EYE TRACKING IN QUESTIONNAIRE PRETESTING

Inauguraldissertation
zur Erlangung des akademischen Grades
einer Doktorin der Sozialwissenschaften
der Universität Mannheim

Vorgelegt von
Cornelia Eva Lilian Neuert

Hauptamtlicher Dekan der Fakultät für Sozialwissenschaften:
Prof. Dr. Michael Diehl

Erstgutachter:
Prof. Dr. Michael Braun

Zweitgutachter:
Prof. Dr. Michael Bošnjak

Drittgutachterin:
Prof. Annelies Blom, Ph.D.

Tag der Disputation: 21. April 2016

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

A large number of people have guided and supported me during the work on this dissertation project. First of all, I wish to thank my supervisors, Prof. Dr. Michael Braun and Prof. Dr. Michael Bošnjak, for providing important guidance and feedback. I also wish to thank Prof. Annelies Blom, Ph.D for reviewing my thesis.

I am very grateful to my colleague and co-author Dr. Timo Lenzner for his support, mentoring and for hours of valuable discussions. For their help in conducting the study I feel especially thankful with Rolf Porst, Katharina Disch, and Sabrina Zolg. I also wish to thank Franziska Adams, Jan Karem Höhne, and Sara Schneider. I have also benefited from the motivating and excellent environment for doing research at GESIS and I am very thankful for discussions and working with Dr. Kathrin Bogner, Prof. Dr. Constanze Beierlein, Bianka Breyer, Dr. Daniel Danner, Dr. Angelika Glöckner-Rist, Katharina Meitinger, Dr. Natalja Menold, Wanda Otto, again Rolf Porst, Prof. Dr. Beatrice Rammstedt, Angelika Stiegler, Cornelia Züll, and all my former and current colleagues at GESIS.

Finally, I would like to thank my family for always being there for me, and particularly my husband Johannes Jarke, for always believing in me and for encouragement and both professional and emotional support throughout the years. Thank you!

1 GENERAL INTRODUCTION

"The questionnaire designer must understand the need to pretest, pretest, and then pretest some more."

AMERICAN STATISTICAL ASSOCIATION (1999, p. 11)

The survey is a cornerstone in the toolbox of the social sciences (Groves et al., 2009). A respondent's answers about facts, perceptions, beliefs, values, opinions, attitudes, or behaviors are not only used to measure public opinion and to understand the workings of a group or society but also to inform political decisions (e.g., Foddy, 1993; Fowler, 2013; Groves et al., 2009). Thus, questions asked in surveys should produce data that are valid, reliable, and unbiased (Fowler, 2013; Fowler & Cannell, 1996; Schober & Conrad, 1997). A critical step to this end is to design the survey in a way that (i) each respondent comprehends the questions, (ii) all respondents understand the questions in the same way, and specifically, (iii) understand them as the researcher intended them to be understood. In addition, the questions should only ask for information that respondents have available and can retrieve. This is the task of survey *pretesting* and *evaluation* (Collins, 2015; De Leeuw, Hox, & Dillman, 2007; Fowler, 1995, 2013; Madans et al., 2011; Miller, 2014).

Survey methodologists have a broad and growing set of methods at their disposal (see section 1.2).[1] Thus, a key question with which any pretester is confronted is which methods, or which compounds of methods, are maximally *productive* (and eventually *efficient*) in detecting potential problems with survey items. The present thesis contributes to the understanding of this vital issue by presenting novel experimental results on the productivity of *eye tracking* in survey pretesting, both as a stand-alone technique and in conjunction with the standard method of cognitive interviewing.

[1] Examples are conventional pretests, cognitive interviews, behavior coding, response latency measurement, formal respondent debriefings, and expert reviews. See Presser et al. (2004) for an overview.

Eye tracking is one of the most recent additions to the survey pretester's toolbox. During eye tracking, the position of respondents' eyes is observed, to detect where they are looking. While being recognized as a promising technique to indicate potential problems with survey items and obtain insights into the underlying cognitive processes (Galesic & Yan, 2011), there is little resilient evidence on its productivity. The research presented in this thesis is designed to address this gap in the current literature.

The core part of the thesis consists of three controlled experiments, which are presented in chapters 2 through 4.[2] The first two studies examine eye tracking in conjunction with cognitive interviewing, which is currently the most frequently used method in survey pretesting (Beatty & Willis, 2007; Presser et al., 2004), such that a joint implementation of eye tracking and cognitive interviewing appears to be a natural point of departure. Chapter 2 reports on a method comparison experiment that is designed to examine whether a cognitive interview supplemented with eye tracking is more productive in detecting potential problems than cognitive interviews alone. Chapter 3 compares two eye-tracking-supported cognitive interviewing techniques with respect to their productivity. The final study (chapter 4) utilizes eye tracking as a stand-alone technique to add novel insights on the cognitive processing of forced-choice vs. check-all-that-apply question formats.

The remainder of the present chapter is devoted to the exposition of a concise framework for the three original contributions. Section 1.1 briefly reviews the fundamental problem that motivates pretesting, namely *measurement error*. Specifically, the cognitive processes that are involved in question response and the associated sources of response error are discussed. Section 1.2 introduces the set of standard pretesting methods. For later reference, a focus is set on eye tracking and cognitive interviewing. Section 1.3 is devoted to a more detailed outline of the main research questions addressed within the thesis, and a summary of its findings. The

[2] Variants of two of these chapters have been published in the *International Journal of Social Research Methodology* (chapter 2) and the *Social Science Computer Review* (chapter 3). A version of chapter 4 is currently under review at *Field Methods*.

final section concludes with a discussion of the utility of eye tracking in survey pretesting and suggests directions for future research.

1.1 Measurement error and the question-response process

There are many factors that can have an impact on the quality of a survey, for example, coverage, data collection, or data processing. In the field of survey methodology, these factors are often framed by the concept of *total survey error* (Groves et al., 2009; Groves & Lyberg, 2010). Basically, this concept differentiates between the quality of measurement and the quality of the representation of the target population (see Figure 1.1).[3] During each of the steps, there is a risk of errors (represented by ellipses, Groves et al., 2009).

One type of error that occurs during the measurement process is *measurement error* or error of observation. Measurement error can appear in the response process while a question is being answered by the survey respondents. It is defined as the deviation of the provided response from the true value that the measurement instrument is designed to measure. These errors could be random or systematic, resulting in variance or bias, respectively. Systematic deviations can result in biased estimates of all respondents or of a specific sub-group of respondents (Groves et al., 2009).

According to Biemer et al. (1991), there are three main sources of measurement error: the questionnaire, the method of data collection, and the respondent.[4] Each of these sources can introduce error separately, but they can also interact. The following section describes how respondents produce an answer by

[3] In an earlier work, Groves (1989) distinguishes between errors of nonobservation (coverage, sampling, nonresponse errors) and errors of observation (errors arising from the mode of data collection, interviewers, measurement instrumentation, and respondents themselves). Groves (1989, p.11) defines observational errors as "deviations of the answer of respondents from their true values of measurement" and non-observational errors as "errors arising because measurements were not taken on part of the whole population." Errors of nonobservation are not discussed here.

[4] Using interviewers introduces a fourth source of error. However, eye tracking is especially useful for visually presented, self-administered questionnaires. For this reason, I will concentrate on the questionnaire, the respondent, and their interaction.

reviewing the cognitive processes underlying survey responses and how this can affect measurement accuracy.

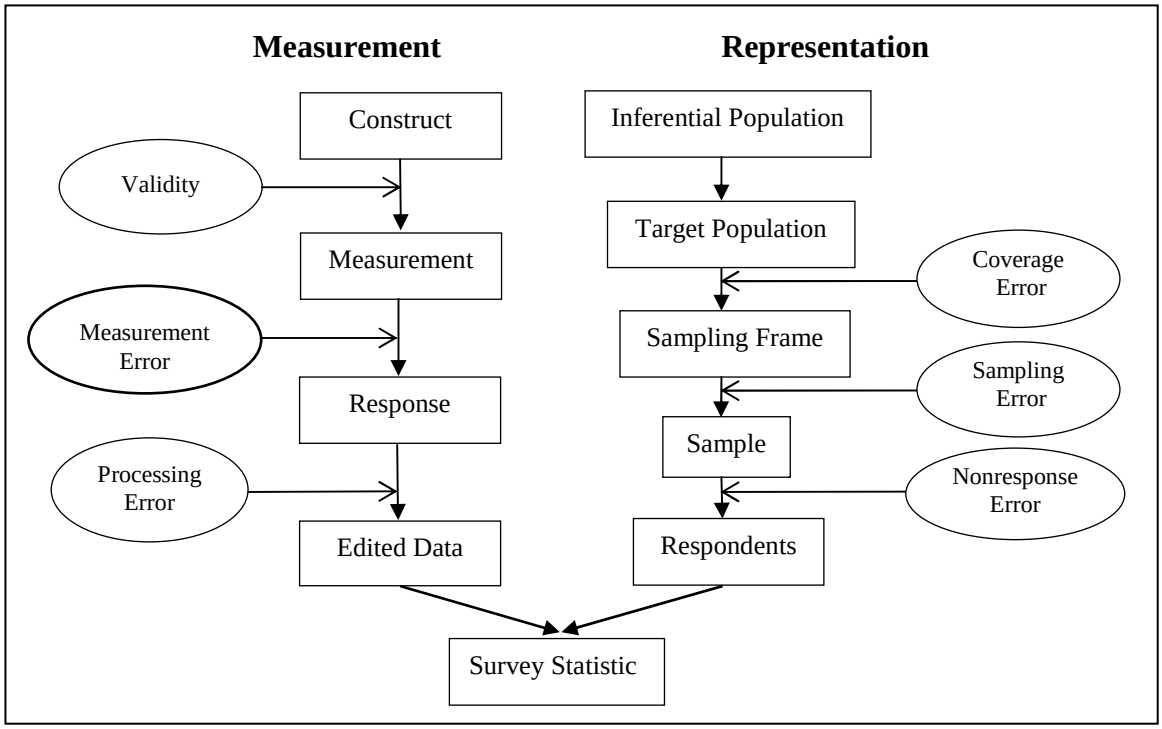

Figure 1.1. Total survey error components linked to steps in the measurement and representational inference process (Groves & Lyberg, 2010, p. 856).

With the entrance of cognitive psychology into the field of survey methodology in the early 1980s – which is typically referred to as "cognitive aspects of survey methodology" (CASM) – more emphasis has been placed on cognitive aspects in question evaluation, to improve the quality of data collection (Fowler, 2013; Miller, 2014). The CASM approach assumes that, when responding to survey questions, respondents are required to go through a series of complex cognitive processes. Understanding these processes is pivotal to question design and the classification and reduction of the different types of response error (Schwarz, 2007; Tourangeau, Rips, & Rasinski, 2000; Willis, 2005).

