Incorporating eye tracking into cognitive interviewing to pretest survey questions

Neuert, Cornelia; Lenzner, Timo

This is an Accepted Manuscript of an article published by Taylor & Francis in
International Journal of Social Research Methodology, Vol. 19, 2016, No. 5, pp. 501-533
available online: http://www.tandfonline.com/10.1080/13645579.2015.1049448

Incorporating eye tracking into cognitive interviewing to pretest survey questions

Cornelia Eva Neuert* and Timo Lenzner

Survey Design and Methodology, GESIS – Leibniz Institute for the Social Sciences, PO. Box 12 21 55, 68072 Mannheim, Germany (Received 3 September 2014; accepted 6 May 2015)

In this study, we investigated whether incorporating eye tracking into cognitive interviewing is effective when pretesting survey questions. In the control condition, a cognitive interview was conducted using a standardized interview protocol that included pre-defined probing questions for about one-quarter of the questions in a 52-item questionnaire. In the experimental condition, participants' eye movements were tracked while they completed an online version of the questionnaire. Simultaneously, their reading patterns were monitored for evidence of response problems. Afterward, a cognitive interview was conducted using an interview protocol identical to that in the control condition. We compared both approaches with regard to the number and types of problems they detected. We found support for our hypothesis that cognitive interviewing and eye tracking complement each other effectively. As expected, the hybrid method was more productive in identifying both questionnaire problems and problematic questions than applying cognitive interviewing alone.

Keywords: cognitive interviews; eye tracking; pretesting; survey questions

Introduction

Questionnaires are the most commonly used tools in the social sciences for collecting data about people's attitudes, values, and behaviors (Groves et al., 2004). To ensure that the data gathered through questionnaires are of high quality, researchers must formulate questions that are easily and consistently interpreted by respondents in the ways intended by the researchers (Collins, 2003; Fowler, 1995). This reasoning is based on the underlying assumption that ‘questions that are easily understood and that produce few other cognitive problems for the respondents introduce less measurement error than questions that are hard to understand or that are difficult to answer for some other reason' (Groves et al., 2004, p. 241). For example, measurement error is introduced into the data if respondents misinterpret words, concepts or entire questions, have difficulties in retrieving the information sought, or encounter problems when formatting their answers (Groves et al., 2004, p. 209). Therefore, survey researchers have to check for cognitive difficulties posed by their survey questions. This is not only important in order to improve data quality, but also to evaluate whether the survey is measuring constructs in an adequate way (Collins, 2003).

*Corresponding author. Email: cornelia.neuert@gesis.org

Today, it is generally acknowledged that new questions or survey instruments require some form of pre-evaluation before they are actually fielded. Survey methodologists have several methods at hand for evaluating survey questions, including conventional pretests, cognitive interviews, behavior coding, response latency measurement, formal respondent debriefings, and expert reviews (Presser et al., 2004). A relatively new approach to evaluating questionnaires is to incorporate eye tracking into cognitive interviewing. Whereas cognitive interviewing has become a well-established and very popular pretesting method over the last few decades (Beatty & Willis, 2007; Presser et al., 2004), eye tracking has only recently been recognized as a promising method for evaluating self-administered questionnaires in academic survey research (Galesic & Yan, 2011). The hybrid method of cognitive interviewing and eye tracking is currently being used by several questionnaire pretesting laboratories such as those at the German Federal Statistical Office (Tries, 2010) and at the United States Census Bureau (Romano & Chen, 2011). Incorporating eye tracking into cognitive interviewing is bound up with the hope that the former method will offer additional insights into question problems that would remain undetected if only cognitive interviews were conducted. A second underlying hope is that the supplementation with eye tracking will increase the degree of accuracy and precision with which problematic questions are detected in cognitive interviews. To our knowledge, however, these underlying assumptions have not yet been tested explicitly in a controlled experiment. The goal of this article was to fill this void in the existing literature.

In this paper, we test whether incorporating eye tracking into cognitive interviewing is indeed more effective in pretesting self-administered questionnaires than conducting standard cognitive interviews. In the following background section, we first present a brief review of both methods and then describe what additional insights eye tracking could provide when incorporated into cognitive interviewing. We then present and discuss the findings from our experimental study in which we compared both approaches with regard to the number and types of problems they detect as well as the number of problematic questions they identify.

Background

Cognitive interviewing

The cognitive interview is typically a semi-structured, in-depth interview that focuses on respondents' thought processes associated with answering survey questions. It is based on the four-stage survey response process model respondents follow when answering survey questions (Tourangeau, 1984; Tourangeau, Rips, & Rasinski, 2000). According to this model, when answering a survey question respondents must (1) understand the question, (2) retrieve relevant information, (3) make use of this information to form a judgment, and (4) select and report an answer that matches the response categories given by the survey question. The goal of cognitive interviewing is to obtain information on these response processes (i.e. how respondents understand a question and how they arrive at an answer) and to identify difficulties respondents have in performing them (Beatty & Willis, 2007; Miller, 2011; Willis, 2004). By identifying problematic questions and providing information about a question's need for revision, cognitive interviewing contributes to decreasing measurement error (Forsyth & Lessler, 1991; Willis, 2005).

The most commonly used techniques for obtaining information about respondents' cognitive processes and about potential question problems are thinking aloud and verbal probing. During thinking aloud, respondents are asked to report everything that comes to their mind while they are forming an answer. During probing, the interviewer asks direct questions or probes, after administering the questions, to obtain more information about how respondents interpreted and answered them. In practice, often a combination of both methods is applied (Willis, 2005).

