—stating that the difficult version would require more fixations—is not supported for Item b in Experiment 1, for Item b in Experiment 4, and for Experiment6, although the results are also not statistically significant.

Pupil Dilation

To investigate Hypothesis 2, it was examined whether the problematic versions of the question pairs would lead to higher peak dilation than the improved versions, indicating more cognitive effort. The results are presented in Table 2. Contrary to the expectations, only the comparison of the two versions of Experiment 4 shows a significant difference in the expected direction, although for most questions, there is a general trend observable that peak dilation is higher for the problematic compared to the improved version. However, for Experiment 3, it is the other way around. Hence, when examining the responses to the different question versions, no differences in the task-evoked pupillary response were found except for Experiment 4 Item a and Experiment 4 Item b.

Discussion

This study was designed to investigate the potential of eye movements and pupil dilation as indicators for evaluating survey questions. The findings, which are summarized in Table 3, show that respondents' fixation times can provide evidence regarding questionnaire difficulty. With regard to pupil dilation, it

Table 2. Peak Dilation Between Problematic Versus Improved Versions for Each Experimental Question Pair.

	Peak Dilation			
Question	Problematic	Improved	*F* Value ($df_1 = 2$)	*p*
Experiment 1a	.112	.109	0.123	.727
Experiment 1b	.105	.088	0.024	.876
Experiment 1c	.042	.028	0.030	.863
Experiment 2	.106	.099	0.006	.937
Experiment 3	.098	.149	3.130†	.079
Experiment 4a	.186	.113	5.909*	.017
Experiment 4b	.142	.087	3.378†	.069
Experiment 5	.117	.117	0.002	.968
Experiment 6	.109	.108	0.004	.951

†$p < .1$. *$p < .05$. **$p < .01$.

Table 3. Summary of the Findings for the Hypotheses.

	Findings		
Hypotheses	Supported	Seemingly Supported	No Supporting Evidence
Hypothesis 1a—Fixation time	Experiment 1a, Experiment 1c, Experiment 2, Experiment 4a, Experiment 5	Experiment 3	Experiment 1b, Experiment 4b, Experiment 6
Hypothesis 1b—Fixation count	Experiment 1c, Experiment 3, Experiment 4a	Experiment 1a, Experiment 5	Experiment 1b, Experiment 2, Experiment 4b, Experiment 6
Hypothesis 2—Pupil dilation	Experiment 4a, Experiment 4b	Experiment 1b, Experiment 1c	Experiment 1a, Experiment 2, Experiment 3, Experiment 5, Experiment 6

was not possible to differentiate between problematic and improved versions for most of the questions—apart from Experiment 4. Most of the differences found were extremely subtle and often not statistically significant. The two problematic item versions of Experiment 4 both include double negations and are therefore syntactically and cognitively very complex (Hoosain, 1973). It is possible that only these questions differed sufficiently in severity to create an observable effect on pupil dilation.

In line with Hypothesis 1a, respondents fixated significantly longer on the problematic versions of Experiment 1a and c, Experiment 2, Experiment 4a, and Experiment 5 and longer on the problematic versions of Experiment 3. However, fixation times were not longer than those for the improved versions on the problematic versions of Experiment 1b, Experiment 4b, and Experiment 6. As with response times, there is no perfect relationship between fixation time and fixation count, on the one hand, and question difficulty on the other (Yan & Tourangeau, 2008). To sum up, the hypothesis that difficult questions produce longer and more numerous fixations could be confirmed for some, but not all, question pairs, while the hypothesis that difficult questions lead to increased pupil dilation could not be confirmed.

However, several potential reasons for these results must be taken into account, as this line of research only begins to explore the value of eye-tracking data for question evaluation and pupillometric data in particular. First, although most of the questions that were selected for this study were taken from textbooks on questionnaire design or from cognitive interviewing projects, it cannot be

ruled out that the difference in problem severity between the two versions was not sufficiently high to discriminate between good and problematic questions. Especially with regard to the question pairs taken from pretesting projects in which the improved questions were revised to minimize interpretation problems, it could be that the improved question better captured the construct to be measured, but that this version was not less cognitively demanding to answer for the respondents. To check whether there is a correlation between the number of problematic questions respondents received and the self-assessment of the severity of the questionnaire, Spearman's rank correlation was calculated. The results show that respondents who were confronted with more problematic questions did not evaluate the questionnaire as more difficult ($r = -.044$, $p = .621$) than respondents who received fewer or no problematic questions (due to randomization). Second, Graesser et al. (2006) found that respondents tended to abandon question processing when they were confronted with questions that had a complex syntax or required a high level of working memory load. The authors interpreted this behavior as early exit strategy. It is possible that respondents in this study also took an "early exit" if they were struggling with a question and thus did not process each item as thoroughly as required. A response strategy of this kind would not be reflected in longer processing time or more fixations. However, the perceived cognitive effort should be observable in peak dilation of the pupil before respondents decide to abandon the question. An indicator which suggests this may be the case is the fact that, while question fixation time and count did not differ notably for Experiment 4b, pupil dilation did.

Therefore, it would be worthwhile attempting to replicate the findings with other questions containing different types of problems, for instance, questions that ask respondents to retrieve information, or evaluating entire questionnaires containing questions with different levels of difficulty.

A third explanation might be that different indicators or parameters are needed for different types of difficulties. It would be worthwhile to examine whether it is possible to define both specific, atypical eye movements and the eye-tracking metrics of fixation time and count and to associate these with specific problem types. If such an association turns out to be possible, one could imagine developing automated support in real time for respondents showing signs of difficulties. It would then be conceivable in future that the person filling out the questionnaire would be asked in a dialog box whether he or she needs help.