Tourangeau's four-stage question-answer model is the most widely cited. It divides the response process into four distinct steps that respondents have to complete in order to answer a question. Respondents must comprehend the question

or item, retrieve relevant information, make use of the information to form a judgment, and report to the question or item by selecting a response (Bradburn, 2004; Collins, 2015; Schwarz, 2007; Tourangeau, 1984). The four steps won't necessarily be followed in a linear sequence, beginning with comprehension of the question and ending with reporting an answer. This process, instead, involves moving back and forth, multiple iterations, and overlaps between these steps (as illustrated in Figure 1.2). Some of the processes may even be skipped completely (Bradburn, 2004; Tourangeau et al., 2000).

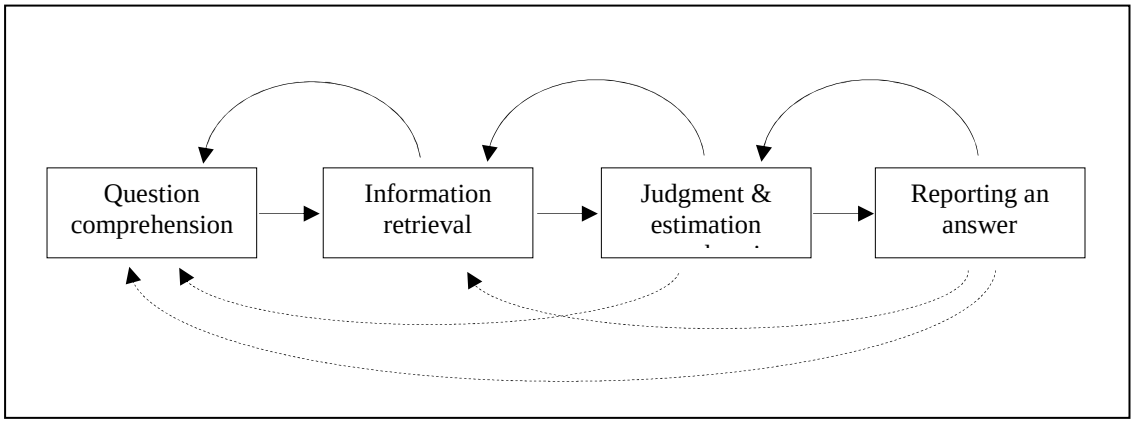

Figure 1.2. A model of the four-step survey response process (Groves et al., 2009, p. 218).

Accurate responses can only be expected when respondents move carefully and thoroughly through all four steps of answering a question (termed *optimizing* respondent behavior, Krosnick & Alwin, 1987; Krosnick, 1991). Depending on the question, this can be quite demanding, requiring substantial cognitive effort from the respondents. In contrast, *satisficing* occurs if respondents take shortcuts and perform the response steps only superficially, for example, when selecting neutral response categories, searching their memories less thoroughly, or when giving random guesses (Tourangeau & Bradburn, 2010). Satisficing can also occur in the form of acquiescence bias: the tendency to agree, regardless of the content (Schuman & Presser, 1981). How much cognitive effort respondents are willing to invest at each of the four stages and the likelihood of satisficing depends on the difficulty and complexity of the task involved (e.g. question difficulty), respondent's cognitive

ability, and respondent's motivation (Biemer & Lyberg, 2003; Cannell et al., 1981; Groves et al., 2009; Krosnick, 1991; Schwarz & Hippler, 1991). Each of the steps and the corresponding difficulties will be outlined in more detail in the following. This provides a useful framework to describe where cognitive pretesting methods have to be applied to uncover sources of error during question completion.

Comprehension. In the first stage of the question-answer process – question comprehension – respondents have to understand and interpret the meaning of the question and the underlying response task. Comprehending a question involves not only decoding the *literal* meaning of the question, but also to infer why the question is being asked (the question designer's intention) and what constitutes an appropriate answer to this question, which is referred to as the *pragmatic* meaning of the question (Clark & Schaeffer, 1989; Schwarz, Groves & Schuman, 1995). Difficulties at this stage may arise because respondents may not notice or understand instructions given to answer the question or, having noticed, they may not bother to read or follow them.

The question may include terms that are unfamiliar to the respondent or terms that are vague or undefined, which can then be understood in different ways by different respondents. If respondents differ in their understanding or interpretation of the question's intent, or of single words, comparisons between their answers will not be valid. Misunderstandings can even occur in questions using common terms, such as "weekday", "children" or "regularly" (Belson, 1981). Additionally, respondents may simply ignore definitions of unfamiliar or technical terms when they are provided with the question (Conrad, Couper, Tourangeau, & Peytchev, 2006). Further types of comprehension problems might occur when words have different meanings (lexical ambiguity) or are used in different ways (structural ambiguity) (Bradburn, 2004).

How well a question is understood also depends on the question length. It is generally recommended to ask short and simple questions and to avoid long or complex questions, to increase question comprehensibility (e.g., Belson, 1981). If a question is too complex, it may simply overload the cognitive resources of the respondents, so that the likelihood that they will be able to perform the

comprehension process accurately and thoroughly decreases (Tourangeau et al., 2000). On the other hand, long questions may help respondents by providing more information, e.g., through explanations or clarifying clauses. Question designers therefore face a trade-off between designing short questions and being more precise and, thus, making the questions more complex and difficult to understand (Bradburn & Sudman, 1991). Another type of problem involves questions that may contain presuppositions or assumptions that are not appropriately or are not accepted by the respondent (so-called faulty presuppositions, Groves et al., 2009).

If respondents have problems in understanding a question, they might interpret all kinds of design features as a source of information that determines what is expected of them and helps them to solve the cognitive tasks required to give an answer; these design features include the position of the question in the questionnaire, the number and order of response options, and the visual design (Schwarz & Hippler, 1991; Sudman & Bradburn, 1982). Respondents also use pictures (Couper, Conrad, & Tourangeau, 2007) or the response categories themselves (e.g., Winkielman, Knäuper, & Schwarz, 1998) as information to clarify the question's meaning. Thereby, respondents consider that the question designer has selected the response categories carefully, to provide more information about what the researcher is interested in and what is appropriate to be reported[5] (Bradburn & Sudman, 1991; Schwarz & Hippler, 1991). An essential component to minimize systematic and variable errors when using questions as measures in surveys is, thus, to ensure that all respondents are able to understand the questions unambiguously (Collins, 2015; Tourangeau & Bradburn, 2010).

Retrieval of information. Once respondents have comprehended the question, they then (usually) have to retrieve the needed information from memory. This step involves adopting a recall strategy, generating retrieval cues that help to recall the information needed, recalling memories, and reconstructing partial memories through inference (Tourangeau et al., 2000). When pretesting survey questions, key

[5] Schwarz et al. (1985) showed that the range of response frequencies, presented either as high- or low-frequency response alternatives, served as information about what is considered to be "normal" TV consumption, which thus affected respondents' estimates.

issues are to assess whether and how well information can be recalled – What types of information does the respondent need to recall, in order to answer the question? – and the recall strategy the respondent uses to retrieve the information (Krosnick & Presser, 2010; Willis, 2005).

Difficulties in providing the information being requested occur because respondents are either not willing or able to expend the cognitive effort necessary to thoughtfully search their memory (Tourangeau et al., 2000). How successful respondents are in retrieving the information required and how accurate their memories are is determined by several factors. First, the distinctiveness or salience of the events is an important aspect. Events that were emotional, important, or remained unique are easier to recall because the memory trace is then stronger and less effort is necessary. Second, it is easier to retrieve information if there is a fit between the terms used in the question and the original experience or event and if the question contains cues that support the respondent's own recall strategies. Further factors that affect the ability to retrieve accurately are the memory sources – is it firsthand experience or not – the recall order, and how long ago the events occurred. Events that happened long ago are generally harder to recall (Tourangeau et al., 2000). In addition, all the information relating to the question (question wording, precoded lists of response alternatives, preceding questions, larger context in which the question is asked, survey material, images, emotions, etc.) serves as retrieval cues that activate and guide the memory search process to the information being requested (Bradburn, 2004; Bradburn & Sudman, 1991). When taking respondents' willingness into account, it cannot be assumed that respondents invest more cognitive effort than necessary. Rather, they simply search their memories for relevant information until they reach some sort of estimate. Commonly, this will be information that is most easily accessible at this moment (Krosnick, 1991).

Judgment and estimation. In the third stage of the question-answer process, respondents have to combine and integrate all information that they have retrieved from memory to come to a judgment. According to Tourangeau at al. (2000), this process involves cognitive tasks such as evaluating the relevance and completeness of the recalled information, drawing inferences based on the information that was

available most directly, adjusting for what is missing (memory gaps), or combining and integrating the information retrieved. If a question asks for chronically accessible information that is well rehearsed, or for which the respondents have a pre-defined position, respondents may retrieve the answers directly. In contrast, when respondents are asked questions about behaviors or attitudes that they have never or rarely thought about, they have to form an opinion or come up with an answer immediately. Further problems may arise if the information being requested is incomplete, for example, due to insufficient recall, or if forming a judgment requires complex estimations or difficult mental calculations. In these cases, respondents may either be unable to provide the information or unwilling to devote sufficient mental effort to answering the question accurately and thoughtfully (Willis, 2005). Consequently, respondents may take short cuts or simply interpret a question superficially. The crucial point at this stage is to check whether respondents are able to provide the information being requested and to decrease task difficulty, to obtain more accurate respondent self-reports (Krosnick & Presser, 2010).

Response. In the final stage of the question-answer process, respondents have to select and communicate an answer. This stage involves two separate processes when responding to a question: *formatting* and *editing* the response (Tourangeau et al., 2000). Having formed a judgment, the respondents are asked to fit their internally generated answer to the response categories provided by the survey question or, although less often, to report it in their own words in an open format. Before reporting the answer, respondents may want to edit it for reasons of consistency (Clark & Schober, 1992), social desirability[6], and self-presentation (Schwarz & Oyserman, 2001).

Even if respondents know how they want to answer to a question, they may encounter problems during the formatting and editing stage (Tourangeau et al., 2000). The internally generated answer might, for example, not fit into the response options given, or the presented options might be too vague or too broad. If the

[6] Social desirability is more common in interviewer-administered modes and can be reduced by having respondents complete a self-administered questionnaire or by procedures such as the randomized response technique (Tourangeau & Yan, 2007).

answer options provided do not match the respondents' judgments or if more than one response option may present a reasonable answer, respondents may choose undesirable approaches, in order to provide a response. For example, they may then choose the first response that seems to be acceptable to meet the question's perceived requirements and then continue with the next question, disregarding the remaining response options (Groves et al., 2009; Krosnick, 1991; Krosnick & Alwin, 1987; Smyth, Dillman, Christian, & Stern, 2006), provide a neutral response (e.g., neither nor), or choose other short cuts such as saying that they simply do not know the answer (Krosnick, 1991; Tourangeau & Bradburn, 2010; Tourangeau et al., 2000). Therefore, how respondents decide to answer a survey question depends strongly on the choice of response options provided (Schwarz & Hippler, 1991). Moreover, the selection of response options may affect the entire question-answer process, in particular, the way in which participants comprehend and interpret the question asked, how they recall information, and which judgment strategies they use (Collins, 2015; Schwarz & Oyserman, 2001).