Eye tracking

Eye tracking refers to the recording of people's eye movements while they interact with objects such as texts, images, humans, computers, or machines. It has long been used to study cognitive processing during reading and other information processing tasks (Rayner, 1998). More recently, the technique has also been introduced into the field of survey methodology to study cognitive processes during survey responding. For example, eye tracking has been used to evaluate visual designs of branching instructions (Redline & Lankford, 2001) and response formats (Lenzner, Kaczmirek, & Galesic, 2014), to investigate response order effects (Galesic, Tourangeau, Couper, & Conrad, 2008), to examine the effects of question wording on question comprehensibility (Graesser, Cai, Louwerse, & Daniel, 2006; Lenzner, Kaczmirek, & Galesic, 2011), and to study cognitive processes in answering rating scale questions (Menold, Kaczmirek, Lenzner, & Neusar, 2014). In survey pretesting, eye tracking makes it possible to observe and record respondents' eye movements in real time while they are completing a survey. Specifically, eye tracking enables the researcher to see where and for how long respondents look when reading and answering questions. This feature can be used to detect questions that are difficult to understand or that are otherwise flawed (Galesic & Yan, 2011).

The link between eye movements and cognitive processing is based upon two assumptions. The immediacy assumption postulates that words or visual objects that are fixated by the eyes are immediately processed. The eye-mind assumption assumes that words or objects are fixated as long as they are being processed (Just & Carpenter, 1980). Taken together, these two assumptions suggest that eye movements provide direct information about what people are currently processing and how much cognitive effort is involved. When text is difficult to process, the frequency of regressions (i.e. backward eye movements) and the duration of fixations increase (Rayner, 1998). Consequently, a question that is difficult to comprehend should take longer to process and this should be reflected in longer fixation times and patterns of repetitive or multiple fixations (Graesser et al., 2006; Lenzner et al., 2011). Additionally, eye tracking allows for a precise observation of participants reading patterns to reveal whether respondents actually read instructions, whether they skip (parts of) questions, and whether they are likely to skim questions or response options rather than read them thoroughly.

The rationale behind incorporating eye tracking into cognitive interviewing

The major strength of cognitive interviewing is that it is an effective tool for identifying problems with question comprehension and – most importantly – for revealing the causes of these problems. Moreover, it provides detailed insights into the cognitive processes underlying survey responding (Collins, 2003). However, both the

techniques commonly used in cognitive interviews (i.e. thinking aloud and verbal probing) as well as the more general behavior of the interviewers can have an impact on the ways respondents answer the questions (Beatty & Willis, 2007; Conrad & Blair, 2009). For example, if an interviewer asks probing questions, even though the respondent answered the survey question without apparent problems, this could affect the question answering process, which had previously occurred automatically, in a way that forces the respondent into a particular (unintended) direction (Conrad & Blair, 2009).

In contrast, eye tracking as an unobtrusive method is basically non-reactive. It allows the detection of respondents' conscious and unconscious reactions to survey questions and provides objective information about how the question and answer process proceeds under natural conditions and without the presence of a (cognitive) interviewer. In practice, respondents can be seated in front of an eye tracker in the laboratory and can be instructed to fill in a questionnaire at their usual pace. Simultaneously, a cognitive interviewer can monitor the respondents' actions and eye movements in real time on a computer screen in an adjacent room and note peculiarities to be discussed after the respondent has completed the survey. Asking probing questions after the eye-tracking session may still potentially introduce reactivity; however, this reactivity is at least triggered by behavior that has actually been observed. This should reduce reactivity bias (Conrad, Blair, & Tracy, 1999). In conclusion, eye tracking can add a non-reactive component to the cognitive interview.

Another limitation of cognitive interviewing is the inability of some respondents to express themselves verbally (Graesser et al., 2006) and to report on their cognitive processes (Willis, 2004). Additionally, respondents may not be consciously aware of all their cognitive processes, so they may sometimes also not be aware of the difficulties or problems they actually have encountered – or they may not want to communicate their difficulties, to avoid appearing ignorant to the interviewer (National Center for Health Statistics [1989] cited in Campanelli [2008]). Consequently, problems that are unconscious for respondents and problems that they cannot or do not want to express verbally have a small chance of being identified in the cognitive interview (Blair & Conrad, 2011).

By contrast, eye tracking is independent of participants' verbal abilities (Galesic & Yan, 2011). For example, eye tracking can help to ascertain whether respondents actually read instructions and definitions that are important for answering a survey question without having to rely on respondents' awareness of or willingness to report whether they have or have not read them. Moreover, eye movements can point to unfamiliar words and complex questions because respondents usually fixate these for a relatively long time and reread them several times (Lenzner et al., 2011).

Finally, the results of cognitive interviews are verbal reports that have to be interpreted by the researcher and which are therefore subjective (Beatty & Willis, 2007; Conrad & Blair, 2009). Similar to behavior coding, which is generally characterized as providing objective and replicable data (Fowler & Cannell, 1996; Groves et al., 2004), eye tracking is a more objective way of collecting information about the response processes (Galesic & Yan, 2011). Therefore, eye tracking could complement cognitive interviewing by providing additional quantitative data. However, for questionnaire pretesting, eye tracking is not suitable as a standalone technique. Eye movements can indicate whether a problem exists, but they do not provide information about what the exact problem is and what causes the problem. For example, repetitive eye movements indicate that a respondent has difficulties

to interpret and/or answer a question; however, this pattern does not reveal whether the difficulties are due to unfamiliar words, vague or ambiguous terms, or other question flaws. Moreover, long fixations and rereadings could indicate problems with the question, but they could also indicate a respondents' increased interest in the question or a relatively conscientious response style (Lenzner et al., 2011). Thus, the eye-tracking data must be enriched with additional information from the respondents, so that researchers can verify their interpretations. Cognitive interviewing is therefore obligatory after eye tracking when pretesting questionnaires. The use of eye tracking in combination with cognitive interviewing methods, such as thinking aloud or probing, has already been employed in other disciplines (e.g. web usability, Van den Haak, De Jong, & Schellens, 2003; communication and media science, Holmquist, Holsanova, Barthelson, & Lundqvist, 2003).