Data Availability and Software Information

The SPSS data file containing the eye-tracking data is available by contacting the author at cornelia.neuert@gesis.org. All data analyses were performed using IBM SPSS Advanced Statistics 24. For gaze analysis, the SMI BeGaze Version 3.6.57 was utilized (https://www.smivision.com).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Reading rate refers to the "average fixation time on the experimental questions." Fixation rate refers to the "average number of fixations on these questions" (see Lenzner et al., 2014, p. 751). Baseline diameter refers to the average pupil diameter on these questions plus the introductory screen.

2. Peak dilation is defined as the "maximum dilation obtained in the measurement interval of interest" and was calculated by subtracting baseline diameter from the maximum value per question. Peak dilation is based on a single value but is independent of the number of data points collected; this could be different, as the processing time per question is individual, and therefore, a different number of data points are generated (Beatty & Lucero-Wagoner, 2000).

References

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*, 276–292.

Beatty, J., & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science*, *5*, 371–372.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. *Handbook of Psychophysiology*, *2*, 142–162.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. Hoboken, NJ: John Wiley.

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

Fowler, F. J. (2001). Why it is easy to write bad questions. *ZUMA Nachrichten*, *25*, 49–66.

Fowler, F. J. (2013). *Survey research methods* (5th ed.). Thousand Oaks, CA: Sage.

Galesic, M., & Yan, T. (2011). Use of eye tracking for studying survey response processes. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 349–370). New York, NY: Routledge.

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.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*, 1190–1192.

Höhne, J. K., & Lenzner, T. (2018). New insights on the cognitive processing of agree/disagree and item-specific questions. *Journal of Survey Statistics and Methodology*, *6*, 401–417.

Hoosain, R. (1973). The Processing of Negation. *Journal of Verbal Learning and Verbal Behaviour*, *12*, 618–626.

Horwitz, R., Kreuter, F., & Conrad, F. (2017). Using mouse movements to predict Web survey response difficulty. *Social Science Computer Review*, *35*, 388–405.

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human–computer interaction. Proceedings of the Conference on Human Factors in Computing Systems (CHI'04), Vienna, pp. 1477–1480.

ISSP Research Group. (2012). International social survey programme: Environment III—ISSP 2010. GESIS Data Archive, Cologne. ZA5500 Data File Version 2.0.0. doi:10.4232/1.11418

ISSP Research Group. (2016). International social survey programme: Citizenship II—ISSP 2014. GESIS Data Archive, Cologne. ZA6670 Data File Version 2.0.0. doi:10.4232/1.12590

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. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology*, *47*, 310–339.

Kamoen, N., Holleman, B., Mak, P., Sanders, T., & Van Den Bergh, H. (2017). Why are negative questions difficult to answer? On the processing of linguistic contrasts in surveys. *Public Opinion Quarterly*, *81*, 613–635.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, *5*, 213–236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567.

Kruger, J. L., Hefer, E., & Matthew, G. (2013). Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual texts. In *Proceedings of the 2013 Conference on Eye Tracking South Africa* (pp. 62–66). ACM.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the reconscious? *Perspectives on Psychological Science*, *7*, 18–27.

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*, 361–373.

Lenzner, T., Kaczmirek, L., & Galesic, M. (2014). Left feels right: A usability study on the position of answer boxes in web surveys. *Social Science Computer Review*, *32*, 743–764.

Lenzner, T., Neuert, C., & Otto, W. (2014). German Internet Panel (GIP) Reforms Monitor (2014). Kognitiver Pretest (GESIS Projektbericht). doi:10.17173/pretest20

Lenzner, T., Otto, W., Adams, F., Disch, K., Neuert, C., & Menold, N. (2015). Conceptions of democracy and preferences over democratic procedures. Kognitiver Online-Pretest (GESIS Projektbericht). doi:10.17173/pretest2

Lenzner, T., Otto, W., Neuert, C., Beitz, C., Schmidt, R., & Stiegler, A. (2016). Comparative Study of Electoral Systems (CSES) Module 5. Cognitive Pretest (GESIS Project Report). doi:10.17173/pretest27

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, *4*, 6–14.

Miller, K. (2014). Introduction. In K. Miller, V. Chepp, S. Willson, & J. L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 1–6). New York, NY: John Wiley.

Neuert, C. E., & Lenzner, T. (2016). Incorporating eye tracking into cognitive interviewing to pretest survey questions. *International Journal of Social Research Methodology*, *19*, 501–519.

Porst, R. (2011). *Fragebogen. Ein Arbeitsbuch* [Questionnaire. A workbook]. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and counting: Insights from pupillometry. *Quarterly Journal of Experimental Psychology*, *60*, 211–229.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly*, *68*, 109–130.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*, 1457–1506.

Siegle, G. J., Steinhauer, S. R., Stenger, V. A., Konecky, R., & Carter, C. S. (2003). Use of concurrent pupil dilation assessment to inform interpretation and analysis of fMRI data. *Neuroimage*, *20*, 114–124.

Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*, 679–692.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, England: Cambridge University Press.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, *22*, 51–68.

Yan, T., Williams, D., Maitland, A., & Tourangeau, R. (2016, May). Use of eye-tracking to measure response Burden. Paper presented at 2016 AAPOR Annual Conference, Austin, Texas.

Author Biography

Cornelia E. Neuert is a senior researcher at the GESIS Pretest Lab, GESIS—Leibniz Institute for the Social Sciences. Her research focuses on questionnaire design and evaluation and eye tracking. She can be reached at cornelia.neuert@gesis.org.