Using questionnaires in a self-completion format implies further sets of pitfalls, because respondents not only have to understand the question but also related instructions, definitions, visual aspects, such as the graphical layout of the questionnaire, and other navigational issues such as skip patterns in paper-and-pencil questionnaires (Jenkins & Dillman, 1997; Tourangeau & Bradburn, 2010). Because an interviewer is not present, no one can clarify questions related to the entire questionnaire or individual questions, provide additional advice, or explain unclear terms. Opportunities to probe for incomplete or ambiguous information are also not available (Biemer & Lyberg, 2003). To avoid errors, several recommendations for the design of self-administered questionnaires have been proposed (e.g. Jenkins & Dillman, 1997; Couper, 2008, for web surveys). Nevertheless, the graphical and visual design of self-administered questionnaires and its potential consequences should be included in question testing (Couper, 2008; Presser et al., 2004; Schwarz & Oyserman, 2001).

To summarize, the only way to minimize respondents' contribution to measurement error is to reduce the respondents' burden and to minimize the chance

of respondents' adopting response strategies that might affect data quality adversely (e.g., satisficing, Krosnick, 1991). This can be achieved by reducing the difficulty and cognitive effort required to comprehend and answer a survey question (Biemer & Lyberg, 2003). Therefore, survey researchers have to check for cognitive difficulties and identify what causes these difficulties by pretesting their questionnaire (De Leeuw et al., 2007; Fowler, 2013; Miller, 2014).

1.2 Cognitive pretesting methods

The awareness of survey researchers about the need to check whether questions are understood as intended, how difficult they are, and whether they pose other cognitive problems for the respondents prior to fielding them has increased in recent decades (Conrad & Blair, 2009; Presser et al., 2004). Whereas conventional pretests assume that problems with questions will be indicated by respondents' answers (e.g. refusals, response of don't know) or by other overt behavior (hesitation, discomfort during responding), cognitive pretesting methods aim at revealing potential problems during the question-response process, so that measurements really meet the intended objectives (Presser et al., 2004). For example, the intent of a question can be misunderstood by a respondent without any signals that indicate that a problem exists. Cognitive question evaluation methods are used to expose these problems and to point to potential solutions (Conrad & Blair, 2009; Yan, Kreuter & Tourangeau, 2012). They classify questions as either problematic – having problems that require revision of the question – or non-problematic (Yan, et al., 2012). To identify question flaws and assess task difficulty, survey methodologists have several methods at hand, such as cognitive interviews, behavior coding, response latency measurement, formal respondent debriefings, interviewer debriefings, and expert reviews (Presser et al., 2004). Each method has a different focus and provides different information about potential question problems (Collins, 2015; Krosnick, 1999). In addition, the methods differ with regard to timing in the data collection process and whether or not they are byproducts of the answer process (Collins, 2015). For example, response latency analysis, which measures the time lapse between the presentation of a

question to the respondent and the indication of a response, can be directly integrated into the data collection process. Response latencies are then used as an indicator of task difficulty. The underlying assumption is that more complex questions, or questions that require more cognitive effort, have longer response latencies (Bassili, 1996; Collins, 2003; Draisma & Dijkstra, 2004). In the following, I will focus, in particular, on eye tracking because it can be used either in conjunction with cognitive interviewing or as a stand-alone method. For later reference, I will also briefly introduce cognitive interviewing.

1.2.1 Eye Tracking

The aim of this section is to describe what eye tracking is and how it can be used for question pretest and evaluation.

Eye tracking is a technique whereby people's eye movements are recorded and measured while they move across visual stimuli such as texts, images, computers, videos, etc., to provide information on the distribution of visual attention and information processing. Eye-tracking data record the exact location of eye gaze, the duration of fixation, and the sequence of eye gazes. It hence provides information where respondents look at any given time, for how long they look at, and in what order (Romano Bergstrom & Schall, 2014).

The use of eye tracking has a long tradition in studying cognitive processing during reading and other information processing tasks, such as scene perception or usability testing (Duchowski, 2003, for a review; Rayner, 1998). More recently, the technique has also been introduced into the field of survey methodological research to study cognitive processes during survey responding. Eye tracking makes it possible to observe and record respondents' eye movements in real-time while they are completing a questionnaire. Specifically, eye tracking enables the researcher to see where and for how long respondents look when reading and answering survey items. This feature can be used to detect questions that are difficult to understand or that are otherwise flawed (Galesic & Yan 2011; Graesser et al., 2006).

The relationship between eye movements and cognitive processing is based on two key assumptions that were presented by Just & Carpenter (1980): the

immediacy assumption and the *eye-mind assumption*. The *immediacy assumption* postulates a close connection between the visual object viewed and the content being thought about, meaning that words or visual objects that are fixated by the eyes are immediately processed (the mind follows the eye). The second assumption, the *eye-mind assumption*, states that words or visual objects are fixated as long as they are being processed. According to this assumption, what is being fixated by the eyes indicates what is being processed in the mind (the eye follows the mind, 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: the time a respondent spends fixating a word or a particular area of the screen can be taken as a measure of the processing time associated with that word or area (Just & Carpenter, 1980; Staub & Rayner, 2007). Or as Just & Carpenter (1980) put it: "Readers interpret a word while they are fixating it, and they continue to fixate it until they have processed it as far as they can" (1980, p. 350).

Consequently, increased cognitive demand or processing difficulties are reflected in patterns of repetitive fixations, fixations located close together, or patterns of increased fixation duration (Rayner et al., 1981). Rayner (1998) observed that, when text is difficult to process, the frequency of regressions (i.e., backward eye movements through the text) and the duration of fixations increase. Furthermore, unusual or low-frequency words are fixated longer, ambiguous or unfamiliar words are read multiple times, and highly predictable words are often skipped (Rayner & Pollatsek, 2006; Rayner, 2009).

For questionnaire pretesting, this means that questions that are difficult to comprehend should take longer to process and this should be reflected in longer fixation times. Respondents trying to make sense of a word or an entire question will re-read it and backtrack as they scan and rescan it. Thus, eye tracking can point to words in a question that take longer to process, perhaps because they are complex or more difficult to understand (Graesser et al., 2006; Holleman & Murre, 2008; Lenzner, Kaczmirek, & Galesic, 2011).

In addition to indicating questions that are difficult to understand, eye tracking can also be used to find out whether respondents read the questions and response options in the intended order, whether they skip (parts of) questions, whether they read questions to the end or rather skim the question text and then move immediately to the response options, and whether respondents read all response options thoroughly, or just quickly scan them, to provide an answer. Eye movements also reveal 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 read them or not.

When evaluating questions, respondents' eye movements can also be used to answer practical usability questions (Galesic & Yan, 2011) or questions regarding the visual layout or specific visual design elements used to create surveys, e.g. the use of colors or pictures, but also where to place important information, how to design the screen or the arrangement of long lists of response options (Couper, 2008). Data about eye movements, furthermore, provide information on how respondents work with a questionnaire and how easy or difficult it is for them to navigate through the questions and to provide the requested information. Additionally, eye-tracking data provide objective information about what visual aspects of a question (e.g. layout of instructions, response options, questions) draw the initial and most attention and helps to identify areas or elements on a screen that are given too much or too little attention (Galesic & Yan, 2011).

Because recording respondents' eye movements is relatively unobtrusive, eye tracking is an objective way of collecting information about how respondents are interacting with a questionnaire (their true response behavior) and how they are processing the response task. Thereby, eye tracking is independent of respondents' memory, verbal abilities, problem awareness, and subjective judgments (Galesic & Yan, 2011; Romano Bergstrom & Schall, 2014).

In survey methodology, eye tracking was first introduced with a study by Redline & Lankford (2001), who evaluated visual designs of routing instructions in a self-administered paper questionnaire. They found that the notification of branching instructions depends on the position and is recognized best if respondents observe the

instruction immediately before or after marking their answers. Galesic et al. (2008) and Kunz & Fuchs (2012) extended work on clarification features in web surveys. By comparing whether definitions of survey concepts should be always visible or only on request when rolling the mouse over a term, Galesic et al. (2008) found that the chances of being read are higher if important definitions are always visible on the screen. Kunz & Fuchs (2012) used eye tracking to investigate the optimal position of definitions, retrieval cues, and formatting instructions for supporting respondents in answering open-ended questions within different stages of the question-answer process. Their results suggested that instructions should be placed directly where respondents need them. Definitions, for example, should be displayed before the question text, whilst formatting instructions should be placed next to the answer options (Kunz & Fuchs, 2012).

Eye tracking has also been used to explore visual attention and design in web surveys, for example, to evaluate response order effects (Galesic et al., 2008), to explore the visual design of response formats (Lenzner, Kaczmirek, & Galesic, 2014), or for comparing how often and long respondents looked at the labels in either fully-labeled or end-labeled rating scales with five or seven categories, respectively (Menold et al., 2014). Thereby, the analysis of eye-tracking patterns provides insights into how respondents' attention can be improved, depending on visual aspects of a questionnaire. Galesic et al. (2008) found that primacy effects occur because respondents spent more time processing response options presented in the first half of a list than response options presented in the second half, regardless of their content. Moreover, they observed that some respondents did not read the last response option at all. With the help of eye-tracking technology, the authors were able to demonstrate visually what had been long thought to occur. Another experimental study used eye movements to examine whether answer boxes should be placed to the left or to the right of the answer options in web surveys (Lenzner et al., 2014). The authors found that placing answer boxes to the left of response options decreases response latencies, decreases fixation times and counts on the answer boxes, and decreases the number of gaze switches between answer boxes and answer options. They concluded that placing the answer boxes to the left enhances usability

by making it easier for respondents to select an answer, which thus facilitates the overall response task.

A few studies have used eye movements to evaluate the effects of question wording on question comprehensibility (Graesser et al., 2006; Lenzner et al., 2011; Kamoen et al., 2011). Graesser et al. (2006), for example, collected eye-tracking data while respondents answered questions that had been identified either as problematic (containing difficult text features) or not. They found that questions identified as problematic were processed differently than the non-problematic ones: Content words with unfamiliar technical terms had longer total fixation times, longer first fixation times, and more fixation counts than words that were defined to be non-technical terms. When questions contained a complex or difficult syntax, respondents tend to give up answering by using an early exit from reading the question. Lenzner et al. (2011) added to this line of research: the authors investigated the processing of two versions of similar questions containing either one of seven problematic text features (e.g., low frequency words, vague or imprecise relative terms, complex syntax) or none (text feature version vs. control version) by examining respondents' fixation times and counts. The results revealed that respondents had longer fixation times and more fixation counts in the text feature questions than in the control questions, which indicates higher cognitive effort. Eye tracking methods were also used by Kamoen et al. (2011) to examine the cognitive processes involved in answering contrastive survey questions. The results revealed that negatively worded questions and their response options were reread more frequently than positively formulated questions (Kamoen et al., 2011).

Recently, Kaminska & Foulsham (2014) explored, in a small feasibility study, the use of real-world eye tracking, to compare visual attention in different survey modes (SAQ, web, and PAPI). Due to changes in posture of participants, which resulted in insufficient data quality, SAQ had to be excluded from data analysis and could not be compared. However, the authors were able to detect some differences in how respondents process survey questions in PAPI and web. They found, for example, that the time spent on question wording does not differ largely

whether an interviewer reads a question out loud or whether a respondent reads a question in web mode.