Method

Design and hypotheses

The aim of this study was to assess whether eye tracking can be an effective supplement to cognitive interviewing in evaluating and improving survey questions. We used a randomized between-subject design with two conditions (eye tracking yes/ no). The dependent variables were the number of problems identified, the types of problems identified, and the number of problematic questions identified. As discussed above, we expected that incorporating eye tracking into cognitive interviewing (treatment condition) would identify more problems (hypothesis 1) and more problematic questions (hypothesis 2) than the application of cognitive interviewing as usual (control condition). With regard to the types of problems identified, we did not expect differences between the two conditions (hypotheses 3) because both approaches are based on cognitive interviewing as the basis pretesting method.

Participants

We conducted this study in October and November 2012 in the pretest laboratory at GESIS – Leibniz-Institute for the Social Sciences in Mannheim, Germany. A total of 66 participants were recruited from the respondent pool maintained by the institute as well as by word of mouth. These participants received a compensation of 30 € after participating in the study. Additionally, 18 colleagues and student assistants who worked primarily in non-scientific departments of the institute participated in the study for free.[1] One participant had to be excluded from the analyses, leaving effectively 83 respondents in the data set (41 in the control and 42 in the treatment condition). Of these, 46% were male, 55% were between 18 and 34 years old, 30% were between 35 and 54 years old, and 15% were between 55 and 76 years of age. Participants' mean age was 36 (SD = 14.3). Sixty-eight percent had received twelve or more years of schooling, twelve percent had received ten years, and twenty-one percent had received nine or less years of schooling.[2] Most participants were experienced computer and Internet users who used computers and the Internet on a daily basis with 88% and 87%, respectively.

The questionnaire

The questionnaire contained 52 closed-ended items on a variety of topics, such as politics, family, social inequality, and leisure time that could be administered to the

general population.[3] Most of the questions were adapted from various existing surveys, such as the International Social Survey Programme (ISSP), the German General Social Survey (ALLBUS), and the Socio-Economic Panel (SOEP). The questionnaire included a variety of question formats: single-choice questions, grid questions, and one check-all-that-apply question. The questions were selected on the basis of anticipated problems with regard to the four stages of the response process. Participants in the treatment condition first answered the questions on a computer and later received a paper version of the questionnaire, with screenshots of the questions, during the cognitive interview. Participants in the control condition only received the paper questionnaire with the screenshots of the questions. The screenshots were printed with the same font size and line height as in the online questionnaire to keep the presentation of the questions comparable across conditions. A maximum of four items were presented per screen to avoid vertical scrolling on the computer and to ensure that the screenshots could be printed on a DIN A4 page of paper. The language of the questionnaire was German.

Eye-tracking equipment

A Tobii T120 Eye Tracker was used to record participants ' eye movements. The Tobii T120 is a remote eye tracker embedded in a 11" TFT monitor (resolution 1280 x 1024) with two binocular infrared cameras placed underneath the computer screen providing unobtrusive recording of respondents' eye movements and permitting for head movements within a range of 30 x 22 x 30 cm. Eye movements were recorded at a sampling rate of 120 Hz. The online questionnaire was programmed with a font size of 18 and 16 pixels and a line height of 40 and 32 pixels for the question text and answer options, respectively.

Interview protocol and interviewer instructions

To conduct the cognitive interviews (in both treatment and control condition), we developed an interview protocol. The interview protocol included pre-scripted, general probing questions, such as ‘Could you please explain your answer a little further?' and ‘How easy or difficult was it for you to come up with your answer?' for 13 (one-quarter) of the 52 items. These 13 items were selected randomly rather than based on theoretical expectations and hypotheses about the presence of problems in individual questions. For the remaining 39 items, the interviewers were instructed to use only conditional probes (i.e. follow-up questions that are only asked if elicited by a particular respondent behavior, Conrad and Blair, [2004]) instead of asking probing questions proactively when they themselves believed that a problem existed. Allowing the interviewers to use only conditional probes for these 39 items has the advantage that the variation in experience and behaviors across interviewers is minimized and that participants have a greater chance to express problems spontaneously and on their own. Probing questions in addition to the ones specified in the interview protocol were only asked if participants seemed to have difficulties in answering a question during the interview (conditional probing) or if – in the treatment condition – peculiar reading patterns were observed during the eye-tracking session. Indicators for difficulties in the cognitive interviews consisted of respondents needing a long time for answering a question, showing signs of uncertainty (e.g. explicit cues such as ‘um', ‘ah', and changing an answer), choosing an objectively wrong answer, or

requesting clarification (Conrad & Blair, 2001; Willis, 2005, p. 91). Peculiar reading patterns in the eye-tracking session were defined as particularly long or repeated fixations on a word, rereadings of specific words or text passages, regressions from answers to question text, correction of the chosen response category, or skipping questions. If peculiar reading patterns were observed during the eye-tracking session, the interviewers were instructed to first ask the general probing questions and to probe the peculiar reading patterns explicitly only if the general probes had not already uncovered the reasons for this peculiar reading behavior.
Interviewers in the treatment condition were provided with a coding scheme for peculiar reading patterns where they had to check a box if they observed one of the five behaviors mentioned above. To assess the intercoder reliability of the peculiar reading patterns, all five interviewers coded a sample of six eye-tracking sessions. Coding reliability was found to be adequate: the overall median Kappa statistic was .64, which is generally classified as 'substantial' reliability (Landis & Koch, 1977). Agreement between individual raters ranged from .51 to .72.