While initial eye-tracking equipment was often invasive and caused discomfort to the users, for example, by placing several electrodes on the skin surrounding the eye or by using contact lenses holding a mirror next to the pupil, there are now apparatuses that do not need any form of special lenses and electrodes (Galesic & Yan, 2011; Hammoud & Mulligan, 2008) and are relatively reliable, less intrusive, and easy to use[7] (Jacob & Karn, 2003). Most of the eye trackers currently used in usability labs are based on the pupil center/corneal reflection method to follow and track the eyes while they move; they are also called *video-based* eye trackers. These eye trackers usually operate with (near-)infrared light and a video camera to image the eye. The camera is placed either underneath or next to a computer monitor on which the participant is performing a task (remotely mounted) or mounted on the participant's head (head-mounted)[8]. With the pupil center/corneal reflection method, near-infrared light is directed into the eye where it meets the retina and causes a reflection. The back-reflected light is then sensed by the infrared-sensitive camera. The image captured by the camera is used to identify the center of the pupil and the location of the corneal reflection. The separation of these two features is analyzed (using advanced image processing algorithms) to determine where the user is looking (Duchowski, 2003; Jacob & Karn, 2003). In order to set the eye tracker up for each respondent and to lessen gaze tracking errors due to individual differences, a calibration procedure is required, in which the respondent looks at dots appearing on the screen. During the personal calibration process, the

[7] Currently, vendors of eye tracking systems provide software for set up of the apparatus, calibration procedures, and for data analysis. This development has made data collection and extraction less time consuming and labor-intensive (Jacob & Karn, 2003).

[8] Head-mounted eye trackers allow more freedom of movement, but are more invasive, whilst remote eye trackers can be completely unobtrusive and are more comfortable for the participants. Moreover, they allow a more natural experience for the users. On the negative side, unobtrusive eye tracking is less precise in recording and it might not be precise enough to determine exactly which words respondents read when the words are presented in normal font size (Galesic & Yan, 2011; Jacob & Karn, 2003). However, the available accuracy and precision are satisfactory for most practical applications (Galesic & Yan, 2011; Hammoud & Mulligan, 2008).

eye tracking system measures characteristics of the user's eyes and records the pupil-center/corneal reflection and the value that corresponds to each gaze position (as x-y coordinates; Duchowski, 2003).

There are different types of eye movements that can be analyzed to understand visual attention (Rayner, 1998). The main measurements typically analyzed are fixations and saccades. Fixations are moments in which the eyes remain relatively motionless and pause on a specific area of the visual field. During fixations, meaningful information is extracted and new information is encoded. Fixations can be measured by the frequency and length of time with which an object is viewed. Saccades are rapid eye movements occurring between fixations. Saccades serve to reorient the eye and to move target words into foveal focus, so that they can be fixated and processed. No information is obtained during saccades (Duchowski, 2003; Rayner, 1998; Staub & Rayner, 2007).

Besides analyzing metrics such as time to first fixation, fixation duration, or fixation count, an eye tracker also allows researchers to generate heat maps that can be used to visualize specific areas of interest, areas that received too little attention, or so-called gaze plots. Gaze plots show the order and sequence of respondents' eye movements as they move across the screen and are useful for illustrating typical behaviors displayed when navigating and completing online questionnaires (Romano Bergstrom & Schall, 2014).

1.2.2 Cognitive Interviewing

Since the mid 1980s, cognitive laboratory techniques, in particular cognitive interviews, have emerged from the CASM movement (Beatty & Willis, 2007; Forsyth & Lessler, 1991; Willis & Miller, 2011). Beatty & Willis (2007) define cognitive interviewing as "the administration of draft survey questions while collecting additional verbal information about survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends" (2007, p. 287). Cognitive interviewing focuses on respondents' thought processes while answering survey questions, and errors that may arise during this process (Beatty & Willis, 2007;

Miller, 2011; Willis & Miller, 2011). The verbal material gathered by the interviews is used to diagnose problems and to evaluate the quality of the questions (Beatty & Willis, 2007; Presser et al., 2004). The goal is to use this information to find better ways of constructing, formulating, and asking survey questions and to find out how they should be modified to make them easier to answer (Forsyth and Lessler, 1991; Willis & Miller, 2011). By identifying problems with particular questions and providing hints on how to revise them, cognitive interviewing contributes to reducing measurement error (Conrad & Blair, 2009; Forsyth & Lessler, 1991; Willis, 2005).

The cognitive interview is typically a semi-structured, in-depth interview with small sample sizes of 10 to 30 people. When conducting cognitive interviews, the most commonly used techniques are *think aloud* and *verbal probing* (Willis, 2005). During think aloud, respondents are asked to report everything that comes to their mind while they are forming an answer. During verbal probing, the interviewer asks direct questions or probes, after administering the questions, to obtain more information about how respondents interpreted and answered them or about how they interpreted specific terms (Beatty & Willis, 2007; Willis, 2005). In practice, often a combination of both variants is applied, as they "fit together very naturally" (Willis, 2005, p. 57). When conducting cognitive interviews, the interviewers normally use a cognitive interview protocol consisting of the questions to be tested and pre-scripted probes to search for problems (Willis & Miller, 2011). The cognitive techniques can either be administered immediately after the subject has answered the targeted survey question (concurrent probing) or at the end of the interview (retrospective probing; Collins, 2003; Willis, 2005; Willis & Miller, 2011). Probing questions are often designed to investigate a specific cognitive process (e.g., there are comprehension probes, recall probes, and so on[9]). In addition to pre-scripted probing questions that are developed prior to the interview, emergent probing questions can be asked in case problems that had been unanticipated arise during the interview. Such probes are flexible and reactive because the interviewer chooses spontaneously what to ask

[9] An example of a probe targeting the response stage is: "How easy or difficult was it to find your answer on that list?" (Willis, 2005).

in response to what the participant says (Willis & Miller, 2011). After the interview, the verbal reports produced have to be analyzed and interpreted to define whether or not a question poses a problem for respondents (Beatty & Willis, 2007).[10] To analyze the data, the comments of the participants are successively aggregated (Willis, 2005) and summarized for each survey item. Occasionally, problem classification schemes (DeMaio & Landreth, 2004; Presser & Blair, 1994) are applied that classify problems according to the four stages of the survey response process (Willis & Miller, 2011).

Although there is general agreement about the value of cognitive interviewing, it has also some limitations (Collins, 2015; Presser et al., 2004). First, it is a qualitative method that produces verbal data that have to be interpreted by the researcher and that are, therefore, subjective (Beatty & Willis, 2007; Conrad & Blair, 2009). Second, some respondents find it difficult to express themselves verbally (Graesser et al., 2006) and to report on their cognitive processes, because not all such processes are conscious (Collins, 2015; Willis, 2004). In particular, respondents with relatively low levels of education and cognitive skills often find it difficult to report on these processes (Galesic & Yan, 2011; Sudman, Bradburn, & Schwarz, 1996). Moreover, respondents may not always themselves be aware of having a problem with answering or comprehending the question (Campanelli, 2008). And, finally, the cognitive techniques and the behavior of the interviewers may have an impact on the ways respondents answer the questions (Beatty & Willis, 2007; Conrad & Blair, 2009; Willis, 2005).

1.3 Contributions at a glance

Using three novel experiments, the next chapters investigate the productivity of eye tracking in question design and problem detection, both in combination with cognitive interviewing or as a stand-alone technique. This section summarizes the three studies and the key results.

[10] For more practical information on cognitive interviewing and its varieties, see Willis (2005) and Collins (2015).

Chapter 2 (“Incorporating eye tracking into cognitive interviewing to pretest survey questions”) and chapter 3 (“A comparison of two cognitive pretesting techniques supported by eye tracking”) are concerned with eye tracking in combination with cognitive interviewing.

The former chapter presents a controlled experiment designed to test whether a joint implementation of eye tracking and cognitive interviewing is more productive in pretesting self-administered questionnaires than standard cognitive interviews alone by comparing both the total number of problems detected and the number of questions identified as flawed. In the control condition, a cognitive interview was conducted using a standardized interview protocol. In the treatment condition, respondents’ eye movements were tracked while they completed an online version of the questionnaire. In the subsequent cognitive interview, interviewers used the data to identify potential problems and ask targeted probing questions in addition to the probes scripted in the interview protocol. The results show that cognitive interviewing and eye tracking complement each other effectively. The hybrid method detected more problems and identified more questions as problematic than applying cognitive interviewing alone. With regard to the types of problems detected, both experimental conditions produced almost identical results.

Chapter 3 builds upon the previous study by examining how eye tracking assists cognitive interviewing most effectively. To this end, two retrospective probing techniques are compared: Retrospective probing based on observed eye movements (as used in chapter 2) and gaze video cued retrospective probing. In the latter, a video of their own eye movements is shown to the respondents during the cognitive interview. The motivating hypothesis is that this technique could be more productive because respondents are reminded of their answer process by the additional visual cue. The two conditions are compared with regard to the number and types of problems identified and the way they stimulate respondents when commenting on their behavior. The results show that both techniques did not differ in terms of the total number of problems identified. However, video cued retrospective probing identified fewer unique problems and fewer types of problems than pure retrospective probing. Additionally, when seeing a video of their own eye

movements, participants commented more on what they were *doing* and less on what they were *thinking* when answering questions.

In chapter 4, eye-tracking data are used to gain information about the cognitive processes underlying respondents' behavior when answering questions in two different response formats (check-all-that-apply vs. forced-choice) and, accordingly, whether and why one of those formats is more susceptible to problems in the response process. Both question formats are compared using the amount of attention paid to the questions and the cognitive effort (operationalized by response latencies, fixation times, and fixation counts) respondents spent while answering one factual and one opinion question, respectively. No difference in cognitive effort spent on the factual question was found, whereas, for the opinion question, respondents invested more cognitive effort in the forced-choice than in the check-all-that-apply condition. The observation of participants' reading behavior did not reveal differences in the number of options read across question formats.

Versions of chapters 2 and 3 have been published or accepted for publication as:

1. Neuert, C. E., & Lenzner, T. (2015). Incorporating eye tracking into cognitive interviewing to pretest survey questions. *International Journal of Social Research Methodology* online first.

2. Neuert, C. E., & Lenzner, T. (2015). A comparison of two cognitive pretesting techniques supported by eye tracking. *Social Science Computer Review* online first.

A version of chapter 4 is under review as:

3. Neuert, C. E. (under review). Processing forced-choice versus check-all-that-apply question formats - Evidence from eye tracking.

1.4 Conclusion

It is generally acknowledged that new questions or survey instruments require some form of pre-evaluation before they are actually fielded, in order to check their validity and minimize measurement error. This is the task of questionnaire pretesting. The present thesis contributes to survey pretesting methodology by examining the productivity of eye tracking in problem detection and question design. Several insights can be drawn from the research presented here.

Overall, the studies provide evidence that eye tracking is a valuable addition to the methodological toolbox in questionnaire design and pretesting. Two reasons are highlighted:

First, eye tracking can supplement cognitive interviewing. With instant access to respondents' eye movements, the cognitive interviewer or survey researcher obtains a richer picture of the response process and is able to ask more targeted probing questions. This contributes to the value of standard cognitive interviewing: it helps to detect problems that are not consciously apparent to the respondents, and illustrates problems visually that are difficult to express verbally by the subjects themselves. Monitoring respondents' eye movements also permits to testing hypotheses regarding response strategies, such as satisficing in the setting of chapter 4.

Second, eye movement recordings are a source of objective data that can be analyzed quantitatively. The verbal data gathered from the cognitive interviews can be compared with the eye-movement data to crosscheck and confirm the conclusions drawn. Additionally, they can be used as an indicator of cognitive effort.

There are some caveats, though[11]. First, the setup costs of an eye tracker are relatively high. It seems advisable to assess whether the expected additional insights are worth the financial investment required. For large, specialized pretesting

[11] There are also some specific limitations to the experiments outlined in chapters 2 to 4. These are addressed in the specific chapter discussions.

laboratories, this is naturally more likely than for small ones or for researchers planning one-shot pretests.

Second, there are technical limitations in recording accuracy for some participants, such as wearers of glasses or contact lenses. This demands more effort for gathering a suitable sample of participants (e.g., older adults are more likely to wear eye glasses and do not track as well as younger participants; Loos & Romano Bergstrom, 2014). Inaccuracy of recordings or systematic shifts of eye movements that prevent a precise (quantitative) analysis of the data can also occur if respondents change their position substantially while filling-in the questionnaire. The calibration and tracking process therefore needs to be carefully monitored in order to minimize such errors.

Third, eye tracking is limited to visually presented stimuli and thus to pretest visual survey instruments, such as web or self-administered questionnaires. In contrast, other pretesting methods such as cognitive interviewing can be used with all modes of questioning (including personal-oral or telephone).

Fourth, the interpretation of eye-movement data is not always straightforward. Eye movements alone can only point to difficulties, but they generally do not provide complete information about the kind and the cause of the problem. Peculiar reading patterns, such as repetitive eye movements could indicate that a problem exists, but this may be due to unfamiliar terms, complex syntactical structures, incorrect presuppositions, or other question flaws. Moreover, peculiar reading patterns are not problematic per se. They could also indicate a respondents' increased interest in the question or a relatively conscientious response style, as is shown by respondents who optimize (Lenzner et al., 2011). Thus, the interpretation of eye movement metrics depends strongly on the context and the underlying task.

Fifth, this thesis is based on the premises that there is a direct link between eye movements and cognitive processes (as presented in section 1.2.1, Just & Carpenter, 1980). Those assumptions are generally accepted in the current literature, and are supported by direct evidence (Aslin & McMurray, 2004; Balota, Pollatsek & Rayner, 1985; Just & Carpenter 1976, 1980; Lass & Lüer, 1990; Poole & Ball, 2005; Rötting, 1999; 2001; Schroiff, 1986; Velichkovsky, 2001). Nevertheless, there may

be some issues, e.g., a covert shift of attention (Findlay, 2005) or when people are "looking without seeing" (Joos, Rötting, & Velichkovsky, 2003). According to Duchowski (2003) "An eye tracker can only track the overt movement of the eyes, however, it cannot track the covert movement of visual attention. Thus, in all eye tracking work, […] we assume that attention is limited to foveal gaze direction, but we acknowledge that this may not always be so" (2003, p. 14).[12]

There are several interesting avenues for future research. First, the use of (i) more specific probes (specifically designed to investigate a particular cognitive process), (ii) different probing techniques (think-aloud, retrospective vs. concurrent probing), or probing styles (standardized vs. more flexible), (iii) the use of different eye-tracking procedures (gaze replays with or without gaze overlay), or (iv) testing survey questions with more complex or dynamic interfaces (e.g. lookup databases) could be examined with the hybrid approach developed in chapter 2.

Second, the approach could also be extended to other forms of method integration, for example, pure eye-tracking sessions followed by a time lag to analyze the quantitative data and delayed follow-up probing techniques designed to gather possible explanations for patterns observed in the quantitative data.

A third line of research worth investigating would be to develop an automatic coding system for peculiar reading patterns to detect problems in survey questions based on the reading behavior. This system could then be used to link eye-tracking measures to types of question problems. Moreover, it could be assessed whether it is possible to define peculiar eye movements and to connect these to specific problem types. These findings could then be integrated within different stages of the question-response process (e.g., first pass reading time of the question text as a measure of the comprehension process) to deepen our understanding of the ongoing processes.

Fourth, it could be fruitful to supplement eye-tracking data with other physiological measures, for example, collecting data on pupil dilation. Many eye trackers are able to collect pupil dilation data and including this data could provide

[12] An interesting avenue for future research would be to test this fundamental hypothesis by means of brain imaging technology.

additional information regarding attention, interest, or mental workload (Iqbal et al., 2004; Tullis & Albert, 2008) compared to eye movement data alone. This could thus provide an even deeper understanding of the response process and the underlying task difficulty when answering survey questions.

Fifth, eye-tracking data could also be used to study the optimal design and the comprehensibility of survey invitations, cover letters, survey instructions (long vs. short, providing much or little information regarding the questionnaire), consent forms, or welcome pages (e.g., what to put on the screen and where to put the most important information). How respondents perceive and interpret various kinds of supplementary material for a survey could provide important information how this affects their general motivation to participate.

Sixth, recent technological advancements, such as eye-tracking glasses or technical solutions for mobile devices, allow for eye-tracking research outside the laboratory in (more) natural settings (see also Kaminska & Foulsham, 2014). This could be especially interesting for surveys using mixed modes, since the respondents' tasks should be identical independent of whether they answer a paper-and-pencil questionnaire or a web questionnaire on either a desktop PC, smartphone, or tablet PC. It would be possible to test whether the task remains the same or which adaptations should be made.

To summarize and conclude, eye tracking is a useful tool in survey pretesting that helps to indicate question difficulties and provides an accurate representation and understanding of respondents' eye movement behavior and the underlying survey response processes. It allows investigators to observe respondents' behavior instead of guessing what could have occurred on basis of a respondent's overt behavior or having to rely on indirect measures (reported responses, response times, and mouse movements). Thereby, eye tracking permits insights that other methods cannot offer and provides added value to test or generate research questions that target uncovering respondents' cognitive processes while responding to survey questions. However, eye tracking will not yield answers to all theoretical questions and will not replace other methods aimed at studying cognitive processes and response behavior.

1.5 References

ASA (1999). Designing a Questionnaire. ASA Series: What is a Survey. *Section on survey research methods. American Statistical Association*. Alexandria, VA.

Aslin, R. N., & McMurray, B. (2004). Automated corneal-reflection eye tracking in infancy: Methodological developments and applications to cognition. *Infancy*, *6*(2), 155-163.

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive psychology*, *17*(3), 364-390.

Bassili, J. N. (1996). The how and the why of response latency measurement in telephone surveys. In N. Schwarz, & S. Sudman (Eds.), Answering questions (pp. 319–346). San Francisco, CA: Jossey-Bass.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71(2)*, 287-311. doi:10.1093/poq/nfm006

Belson, W. A. (1981). *The design and understanding of survey questions*. Aldershot: Gower.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A. & Sudman, S. (1991). *Measurement errors in surveys.* Hoboken, NJ, USA: John Wiley & Sons.

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

Bradburn, N. M. (2004). Understanding the question-answer process. *Survey Methodology*, *30*(1), 5-15.

Bradburn, N. M., & Sudman, S. (1991). The current status of questionnaire research. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 27-40). Hoboken, NJ, USA: John Wiley & Sons. doi: 10.1002/9781118150382.ch2

Campanelli, Pamela (2008). Testing survey questions. In: E. De Leeuw, J.J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 176-200). New York, London: Taylor & Francis

Cannell, C. F., Miller, P. V. & Oksenberg, L. (1981). Research on interviewing techniques. *Sociological methodology*, *12*(4), 389-437.

Clark, H. H., & Schaeffer, E. F. (1989). Contributing to discourse. *Cognitive science, 13*(2), 259-294.

Clark, H. H., & Schober, M. F. (1992). Asking questions and influencing answers. In J. M. Tanur (Ed.), *Questions about questions. Inquiries into the cognitive bases of surveys* (pp. 15-48). New York: Russell Sage Foundation.

Collins, D. (2003). Pretesting survey instruments: an overview of cognitive methods. *Quality of Life Research, 12(3)*, 229-238.

Collins, D. (2015). *Cognitive interviewing practice*. Thousand Oaks: Sage.

Conrad, F. G., & Blair, J. (2009). Sources of error in cognitive interviews. *Public Opinion Quarterly, 73(1)*, 32-55.

Conrad, F. G., Couper, M. P., Tourangeau, R., & Peytchev, A. (2006). Use and non-use of clarification features in web surveys. *Journal of Official Statistics, 22*(2), 245-269.

Couper, M. P., (2008). *Designing effective web surveys*. New York: Cambridge University Press.

Couper, M. P., Conrad, F. G., & Tourangeau, R. (2007). Visual context effects in web surveys. *Public Opinion Quarterly, 71*(4), 623-634.

De Leeuw, E. D., Hox, J. & Dillman, D. A. (2012). *International handbook of survey methodology*. New York, London: Taylor & Francis.

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). Hoboken, NJ: John Wiley & Sons. doi: 10.1002/0471654728.ch5

Draisma, S., & Dijkstra, W. (2004). Response latency and (para)linguistic expressions as indicators of response error. 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. 131–147). Hoboken, NJ: Wiley.

Duchowski, A.T. (2003). *Eye tracking methodology: Theory and practice.* London: Springer Verlag.

Findlay, J. M. (2005). Covert attention and saccadic eye movements. In I. Laurent, R. Geraint & K. T. John (Eds.), *Neurobiology of attention* (pp. 114-116). Burlington: Academic Press

Foddy, W. (1993). *Constructing questions for interviews*. Cambridge University Press.

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). Hoboken, NJ, USA: John Wiley & Sons. doi: 10.1002/9781118150382.ch20

Fowler, F. J. (1995). *Improving survey questions.* Thousand Oaks: Sage.

Fowler, F. J. (2013). *Survey research methods* (5th edition). 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: 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*(5), 892–913.

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: 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(1),* 3–22. doi:10.1093/poq/nfj01

Groves, R.M. (1989). *Survey error and survey costs*. New York: Wiley.

Groves, R. M., Fowler Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd edition). New York: Wiley.

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849-879.

Hammoud, R. I., & Mulligan, J. F. (2008). Introduction to eye monitoring. In R. I. Hammoud (Ed.), *Passive eye monitoring: Algorithms, applications and experiments* (pp. 1-19). Berlin, Heidelberg: Springer Science & Business Media.

Holleman, B. C., & Murre, J. M. (2008). Getting from neuron to checkmark: Models and methods in cognitive survey research. *Applied Cognitive Psychology*, *22*(5), 709-732.