Procedure

Respondents in the treatment condition were seated in front of the eye tracker. They were instructed to fill in the questionnaire as they would in their normal environment and to articulate problems or difficulties at any time they occurred. After completion of a standard calibration procedure and two warm-up questions, the actual survey started and participants' eye movements were tracked. Simultaneously, their reading patterns were monitored in real time by an interviewer on a second screen in an adjacent room. The interviewer used the coding scheme described above to document any peculiar reading pattern he or she observed.
Immediately after respondents had completed the online survey, a cognitive interview was conducted. In addition to probing the questions specified in the cognitive interview protocol, interviewers were instructed to probe those questions for which they had noted peculiar reading patterns during the eye-tracking session. Because probing questions were not asked immediately after they had responded to the questions in the web survey, participants were asked to answer those questions that had been selected for probing once again, on paper, before being asked to respond to the probing questions. This procedure was used to remind the participants of their initial thoughts. In the control condition, only a cognitive interview was conducted. Respondents first received the questions on paper, one question at a time. If probing questions for the individual questions were specified in the interview protocol, these were asked immediately after participants had provided an answer. In addition, conditional probing (for other questions) was applied if respondents needed a long time to answer a question, showed signs of uncertainty, chose an objectively wrong answer, or requested clarification.
The interviews were conducted by five interviewers (three researchers and two student assistants) which had between 1 and 10 years of experience in using cognitive interviewing methods. The interviewers received specific training on coding peculiar reading patterns with a training video. The individual interviewers each conducted between 14 and 20 interviews and carried out an equal number of interviews in both conditions. The average interview length was 44 min in the control condition and 60 min in the treatment condition, including the completion of the online survey with a mean answer time of almost 13 min. All cognitive interviews were videotaped.

Results

The analysis described below centers on three basic issues: the number of problems, types of problems, and problematic questions identified by each method. Moreover, we take a closer look at the severity of the problems identified by only one of the two methods and examine whether the quantitative eye-tracking data confirm the results from the cognitive interviews.

Number and types of problems

For problem identification, all videotapes of the cognitive interviews were reviewed by the first author and each questionnaire item, for each interview, was given a dichotomous score that reflected whether a problem was identified in the question (1) or not (0). A student assistant coded 10% of the interviews for estimating interrater reliability. Agreement between these two raters was 93% and the Kappa statistic (Cohen, 1960), which accounts for chance, was found to be Kappa = .69, which is generally classified as 'substantial' reliability (Landis & Koch, 1977). If an item was perceived as problematic, short descriptions about the nature of the problem(s) were noted. In the next step, these descriptions were coded into problem types using a problem classification scheme adopted from various existing schemes (DeMaio & Landreth, 2004; Presser & Blair, 1994). The problem classification scheme included a total of 30 problem codes, which were grouped into the four stages of the survey response process (comprehension, retrieval, judgment, response selection; Tourangeau, 1984; Tourangeau et al., 2000) and an additional category for navigational problems (see Appendix 1). Individual items could be assigned multiple problem codes.

Table 1 shows the total number of problems identified by each method and the variants of probing that lead to the identification of these problems. Comparing the total number of problems across treatments revealed that incorporating eye tracking into cognitive interviewing (treatment condition) detected more problems than cognitive interviewing (control condition) alone, but this difference was not statistically significant ($x^2 = 2.08$, df = 1, p = .188).[4] In the next step, we examined whether the problems found were identified by pre-scripted probes or by conditional

probing based either on peculiar reading patterns or on peculiar response behaviors. If most problems were identified by conditional probing based on peculiar reading patterns, this would suggest that eye tracking indeed offers additional insights into question problems. Overall, 30.8% of the problems found where identified by prescripted probes and 69.2% were identified by conditional probing based on peculiar

Table 1. Number of problems identified by method and by types of probing questions.

Types of probes	Cognitive interviewing	Eye-tracking and cognitive interviewing	Total number of problems
Pre-scripted	125 (36.2%)	102 (26.0%)	227 (30.8%)
Conditional based on peculiar response behavior	220 (63.8%)	-	220 (29.9%)
Conditional based on peculiar reading patterns	-	290 (74.0%)	290 (39.3%)
Total number of problems	345 (100%)	392 (100%)	737 (100%)

Table 2. Types of problems identified by method.

Types of problems	Cognitive interviewing	Eye-tracking and cognitive interviewing	Total number of problems
Comprehension	84.6% (292)	86.5% (339)	85.6% (631)
Retrieval	2.3% (8)	1.0% (4)	1.6% (12)
Judgment	4.1% (14)	4.6% (18)	4.3% (32)
Response selection	9.0% (31)	7.4% (29)	8.6% (60)
Navigation	.0% (0)	.5% (2)	.3% (2)
Total	345	392	737

response behavior in the control condition (29.9%) or based on peculiar reading patterns in the treatment condition (39.3%). Significantly more problems were identified by conditional probing in the treatment condition than in the control condition ($x^2 = 8.98$, df = 1, p = .005). These findings suggest that respondents' eye movements indeed hint at question problems that would remain undetected if no eye tracker was used.

With regard to the types of problems identified, the vast majority of problems were classified as comprehension problems in both conditions and the second largest group of problems – only around one-tenth of the size of the largest group – was related to response selection (see Table 2), which is in line with previous research (e.g. DeMaio & Landreth, 2004; Presser & Blair, 1994). Here, no statistically significant difference was found between the two conditions ($x^2 = 4.42$, df= 4, p = .352).

Number of problematic questions

In our next analysis step, we evaluated whether one method is more effective than the other in identifying problematic questions. Specifically, we examined whether both methods identify the same or different questions as problematic. To compare the number of problematic questions across conditions, we had to decide on a quantitative threshold at which we defined a question as problematic.[5] In accordance with recommendations from behavior coding (Blair & Srinath, 2008; Fowler, 1992), we coded a question as problematic if at least 15% of the respondents had a problem with the item.[6]

Table 3 shows the total number of problematic questions identified by each method and whether these questions were identified by pre-scripted or conditional probing. A larger number of problematic questions were identified in the treatment condition than in the control condition. In the control condition, 20 flawed questions

Table 3. Number of problematic questions identified by method and by types of probing questions.

Types of probes	Cognitive interviewing	Eye-tracking and cognitive interviewing	Identified by both methods
Pre-scripted probes	9	11	9
Conditional probes	11	14	9
Total number of problematic questions	20	25	18

were identified (16 attitudinal, 4 factual questions), whereas in the treatment condition, 25 problematic questions were detected (21 attitudinal, 4 factual questions).
This difference, however, was not statistically significant ($x^2 = .98$, df= 1, $p = .645$). In total, 18 of the flawed questions were identified in both conditions, nine by prescripted probing questions and nine by conditional probing, respectively. In the control condition, two questions that showed no flaws in the treatment condition were identified (by conditional probing); in the treatment condition, seven questions were detected that were not identified in the control condition. Of these seven questions, five were identified by conditional probing triggered by the observation of peculiar reading patterns. Those questions would not have been identified if only a cognitive interview was conducted. The remaining two questions were identified by pre-defined probes. Hence, identification of these latter two problematic questions does not constitute a contribution of eye tracking.

Severity of problems

Given that some questions were only identified as problematic by one but not the other method, the question arose whether these were serious or only relatively minor (and probably neglectable) problems. Thus, in an additional exploratory analysis step, we examined whether the problems identified by only one of the two methods vary in their severity (Blair & Conrad, 2011; Presser & Blair, 1994). Severity was defined as the effect of a question problem on each measurement (Blair & Conrad, 2011) and quantified according to the approach of Blair and Conrad (2011): three questionnaire design experts independently rated the problems identified in those (nine) questions which were detected in one but not both conditions on a scale of one (no or minor effects) to ten (extremely serious effects).[7] Subsequently, the ratings were averaged across the experts.[8]
Table 4 lists the respective questions together with their severity ratings, sorted by average question severity score (ranging from Ø2.5 to Ø7.3). Problem scores for the individual types of problems per question range from 1.0 (in Q11.1) to 8.7 (in Q10.1) and we divided the problems into severity quartiles in which first-quartile problems were defined as non-crucial or weak problems and fourth-quartile problems were defined as severe problems. One (Q10.1) of the two questions which were only identified in the control condition received a high average score (Ø6.7) and contained the most serious problem, with a score of 8.7, namely that the term ‘corrupt' was unknown/unfamiliar to some respondents. The remaining types of problems in question Q10.1 were middle-quartile problems.
The second problematic question (Q8) that was exclusively identified in the control condition received a comparatively low average severity score and contained two types of problems that were both in the lowest quartile (Ø2.5). One of the problems concerned an unclear respondent instruction (severity = 2.0). The question was a check-all-that-apply question and several participants asked whether they are allowed to tick more than one answer. The other problem concerned one of the response categories [sign a petition] and was classified as undefined/vague term and rated with a severity score of 3.0. In German, ‘sign a petition' [Beteiligung an einer Unterschriftensammlung] could be either interpreted as signing a petition or as collecting signatures for a petition, although this is not the case in the English translation of the response category.

Table 4. Severity rating and problems identified by method.

Question		Identified in	Problem (code)	Severity Ø
Q8	If you wanted to have political influence or to make your point of view felt on an issue which was important to you: which of the possibilities listed on these cards would you use? Which of them would you consider? *Please select all that apply.* • Express your opinion to friends and acquaintances and at work • … • Boycott or buy goods for political, ethical or environmental reasons	Control condition		**2.5**
			Undefined/vague term [sign a petition] (4)	3.0
			Unclear respondent instruction (9)	2.0
Q6.7	By and large, economic profits are nowadays distributed fairly in Germany	Experimental condition		**3.4**
			Vague/unclear question (1)	4.7
			Knowledge may not exist (5)	3.7
			Question is misunderstood (1)	2.7
			Undefined/vague term [fairly] (4)	2.7
Q11.6	People worry too much about human progress harming the environment Agree strongly – agree – neither agree nor disagree – disagree – disagree strongly	Experimental condition		**4.3**
			Vague/unclear question (1)	6.7
			Undefined/vague term [human progress] (4)	4.3
			The response of others or of the general public is asked (15)	4.0
			Too detailed or broad response categories (24)	2.0
Qll.l	We believe too often in science, and not enough in feelings and faith	Experimental condition		**4.8**
			Knowledge may not exist (5)	7.0
			Vague/unclear question (1)	6.0
			Boundary lines (6)	5.7
			Undefined/vague term [Science] (4)	5.3
			Undefined/vague term [faith] (4)	3.7
			Unclear respondent instruction (9)	1.0

(Continued)

Table 4. *(Continued)*.

Question		Identified in	Problem (code)	Severity Ø
Q11.2	Overall, modem science does more harm than good	Experimental condition		**5.7**
			Knowledge may not exist (5)	8.0
			Vague/unclear question (1)	4.7
			Undefined/vague term [modem science] (4)	4.3
Q7	Suppose a law were being considered by [the German Bundestag] that you considered to be unjust or harmful. If such a case arose, how likely is it that you, acting alone or together with others, would be able to try to do something about it?	Experimental condition		**6.1**
			Boundary lines (6)	7.7
			Undefined/vague term [do something about it] (4)	6.3
			Undefined/vague term [unjust or harmful] (4)	5.3
			Complex or awkward syntax (11)	5.0
Q6.3	The State has to make sure that everyone has a job and that prices remain stable, even if the freedom of entrepreneurs has to be curtailed because of this	Experimental condition		**6.6**
			Vague/unclear question (1)	7.3
			Vague/unclear question/question is misunderstood (1)	6.7
			Information overload, question too long (10)	6.7
			Several questions in one or multiple subjects (14)	6.7
			Complex topic (2)	6.0
			Knowledge may not exist (5)	6.0
Q10.1	To get all the way to the top in Germany today, you have to be corrupt Strongly agree – agree – neither agree nor disagree disagree – strongly disagree	Control condition		**6.7**
			Undefined/vague term [corrupt] (4)	8.7
			Vague/unclear question (1)	6.7
			Objectively wrong answer/question is misunderstood (7)	6.7
			Response categories not appropriate to question (23)	6.0
			Knowledge may not exist (5)	5.3
Q10.4	In Germany people have the same chances to enter university, regardless of their gender, ethnicity or social background	Experimental condition		**7.3**
			Objectively wrong answer/question is misunderstood (7)	8.3
			Several questions in one or multiple subjects (14)	7.3
			Uncertainty which answer category reflects own opinion (29)	7.3
			Vague/unclear question (1)	6.0

Note: The original questions (in German) are available upon request. Bold figures are averaged question severity scores.