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI'04 Extended abstracts on human factors in computing systems* (pp. 1477-1480). ACM Press.

Jacob, R. J. K., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: ready to deliver the promises. In J. Hyönä, R. Radach & H. Deubel (Eds.), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research* (pp. 573-605). Amsterdam: Elsevier.

Jenkins, C. R., & Dillman, D. A. (1997). A theory of self-administered questionnaire design. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 165-196). New York: Wiley-Inter Science.

Joos, M., Rötting, M., & Velichkovsky, B. M. (2003). Die Bewegungen des menschlichen Auges: Fakten, Methoden und innovative Anwendungen. In G. Rickheit, T. Herrmann & W. Deutsch (Eds.), Psycholinguistics / Psycholinguistik. An international Handbook / Ein internationales Handbuch (pp. 142-168). New York, Berlin: de Gruyter.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology, 8*(4), 441–480.

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

Kaminska, O., & Foulsham, T. (2014). Real-world eye-tracking in face-to-face and web modes. *Journal of Survey Statistics and Methodology*, *2*(3), 343-359.

Kamoen, N., Holleman, B., Mak, P., Sanders, T., & Van Den Bergh, H. (2011). Agree or disagree? Cognitive processes in answering contrastive survey questions. *Discourse Processes, 48*(5), 355-385.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology, 5*(3), 213-236.

Krosnick, J. A. (1999). Survey research. *Annual review of psychology, 50*(1), 537-567.

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

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 263-314). Bingley, UK: Emerald Group Publishing (2nd edition).

Kunz, T., & Fuchs, M. (2012). *Positioning of clarification features in web surveys: evidence from eye tracking data*. Paper presented at the 66th annual conference of the American Association for Public Opinion Research (AAPOR), Orlando, FL.

Lass, U., & Lüer, G. (1990). Blickbewegungsverhalten als Indikator für Gedächtnisbildung. In H. Mühlendyck & W. Rüssmann (Eds.), *Augenbewegung und visuelle Wahrnehmung. Physiologische, psychologische und klinische Aspekte* (pp. 79-82). Stuttgart: Ferdinand Enke.

Lenzner, T., Kaczmirek, L., & Galesic, M. (2011). Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension. *International Journal of Public Opinion Research, 23*(3), 361-73. 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*(6), 743-764. doi:10.1177/0894439313517532

Loos, E., & Romano Bergstrom, J. C. (2014). Older adults. In J. Romano Bergstrom & A. Schall (Eds.), *Eye Tracking in User Experience Design* (pp. 313-329). San Francisco, CA: Morgan Kaufmann.

Madans, J., Miller, K., Maitland, A. & Willis, G. B. (2011). Introduction. In J. Madans, K. Miller, A. Maitland & G. B. Willis (Eds.), *Question evaluation methods: Contributing to the science of data quality* (pp. 1-4). New York: Wiley.

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? *Field Methods*, 26(*1)*, 21-39. doi:10.1177/1525822X13508270

Miller, K. (2011). Cognitive interviewing. In J. Madans, K. Miller, A. Maitland, G. B. Willis (Eds.), *Question evaluation methods: Contributing to the science of data quality* (pp. 51-75). New York: Wiley.

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

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*(1), 74-97.

Poole, A., & Ball, L. J. (2005). Eye tracking in human-computer interaction and usability research: current status and future. In C. Ghaoui (Ed.), *Encyclopedia of Human Computer Interaction* (pp. 211-219). Pennsylvania: Idea Group, Inc.

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results? *Sociological methodology*, *24*(1), 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*(1), 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.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, *62*(8), 1457-1506.

Rayner, K., Inhoff, A. W., Morrison, R. E., Slowiaczek, M. L., & Bertera, J. H. (1981). Masking of foveal and parafoveal vision during eye fixations in reading. *Journal of Experimental Psychology: Human Perception and Performance, 7*(1), 167-179. http://dx.doi.org/10.1037/0096-1523.7.1.167

Rayner, K., & Pollatsek, A. (2006). Eye-movement control in reading. In M. J. Traxler, & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (pp. 613–658). Amsterdam: Elsevier.

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 Bergstrom, J. R., & Schall, A. (2014). Introduction to eye tracking. In J. R. Bergstrom & A. Schall (Eds.), *Eye tracking in user experience design* (pp. 3-46). San Francisco, CA: Morgan Kaufmann.

Rötting, M. (1999). Typen und Parameter von Augenbewegungen. In M. Rötting & K. Seifert, *Blickbewegungen in der Mensch-Maschine-Systemtechnik* (pp. 1-18). Sinzheim: Pro Universitate.

Rötting, M. (2001). *Parametersystematik der Augen-und Blickbewegungen für arbeitswissenschaftliche Untersuchungen*. Aachen: Shaker.

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 61(*4)*, 576-602.

Schuman, H., & Presser, S. (1981*). Questions and answers in attitude surveys: experiments on question form, wording, and context*. New York: Academic Press.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology*, 21, 277-287.

Schwarz, N., Groves, R.M. & Schuman, H. (1995). Survey methods. Survey Methodology Program Working Paper Series. Ann Arbor, Michigan, Institute for Survey Research, University of Michigan.

Schwarz, N., & Hippler, H. J. (1991). Response alternatives: The impact of their choice and presentation order. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 41-56). Hoboken, NJ, USA: John Wiley & Sons. doi: 10.1002/9781118150382.ch3

Schwarz, N., Hippler, H. J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly, 49*(3), 388-395.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation*, *22*(2), 127-160.

Schroiff, H.-W. (1986). Zum Stellenwert von Blickbewegungsdaten bei der Mikroanalyse kognitiver Prozesse. In: L.-J. Issing, H.D. Mickasch & J. Haack (Hrsg.). *Blickbewegung und Bildverarbeitung* (pp. 57-82). Frankfurt a. M.: Peter Lang.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly, 70*(1), 66-77.

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. In: G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 327-342). Oxford, UK: Oxford University Press.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco, CA: Jossey-Bass.

Sudman, S., Bradburn, N., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: CA: Jossey-Bass.

Tourangeau, R. (1984). Cognitive sciences and survey methods. A cognitive perspective. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). Washington DC: National Academy Press.

Tourangeau, R., & Bradburn, N. M. (2010). The psychology of survey response. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 315-346). Bingley, UK: Emerald Group Publishing (2nd edition).

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

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin*, *133*(5), 859.

Tullis, T., & Albert, W. (2010). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics* (pp. 167-188). Morgan Kaufmann.

Velichkovsky, B. M. (2001). Levels of processing: Validating the concept. In M. Naveh-Benjamin, M. Moscovitch & H.L. Roediger (Eds.), *Perspectives on human memory and cognitive aging: Essays in honor of Fergus I. M. Craik* (pp. 48-70). Philadelphia: Psychology Press.

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: Wiley. doi:10.1002/0471654728.ch2

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

Willis, G. B., & Miller, K. (2011). Cross-cultural cognitive interviewing: Seeking comparability and enhancing understanding. *Field Methods, 23*(4), 331-341. doi: 0.1177/1525822X11416092.

Winkielman, P., Knäuper, B., & Schwarz, N. (1998). Looking back at anger: Reference periods change the interpretation of emotion frequency questions. *Journal of Personality and Social Psychology*, *75*(3), 719.

Yan, T., Kreuter, F., & Tourangeau, R. (2012). Evaluating survey questions: A comparison of methods. *Journal of Official Statistics*, *28*(4), 503-529.

2 INOCORPORATING EYE TRACKING INTO COGNITIVE INTERVIEWING TO PRETEST SURVEY QUESTIONS[13]

2.1 Abstract

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.

2.2 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

[13] A version of this chapter has been published as:
Neuert, C. & Lenzner, T. (2015). Incorporating eye tracking into cognitive interviewing to pretest survey questions. *International Journal of Social Research Methodology* online first.
Parts of this chapter were presented at the 5th Conference of the European Survey Research Association (ESRA), July 15-19, 2013, Ljubljana, Slovenia, and at the QUEST Workshop, April 09-11, 2013, Washington DC.

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).

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.

2.3 Background

2.3.1 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).

2.3.2 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.

2.3.3 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 stand-alone 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).

2.4 Method

2.4.1 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.

2.4.2 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.[14] 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.[15] Most participants were experienced computer and Internet users who used computers and the Internet on a daily basis with 88% and 87%, respectively.

2.4.3 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[16]. Most of the questions were adapted from various existing surveys, such as the International Social Survey Programme (ISSP), the German

[14] Excluding these participants does not alter our conclusions. The relevant results are available upon request.
[15] Chi-squared analysis revealed no statistically significant differences between both experimental conditions regarding socio-demographic characteristics, such as gender (χ^2 = .115, df = 1, p = .734), age (χ^2 = 3.696, df = 2, p = .158), and education (χ^2 = .733, df = 2, p = .693).
[16] The questionnaire can be found in Appendix B (section 2.9).

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.

2.4.4 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 17″ 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.

2.4.5 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 & 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.

2.4.6 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.

2.5 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.

2.5.1 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 section 2.8 Appendix A). Individual items could be assigned multiple problem codes.

Table 2.1 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 ($\chi^2 = 2.08$, df = 1, $p = .188$).[17] 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 pre-scripted probes and 69.2% were identified by conditional probing based on peculiar response behavior in the control condition (29.9%) or based on peculiar reading patterns in the treatment condition (39.3%).

Table 2.1. 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%)

Significantly more problems were identified by conditional probing in the treatment condition than in the control condition ($\chi^2 = 8.98$, df = 1, $p = .005$). These findings

[17] We did not expect our results to achieve statistical significance. A power analysis (χ^2 test, $\alpha = .05$) indicated that a minimum sample size of N = 1300 would be required to detect any significant effects of low size (0.1) or a minimum sample size of N = 145 to detect effects of medium size (0.3) (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.

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 2.2), 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 ($\chi^2 = 4.42$, df = 4, $p = .352$).

Table 2.2. 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% (0)	0.5% (2)	0.3% (2)
Total	345	392	737

2.5.2 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.[18] In accordance with recommendations from behavior coding (Blair & Srinath, 2008; Fowler, 1992),

[18] 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).

we coded a question as problematic if at least 15% of the respondents had a problem with the item.[19]

Table 2.3 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 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 ($\chi^2 = 0.98$, df = 1, $p = .645$). In total, 18 of the flawed questions were identified in both conditions, nine by pre-scripted 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 predefined probes. Hence, identification of these latter two problematic questions does not constitute a contribution of eye tracking.

Table 2.3. 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

[19] 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).

2.5.3 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).[20] Subsequently, the ratings were averaged across the experts.[21]

Table 2.4 lists the respective questions together with their severity ratings, sorted by average question severity score (ranging from ∅ 2.5 to ∅ 7.3). 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 (∅ 6.7) 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

[20] 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 our purpose, the results are more intuitively interpretable if only a scale from one to ten 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.

[21] The intraclass correlation between experts was ICC = .44, which is classified as fair agreement (Cicchetti, 1994).

two types of problems that were both in the lowest quartile (∅ 2.5). 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.