Five (Q11.1, Q11.2, Q7, Q6.3, and Q10.4) of the seven problematic questions that were identified only in the treatment condition exhibited (up to three) fourth-quartile problem types and four of these received an above-average score (except Q11.1). The remaining two questions (Q6.7, Q11.6) received comparatively low
average scores (Ø3.4; 4.3, respectively), and the types of problems identified in these questions were mainly defined as lowest quartile problems. As an example of a severely problematic question, consider question Q10.4 which received the highest problem severity rating (Ø7.3) across all questions. In this question, the raters considered the fact that the question was misunderstood as there was a misfit between the response option chosen and the explanation given as the most serious problem (severity = 8.3). Additional flaws were that the question contained several questions in one (severity = 7.3), the respondents did not know which answer category reflected their own opinion appropriately (severity = 7.3), and the question was found to be vague/unclear (severity = 6.0).
Overall, these results show that both methods identify problems that are considered to have serious effects on data quality, as evaluated by three questionnaire experts. Whereas in the control condition, one of two questions (50%) was found to contain severe problems, five of seven questions (71%) contained such problems in the treatment condition.

Quantitative eye-tracking data

The final question we investigated was whether the quantitative eye-tracking data confirmed the results from the cognitive interviews. If this is the case, both cognitive interviewing and eye-tracking data should identify the same questions as problematic and verify each other. As an indicator of question difficulty, we used the eye-tracking metric question fixation time[9] in the Tobii Studio 3.2.1 software and examined the total time participants spent fixating a question (including the response options and possible instructions). A perfect relationship between problematic questions (as identified during the cognitive interview) and question fixation time would mean that all problematic questions would have longer fixation times than non-problematic questions.
If participants exhibited data with too many data gaps due to miscalibration or substantial positional changes while filling-in the questionnaire, they were excluded from the fixation times analysis of the respective questions. This procedure left between 35 and 41 participants per question in the analysis. In order to compare the eye-movement data with the findings from the cognitive interviews, we sorted the items by total fixation duration and divided them into quartiles: The top quartile contained questions with relatively long fixation times and the lowest quartile with short fixation times. When looking at questions in the top and bottom quartiles, we found an agreement between question problems and fixation time of 77%, respectively: The vast majority of questions in the upper quartile were identified as problematic in the cognitive interview (10 of 13), while in the lower quartile, the vast majority were considered unproblematic (10 of 13). Although this is not a perfect relationship, the results of the eye-tracking analyses reveal that the problems found in the cognitive interviews are actually grounded in the eye-movement behavior of the participants. On the one hand, this gives more confidence to the (real time) coding judgments of the interviewers and, on the other hand, to the interpretation and analysis of the qualitative data, which can be considered to be more valid.

Discussion and conclusion

The aim of this study was to test whether eye tracking is an effective supplement to cognitive interviewing in evaluating and improving survey questions. We found support for our hypotheses that incorporating eye tracking into cognitive interviewing is more productive in identifying both questionnaire problems (hypothesis 1) and problematic questions (hypothesis 2) than using cognitive interviewing alone. Given that problem detection is the primary objective of most pretesting methods (Conrad & Blair, 2004) and also an important indicator for the evaluation of pretesting methods, our results indicate that eye tracking and cognitive interviewing complement each other effectively.

With regard to the types of problems, both experimental conditions produced almost identical results. This is in line with hypothesis 3 and, actually, not surprising, given that in both conditions, cognitive interviewing is the basic method used to gain information about the causes of question problems. Finally, we did not find differences between both conditions with respect to the severity of the problems identified. With regard to those questions that were identified as problematic in one condition but not in the other, both methods identified problems that were considered to have serious effects on data quality. In the treatment condition, five of seven questions were judged to exhibit severe problems. Hence, incorporating eye tracking into cognitive interviewing helps to detect severely problematic questions that would remain unnoticed if only cognitive interviewing was conducted.

Apart from our findings that the hybrid method of cognitive interviewing and eye tracking identified both more questionnaire problems and more problematic questions, there are considerable benefits from incorporating eye tracking into cognitive interviewing when testing survey questions. First, as interviewers observe the eye movements of the respondents in real time, they obtain a better understanding of the participant's answer process and problems that have arisen while answering. This is advantageous in several respects for the subsequent cognitive interview. First, providing interviewers with additional insights into participants' behavior helps them to use relevant conditional probes. Second, although participants might not point to a problem because they are either not aware of it or it is too demanding to verbalize it, their eye movements provide interviewers with information that point to difficulties. Thereby, eye tracking contributes to identifying problems that are not consciously apparent to participants and have a small chance of being detected in the cognitive interview. As an additional benefit, asking probing questions in a more targeted way also increases the efficiency of pretesting, because it allows for testing a much larger set of items within a given period of time. And, finally, analyzing eye-tracking metrics quantitatively, such as the total time participants fixated on a question, enables researchers to compare objective eye-movement data with the verbal data gathered from the cognitive interviews. Linking results from different data sources permits researchers to compare and confirm the conclusions made and to achieve more objective and valid results.

Alongside these advantages, however, the use of eye tracking also brings certain challenges with it. First, the setup costs of an eye tracker are comparatively high. When using eye tracking, one needs to decide whether gaining additional information about potential question problems pays off against the financial investment required. A further limitation is that not everyone's eye movements can be recorded accurately, for example, wearers of glasses. And finally, eye movements alone can

only hint at problems but do not tell us what exactly the problem is. Therefore, conducting a cognitive interview after the eye-tracking session is obligatory.