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 (∅ 3.4; 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 (∅ 7.3) 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.

Table 2.4. Severity rating and problems identified by method.

Question		Identified in	Problem (code)	Severity ∅
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 • …	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. Completely agree – tend to agree – tend to disagree – completely disagree	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
Q11.1	We believe too often in science, and not enough in feelings and faith. Agree strongly – agree – neither agree nor disagree – disagree – disagree strongly	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)

Question		Identified in	Problem (code)	Severity ∅
Q11.2	Overall, modern science does more harm than good. Agree strongly – agree – neither agree nor disagree – disagree – disagree strongly	Experimental Condition		**5.7**
			Knowledge may not exist (5)	8.0
			Vague/unclear question (1)	4.7
			Undefined/vague term [modern 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. Completely agree – tend to agree – tend to disagree – completely disagree	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. Strongly agree – agree – neither agree nor disagree – disagree – strongly disagree	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) can be found in Appendix B (section 2.9). Bold figures are averaged question severity scores.

2.5.3 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*[22] 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.

[22] 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).

2.6 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.

2.7 References

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71(2)*, 287-311. doi:10.1093/poq/nfm006

Blair, J., & Conrad, F.G. (2011). Sample size for cognitive interview pretesting. *Public Opinion Quarterly, 75(4)*, 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(1)*, 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(4)*, 284–290.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20(1)*, 37-46.

Collins, D. (2003). Pretesting survey instruments: an overview of cognitive methods. *Quality of Life Research, 12(3)*, 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: Wiley. doi: 10.1002/0471654728.ch4

Conrad, F. G., & Blair, J. (2009). Sources of error in cognitive interviews. *Public Opinion Quarterly, 73(1)*, 32-55.

Conrad, F.G., Blair, J. & Tracy, E. (1999). Verbal reports are data! A theoretical approach to cognitive interviews. In *Proceedings of the Federal Committee on Statistical Methodology Research Conference (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: Wiley.

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: Wiley.

Fowler, F. J. (1992). How unclear terms affect survey data. *Public Opinion Quarterly 56(2)*, 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(5),* 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(1),* 3–22. doi:10.1093/poq/nfj012

Groves, R. M., Fowler Jr, F. J., 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(3),* 361-73. 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(6),* 743-764. doi:10.1177/0894439313517532

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? *Field Methods*, 26(*1),* 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. 51-75). 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(1),* 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(1),* 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 D.C.: 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(5),* 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: Wiley. doi:10.1002/0471654728.ch2

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

2.8 APPENDIX A. Classification scheme

Comprehension	**Retrieval**
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. Objectively wrong answer, question is misunderstood *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 *Reference period* 16. Reference periods are missing or undefined 17. Reference period carried over from earlier question	*Retrieval from memory* 18. High detail required or information unavailable 19. Long recall or reference period
Judgment	**Response Selection**
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 *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	

2.9 APPENDIX B. Questionnaire

Warm-up questions:

Q1. Wie beurteilen Sie ganz allgemein die wirtschaftliche Lage in Deutschland?
Sehr gut – Gut – Teils gut/ teils schlecht – Schlecht – Sehr schlecht

Q2. Und was glauben Sie, wie wird die allgemeine wirtschaftliche Lage in einem Jahr sein?
Wesentlich besser – Etwas besser – Unverändert – Etwas schlechter – Wesentlich schlechter

Experimental questions:

Q3. Alles in allem – wie zufrieden sind Sie mit den demokratischen Einrichtungen in unserem Land?
0 = Ganz und gar unzufrieden – 10 = Ganz und gar zufrieden

Q4. Wie zufrieden sind Sie mit der Krankenversicherung, der Arbeitslosen- und Rentenversicherung in der Bundesrepublik, also mit dem, was man das „Netz der sozialen Sicherung“ nennt?
0 = Ganz und gar unzufrieden – 10 = Ganz und gar zufrieden

Q5. Inwieweit stimmen Sie den folgenden Aussagen zu?
Bitte in jeder Zeile eine Antwort auswählen.

1. Privatwirtschaft ist das beste Mittel zur Lösung der wirtschaftlichen Probleme Deutschlands.
2. Es ist die Aufgabe des Staates, die Einkommensunterschiede zwischen den Leuten mit hohem Einkommen und solchen mit niedrigem Einkommen zu verringern.

   Stimme voll und ganz zu – Stimme zu – Weder noch – Stimme nicht zu – Stimme überhaupt nicht zu

Q6. Hier sind einige Meinungen über Staat und Wirtschaft in Deutschland. Inwieweit stimmen Sie den folgenden Meinungen zu oder nicht zu?
Bitte in jeder Zeile eine Antwort auswählen.

1. In unserer Gesellschaft muss jeder für sich schauen, dass er auf einen grünen Zweig kommt. Es hilft nicht viel, sich mit anderen zusammenzuschließen, um politisch oder gewerkschaftlich für seine Sache zu kämpfen.
2. Die Wirtschaft funktioniert nur, wenn die Unternehmer gute Gewinne machen. Und das kommt letzten Endes allen zugute.

3. Der Staat muss dafür sorgen, dass jeder Arbeit hat und die Preise stabil bleiben, auch wenn deswegen Freiheiten der Unternehmer eingeschränkt werden müssen.
4. Der Staat muss dafür sorgen, dass man auch bei Krankheit, Not, Arbeitslosigkeit und im Alter ein gutes Auskommen hat.
5. Wenn die Leistungen der sozialen Sicherung, wie Lohnfortzahlungen im Krankheitsfall, Arbeitslosenunterstützung und Frührenten, so hoch sind wie jetzt, führt dies nur dazu, dass die Leute nicht mehr arbeiten wollen.
6. Alles in allem gesehen, kann ich in einem Land wie Deutschland gut leben.
7. Die wirtschaftlichen Gewinne werden heute in Deutschland im Großen und Ganzen gerecht verteilt.
8. Selbst wenn man es wollte, könnte man die sozialen Ungleichheiten kaum geringer machen, als sie bei uns in Deutschland sind.

   Stimme voll und ganz zu – Stimme zu – Weder noch – Stimme nicht zu – Stimme überhaupt nicht zu

Q7. Stellen Sie sich vor, der Bundestag berät ein Gesetz, dass Sie für ungerecht oder schädlich halten. Was meinen Sie, wie wahrscheinlich ist es, dass Sie, allein oder mit anderen zusammen, versuchen würden, etwas dagegen zu unternehmen?

   Sehr wahrscheinlich – Einigermaßen wahrscheinlich – Nicht sehr wahrscheinlich – Überhaupt nicht wahrscheinlich

Q8. Wenn Sie politisch in einer Sache, die Ihnen wichtig ist, Einfluss nehmen, Ihren Standpunkt zur Geltung bringen wollten: Welche der folgenden Möglichkeiten würden Sie dann nutzen, was davon käme für Sie in Frage? *Bitte alles auf Sie Zutreffende auswählen.*

- Seine Meinung sagen, im Bekanntenkreis und am Arbeitsplatz
- Sich an Wahlen beteiligen
- Sich in Versammlungen an öffentlichen Diskussionen beteiligen
- Mitarbeit in einer Bürgerinitiative
- In einer Partei aktiv mitarbeiten
- Teilnahme an einer Demonstration
- Sich aus Protest nicht an Wahlen beteiligen
- Aus Protest einmal eine andere Partei wählen als die, der man nahesteht
- Beteiligung an einer Unterschriftensammlung
- Aus politischen, ethischen oder Umweltgründen Waren boykottieren oder kaufen

Q9. Hier ist eine Liste mit verschiedenen Auffassungen darüber, wie es in Deutschland mit den sozialen Unterschieden tatsächlich aussieht und wie es sein sollte. Inwieweit stimmen Sie den folgenden Aussagen zu oder nicht zu? *Bitte in jeder Zeile eine Antwort auswählen.*

1. Was man im Leben bekommt, hängt gar nicht so sehr von den eigenen Anstrengungen ab, sondern von der Wirtschaftslage, der Lage auf dem Arbeitsmarkt, den Tarifabschlüssen und den Sozialleistungen des Staates.
2. Das Einkommen sollte sich nicht allein nach der Leistung des Einzelnen richten. Vielmehr sollte jeder das haben, was er mit seiner Familie für ein anständiges Leben braucht.
3. Nur wenn die Unterschiede im Einkommen und im sozialen Ansehen groß genug sind, gibt es einen Anreiz für persönliche Leistungen.
4. Die Rangunterschiede zwischen den Menschen sind akzeptabel, weil sie im Wesentlichen ausdrücken, was man aus den Chancen, die man hatte, gemacht hat.
5. Ich finde die sozialen Unterschiede in unserem Land im Großen und Ganzen gerecht.

   Stimme voll zu – Stimme eher zu – Stimme eher nicht zu – Stimme überhaupt nicht zu

Q10. Inwieweit stimmen Sie den folgenden Aussagen zu oder nicht zu? *Bitte in jeder Zeile eine Antwort auswählen.*

1. Um in Deutschland heute ganz nach oben zu kommen, muss man korrupt sein.
2. In Deutschland haben nur Schüler der besten Gymnasien gute Chancen zu studieren.
3. In Deutschland können nur die Reichen ein Studium bezahlen.
4. In Deutschland haben alle Menschen die gleichen Chance zu studieren, unabhängig von Geschlecht, nationaler oder ethnischer Herkunft oder sozialer Schicht.

   Stimme voll zu – Stimme eher zu – Stimme eher nicht zu – Stimme überhaupt nicht zu

Q11. Inwieweit stimmen Sie den folgenden Aussagen zu oder nicht zu? *Bitte in jeder Zeile eine Antwort auswählen.*

1. Wir vertrauen zu sehr der Wissenschaft und nicht genug in unseren Gefühlen und dem Glauben.
2. Alles in allem schadet die moderne Wissenschaft mehr als sie nützt.

3. Die moderne Wissenschaft wird unsere Umweltprobleme bei nur geringer Veränderung unserer Lebensweise lösen.
4. Wir machen uns zu viele Sorgen über die Zukunft der Umwelt und zu wenig um Preise und Arbeitsplätze heutzutage.
5. Fast alle, was wir in unserer modernen Welt tun, schadet der Umwelt.
6. Die Leute machen sich zu viele Sorgen, dass der menschliche Fortschritt der Umwelt schadet.

   Stimme voll und ganz zu – Stimme zu – Weder noch – Stimme nicht zu – Stimme überhaupt nicht zu

Q12. Inwieweit fänden Sie es für sich persönlich akzeptabel, Abstriche von Ihrem Lebensstandard zu machen, um die Umwelt zu schützen?

Sehr akzeptabel – Eher akzeptabel – Weder akzeptabel noch inakzeptabel – Eher inakzeptabel – Sehr inakzeptabel

Q13. Glauben Sie, dass man eine Familie braucht, um wirklich glücklich zu sein, oder glauben Sie, man kann alleine glücklich leben?