One could argue that comparing only cognitive interviewing to only eye tracking would have been a more clear-cut approach for examining the effectiveness of both methods. Similarly, testing one group of participants with eye tracking only and one group with cognitive interviewing only may shorten the time required for conducting the individual interview sessions. However, as was mentioned above, eye tracking is hardly usable as a stand-alone pretesting method because it is not able to reveal the causes of question problems. Additionally, one of the biggest benefits of combining both methods, namely giving cognitive interviewers additional cues about what questions or question aspects they should probe, would be lost if eye tracking was used exclusively.

A limitation of this study is that the two conditions differed somewhat with regard to the mode in which the questions were administered (interviewer present and concurrent probing in control condition vs. interviewer absent during eye-tracking session and hybrid of retrospective and concurrent probing in treatment condition). From a theoretical perspective, it would have been desirable to apply identical procedures in both conditions. However, our design decision was primarily guided by practical considerations about the ways we would normally conduct cognitive interviews (concurrent probing by an interviewer) and how we envisioned the application of cognitive interviewing supplemented with eye tracking (hybrid of retrospective and concurrent probing with the interviewer being absent during the eye-tracking session). In order to evaluate the strengths of both methods under realistic conditions (and thereby to increase the external validity of the experiment), we had to accept the risk that the different settings may differently affect participants' response processes. For example, while the typical cognitive interview setting encourages respondents to spontaneously comment on the questions, the eye-tracking setting (without an interviewer present) does not. It is possible that the cognitive interview in the treatment condition did not provide an account of all the problems participants encountered. By the time the cognitive interview was conducted, some respondents might already have resolved (or at least think they have resolved) some of the problems they experienced during the eye-tracking session.

To mitigate this effect, respondents in the treatment condition were encouraged to articulate any problems they encountered immediately while completing the web questionnaire. Moreover, any difficulties the respondents experienced during the eye-tracking sessions should be reflected in their eye movements and thus followed up on later in the cognitive interviews.

The current study clearly calls for future research. First, it would be worthwhile to investigate the use of different eye-tracking techniques and procedures when incorporating it into cognitive interviews. For example, is there an additional benefit if respondents are shown a video of their eye movements during the cognitive interview and are reminded of their answer process? A second line of research worth investigating might be the development of an automatic coding system for peculiar reading patterns to detect problems in survey questions based on the participants' reading behavior.

Acknowledgement

We are grateful to Rolf Porst, Katharina Disch, and Sabrina Zolg for their help in conducting the cognitive interviews.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. Excluding these participants does not alter our conclusions. The relevant results are available upon request.
2. Chi-squared analysis revealed no statistically significant differences between both experimental conditions regarding socio-demographic characteristics, such as gender ($x^2 = .115$, df = 1, p = .734), age ($x^2 = 3.696$, df = 2, p = .158), and education ($x^2 = .733$, df= 2, p = .693).
3. The questionnaire is available from the authors on request.
4. We did not expect our results to achieve statistical significance. A power analysis (x^2 test, $\alpha = .05$) indicated that a minimum sample size of N = 1300 would be required to detect any significant effects of low size (.1) or a minimum sample size of N= 145 to detect effects of medium size (.3) (G*Power 3, Faul, Erdfelder, Lang, & Buchner, 2007). Recruiting and testing so many participants would be highly inefficient in an eye-tracking study. Nevertheless, we use statistical tests for heuristic purposes.
5. Although Beatty and Willis (2007) state that there is no link between the evidence of problems and the number of participants who indicate a problem, we follow the reasoning of Conrad and Blair (2009) that 'over a set of interviews, seriously flawed questions should produce more evidence of problems than questions without flaws' (Conrad & Blair, 2009, p. 51).
6. To check the robustness of our results, we also examined the results using cutoffs at 10 and 20%. In both cases, more problematic questions were identified in the treatment condition. Using the lower cutoff, a larger number of problematic questions were detected, whereas at the higher cutoff, fewer problematic questions were detected (in both conditions, respectively).
7. In contrast to Blair and Conrad (2011), who ask their experts to rate the impact on data quality on two dimensions, namely prevalence and severity, we deviate from their approach for three reasons: first, we are particularly concerned with a problem's severity and not with its prevalence. Second, for purpose, the results are more intuitively interpretable if only a scale from 1 to 10 is used and the resulting values are not blurred by multiplying the ratings for severity and prevalence. Third, the evaluation of prevalence seems to be more subjective and difficult for experts to rate than the severity of the effect of a problem.
8. The intraclass correlation between experts was ICC = .44, which is classified as fair agreement (Cicchetti, 1994).
9. We also reran the analysis with the eye-tracking metric question fixation count. All of our conclusions remained unchanged (the results are available on request).

Notes on contributors

Cornelia Eva Neuert is a researcher at the Survey Design and Methodology Department at GESIS – Leibniz Institute for the Social Sciences in Mannheim, Germany. Her research interests include question evaluation, eye-tracking, and survey methodology.

Timo Lenzner is a senior researcher at the Survey Design and Methodology Department at GESIS – Leibniz Institute for the Social Sciences, Germany. His research focuses on questionnaire design and evaluation, Web surveys, and eye-tracking. He has authored and co-authored several papers on these topics published in journals such as Sociological Methods & Research, Social Science Computer Review, Applied Cognitive Psychology, and Field Methods.

References

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. Public Opinion Quarterly, 71, 287-311. doi:10.1093/poq/nfm006

Blair, J., & Conrad, F. G. (2011). Sample size for cognitive interview pretesting. Public Opinion Quarterly, 75, 636-658. doi:10.1093/poq/nfr035

Blair, J., & Srinath, K. P. (2008). A note on sample size for behavior coding pretests. Field Methods, 20,85-95. doi:10.1177/1525822X07303601

Campanelli, P. (2008). Testing survey questions. In J. Hox, E. de Leeuw, & D. Dillman (Eds.), International Handbook of Survey Methodology (pp. 176-200). New York. NY: Erlbaum/ Taylor & Francis.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6, 284-290.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20,37-46.

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. Quality of Life Research, 12, 229-238.

Conrad, F. G., & Blair, J. (2001). Interpreting verbal reports in cognitive interviews: Probes matter. In Proceedings of the American Statistical Association, Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Conrad, F. G., & Blair, J. (2004). Data quality in cognitive interviews: The case of verbal reports. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), Methods for testing and evaluating survey questionnaires (pp. 67-87). New York, NY: Wiley. doi:10.1002/0471654728.ch4

Conrad, F. G., & Blair, J. (2009). Sources of error in cognitive interviews. Public Opinion Quarterly, 73,32-55.

Conrad, F. G., Blair, J., & Tracy, E. (1999). Verbal reports are data! A theoretical approach to cognitive interviews. Proceedings of the Federal Committee on Statistical Methodology Research Conference (pp. 11-20). Arlington, VA.

DeMaio, T. J., & Landreth, A. (2004). Do different cognitive interview techniques produce different results? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J.

Martin, & E. Singer (Eds.), Methods for testing and evaluating survey questionnaires (pp. 89-108). New York, NY: Wiley. doi:10.1002/0471654728.ch5

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175-191.

Forsyth, B. H., & Lessler, J. T. (1991). Cognitive laboratory methods: A taxonomy. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), Measurement errors in surveys (pp. 393-418). New York, NY: Wiley.

Fowler, F. J. (1992). How unclear terms affect survey data. Public Opinion Quarterly, 56, 218-231. doi:10.1086/269312

Fowler, F. J. (1995). Improving survey questions. Design and evaluations. Thousand Oaks: Sage.

Fowler, F. J., & Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In N. Schwarz & S. Schuman (Eds.), Answering questions.

Methodology for determining cognitive and communicative processes in survey research (pp. 15-36). San Francisco, CA: Jossey-Bass.

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. doi:10.1093/poq/nfn059

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. doi:10.1093/poq/nfj012

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). Survey methodology. Hoboken, NJ: Wiley.

Holmquist, K., Holsanova, J., Barthelson, M., & Lundqvist, D. (2003). Reading or scanning? A study of newspaper and net paper reading. In J. Hyöna, R. Radach, & H. Deubel (Eds.), The mind's eye. Cognitive and applied aspects of eye movement research (pp. 657-670). Amsterdam: North-Holland.

Just, M. A. & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review, 87, 329-354.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

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. doi:10.1093/ijpor/edq053

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.

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? Field Methods, 26, 21-39. doi:10.1177/1525822X13508270

Miller, K. (2011). Cognitive interviewing. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), Question evaluation methods: Contributing to the science of data quality (pp. 5175). New York, NY: Wiley.

National Center for Health Statistics. (1989). Questionnaire design in the cognitive research laboratory, Series 6: Cognition and survey measurement, no 1 (DHHS Publication No. PHS 89-1076). Hyattsville, MD: US Department of Health and Human Services.

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results? Sociological Methodology, 24, 73-104.

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. doi:10.1093/poq/nfh008

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124, 372-422.

Redline, C. D., & Lankford, C. P. (2001). Eye-movement analysis: A new tool for evaluating the design of visually administered instruments (paper and web). Paper prepared for presentation at the annual meeting of the American Association for Public Opinion Research, Montreal.

Romano, J. C., & Chen, J. M. (2011). A usability and eye-tracking evaluation of four versions of the online national survey of college graduates (NSCG): Iteration 2. Study Series: Survey Methodology 2011-01, Washington, DC: U.S. Census Bureau.

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. New York, NY: Cambridge University Press.

Tries, S. (2010). Usability tests of online questionnaires. In Federal Statistical Office (Ed.), Methods, approaches, developments: Information of the German Federal Statistical Office (pp. 5-8). Wiesbaden: Federal Statistical Office.

Van den Haak, M., De Jong, M., & Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. Behaviour & Information Technology, 22, 339-351. doi:10.1080/0044929031000

Willis, G. B. (2004). Cognitive interviewing revisited: A useful technique, in theory? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), Methods for testing and evaluating survey questionnaires (pp. 23-43). New York, NY: Wiley. doi:10.1002/0471654728.ch2

Willis, G. B. (2005). Cognitive interviewing: A tool for improving questionnaire design. London: Sage.

Appendix 1 Classification scheme

Comprehension	Retrieval	Judgment	Response Selection
Question content 1. Vague/unclear question 2. Complex topic 3. Topic carried over from earlier question 4. Undefined/vague term 5. Knowledge may not exist 6. Boundary lines 7. Misfit between response option chosen and explanation given, question is misunderstood/erroneous answer	*Retrieval from memory* 18. High detail required or information unavailable 19. Long recall or reference period	*Judgment and evaluation* 20. Complex estimation, difficult mental calculation required 21. Potentially sensitive/ desirability bias	*Response terminology* 22. Undefined/vague term *Response Units* 23. Response categories not appropriate to question 24. Too detailed or broad response categories 25. Vague response categories
Question structure 8. Transition needed 9. Unclear respondent instruction 10.Information overload, question too long 11. Complex or awkward syntax 12. Erroneous/inappropriate assumption 13. Assumes constant behavior 14. Several questions in one, multiple subjects 15. The response of others or of the general public is asked for			*Response structure* 26. Overlapping response categories 27. Missing response categories 28. No formally adequate answer 29. Uncertainty which answer category reflects own opinion *Questionnaire Navigation* 30. Questionnaire navigation
Reference period 16. Reference periods are missing or undefined 17. Reference period carried over from earlier question			