Braucht Familie – alleine genauso glücklich – Alleine glücklicher – Unentschieden

Q14. Kinderreiche Familien sind selten geworden. Was denken Sie, ist das Image von Kinderreichen in unserer Gesellschaft. Bitte geben Sie dazu an, inwieweit die folgenden Aussagen zutreffen oder nicht zutreffen?
Bitte in jeder Zeile eine Antwort auswählen.

1. Kinder zu haben ist etwas Wundervolles, davon kann man nie genug haben.
2. Kinderreiche gelten als asozial.
3. Mit vielen Kindern leben ist wie in den guten alten Zeiten.

   Trifft voll und ganz zu – Trifft eher zu – Trifft eher nicht zu – Trifft überhaupt nicht zu

Q15. Wie oft waren Sie insgesamt in den letzten 12 Monaten über Nacht nicht zu Hause, weil Sie im Urlaub waren oder auf Besuch bei Freunden, Verwandten usw.?

Ich war nicht über Nacht fort – 1-5 Nächte – 6-10 Nachte – 11-20 Nächte – 21-30 Nächte – Mehr als 30 Nächte

Q16. Mit wie vielen Menschen haben Sie im Durchschnitt an einem normalen Wochentag Kontakt? Wir meinen Kontakte mit einzelnen Personen, also wenn sie mit jemandem reden oder diskutieren. Dies kann persönlich, telefonisch, brieflich oder über das Internet sein. Zählen Sie nur die Menschen, die Sie kennen, und denken Sie bitte auch an die, mit denen Sie zusammenwohnen.

0-4 Personen – 5-9 Personen – 10-19 Personen – 20-49 Personen – 50 oder mehr Personen

Q17. An wie vielen Tagen sehen Sie im Allgemeinen in einer Woche – also an den 7 Tagen von Montag bis Sonntag – fern?

An allen 7 Tagen in der Woche – An 6 Tagen in der Woche – An 5 Tagen in der Woche – An 4 Tagen in der Woche – An 3 Tagen in der Woche – An 2 Tagen in der Woche – An 1 Tag in der Woche – Seltener – Nie

Q18. An wie vielen Tagen sehen Sie im Allgemeinen in einer Woche Nachrichtensendungen von ARD oder ZDF?

An allen 7 Tagen in der Woche – An 6 Tagen in der Woche – An 5 Tagen in der Woche – An 4 Tagen in der Woche – An 3 Tagen in der Woche – An 2 Tagen in der Woche – An 1 Tag in der Woche – Seltener – Nie

Q19. An wie vielen Tagen sehen Sie im Allgemeinen in einer Woche Nachrichtensendungen von den privaten Fernsehsendern?

An allen 7 Tagen in der Woche – An 6 Tagen in der Woche – An 5 Tagen in der Woche – An 4 Tagen in der Woche – An 3 Tagen in der Woche – An 2 Tagen in der Woche – An 1 Tag in der Woche – Seltener – Nie

Q20. Wie oft nutzen Sie im Allgemeinen das Internet, um sich über Politik zu informieren?

Täglich – Mindestens einmal jede Woche – Mindestens einmal jeden Monat – Seltener – Nie

Q21. Wie oft würden andere Leute bei passender Gelegenheit versuchen, Sie zu übervorteilen oder aber versuchen, sich Ihnen gegenüber fair zu verhalten? Andere Leute würden …

fast immer versuchen, mich zu übervorteilen – meistens versuchen, mich zu übervorteilen – meistens versuchen, sich mir gegenüber fair zu verhalten – fast immer versuchen, sich mir gegenüber fair zu verhalten.

Q22. Ganz allgemein, was meinen Sie: Kann man Menschen vertrauen oder kann man im Umgang mit Menschen nicht vorsichtig genug sein? Man kann …

Menschen fast immer vertrauen – Menschen normalerweise vertrauen – normalerweise nicht vorsichtig genug sein im Umgang mit Menschen – fast nie vorsichtig genug sein im Umgang mit Menschen.

Q23. Inwieweit achten Sie auf gesundheitsbewusste Ernährung?

Sehr stark – Stark – Ein wenig – Gar nicht

Q24. Wie häufig trinken Sie die folgenden alkoholischen Getränke?
Bitte in jeder Zeile eine Antwort auswählen.

1. Bier
2. Wein, Sekt
3. Spirituosen (Schnaps, Weinbrand, etc.)
4. Mischgetränke (Alkopops, Cocktails, etc.)

Regelmäßig – Ab und zu – Selten – Nie

Q25. Es gibt unterschiedliche Meinungen zum Sport. Inwieweit stimmen Sie den folgenden Aussagen zu oder nicht zu?
Bitte in jeder Zeile eine Antwort auswählen.

1. Sport zu treiben fördert die Charakterentwicklung von Kindern.
2. Im Fernsehen kommt zu viel Sport.
3. Sport bringt unterschiedliche Gruppen in Deutschland einander näher, etwa Gruppen verschiedener nationaler oder ethnischer Herkunft.
4. Internationale Sportwettkämpfe erzeugen mehr Spannungen zwischen den Ländern als positive Gefühle.
5. In Deutschland sollte der Sport mehr durch öffentliche Mittel gefördert werden.

Stimme voll und ganz zu – Stimme zu – Weder noch – Stimme nicht zu – Stimme überhaupt nicht zu

3 A COMPARISON OF TWO COGNITIVE PRETESTING TECHNIQUES SUPPORTED BY EYE TRACKING[23]

3.1 Abstract

In questionnaire pretesting, supplementing cognitive interviewing with eye tracking is a promising new method that provides additional insights into respondents' cognitive processes while answering survey questions. When incorporating eye tracking into cognitive interviewing, two retrospective probing techniques seem to be particularly useful. In the first technique – retrospective probing – participants complete an online questionnaire, while cognitive interviewers monitor participants' eye movements in an adjacent room and note down any peculiarities in their reading patterns. Afterward, the interviewers ask targeted probing questions about these peculiarities in a subsequent cognitive interview. In the second technique – gaze video cued retrospective probing – respondents are additionally shown a video of their eye movements during the cognitive interview. This video stimulus is supposed to serve as a visual cue that may better enable respondents to remember their thoughts while answering the questions. We examine whether one of the two techniques is more effective when it comes to identifying problematic survey questions. In a lab experiment, participants' eye movements (n = 42) were tracked while they completed six questions of an online questionnaire. Simultaneously, their reading patterns were monitored by an interviewer for evidence of response problems. After completion of the online survey, a cognitive interview was conducted. In the retrospective probing condition, probing questions were asked if peculiar reading patterns were observed during the eye-tracking session (e.g., re-readings of specific words or text passages). In the other condition, participants were shown a video of their recorded eye movements, in addition to receiving probing

[23] A version of this chapter has been published as:
Neuert, C.E. & Lenzner, T. (2015).A comparison of two cognitive pretesting techniques supported by eye tracking. *Social Science Computer Review* online first.
Parts of this chapter were presented at the VI European Congress of Methodology, July 23-25, 2014, Utrecht, Netherlands, and at the 16th General Online Research Conference (GOR15), March 18-20, 2015, Cologne, Germany.

questions about the questions displayed. Results show that both techniques did not differ in terms of the total number of problems identified. However, gaze video cued retrospective probing identified fewer unique problems and fewer types of problems than pure retrospective probing.

3.2 Introduction

The general goal of cognitive interviewing is to obtain information about the cognitive processes underlying survey responding and to identify difficulties respondents have in answering them. By identifying problematic questions and providing information about how a question could be revised, cognitive interviewing contributes to a better understanding of questions by respondents and thus decreases measurement error (Forsyth & Lessler, 1991; Willis, 2005). 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).

In questionnaire pretesting, supplementing cognitive interviewing with eye tracking is a novel and promising approach that might provide additional insights into respondents' cognitive processes while answering survey questions (Galesic & Yan, 2011). Whereas cognitive interviews initially took place in pretesting laboratories equipped with video and audio recording equipment, these labs are, today, often additionally equipped with eye-tracking technology (Campanelli, 2008); for instance, 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 based on the idea of a direct relationship between eye movements and cognitive processing. The so-called eye-mind hypothesis of Just and Carpenter (1980) assumes a link between what people are looking at and what they are thinking. It postulates that words or objects are fixated as long as they are being processed (Just & Carpenter, 1980). According to this assumption, eye tracking appears to be a natural supplement to cognitive interviewing, because cognitive interviewing is about obtaining information about peoples' thoughts while answering

a questionnaire (Willis, 2005). Observing the eye movements – where and for how long respondents look when reading and answering questions – helps to reach a better understanding of the participant's answer process and can be used to detect difficulties that may have arisen while answering (Neuert & Lenzner, 2015). Because eye tracking allows the detection of conscious and unconscious reactions to survey questions (Tries, Nebel & Blanke, 2012), it might also point to difficulties that are not consciously apparent to participants and have a small chance of being detected (Blair & Conrad, 2011). As we have demonstrated in a previous study, incorporating eye tracking into cognitive interviewing is indeed more productive in identifying questionnaire problems than using cognitive interviewing alone (Neuert & Lenzner, 2015).

In the present article, we are interested in how eye tracking can be implemented most effectively into cognitive survey pretesting studies. We compare two eye tracking supported cognitive pretesting techniques: Retrospective probing based on observed eye movements and retrospective probing, which incorporates a gaze video cue, that is, a video that shows the participants' eye movements while they filled in an online questionnaire.

3.3 Background

The term "cognitive interviewing" usually refers to administering draft questions of a survey instrument to respondents who provide additional verbal material about their responses and their thoughts (Beatty & Willis, 2007). Cognitive interviewing aims to understand and to obtain information on respondents' thought processes while answering these questions (i.e., how respondents understand the questions, as well as how they arrive at an answer) and to identify specific difficulties respondents have with the questionnaire (Beatty, 2004; Beatty & Willis, 2007). The verbal material about respondents' thought processes that is gathered in the cognitive interviews is used to evaluate the quality of the questions and to provide information about whether a question needs revision (Beatty & Willis, 2007).

One of the most common techniques used in cognitive interviews is “verbal probing”. Probes are follow-up questions about what respondents were thinking and how they interpreted the questions or specific terms used in the questionnaire (Willis, 2005). During cognitive interviews, participants typically first answer the survey questions and then respond to a series of probing questions (Willis, 2005; Willis & Miller, 2011). Follow-up probing can occur either immediately after the subject has answered the target survey question (concurrent probing) or at the end of the interview, during a debriefing session (retrospective probing; Willis, 2005). In current practice, concurrent probing is used more frequently, although, under certain circumstances, retrospective probing may be the more efficient technique, for example, when testing self-administered questionnaires, in which the respondent should not be disturbed, to determine whether he or she can handle the instrument alone (Willis, 2005).

When conducting cognitive interviews in combination with eye tracking, it is sensible to probe only retrospectively. In eye-tracking supported cognitive pretesting studies, respondents are seated in front of an eye tracker in the laboratory and are instructed to fill in a questionnaire at their usual pace. Simultaneously, a cognitive interviewer monitors the respondents’ actions and eye movements, in real time, on a computer screen in an adjacent room and notes any peculiarities in their reading patterns (e.g., long or repeated fixations or multiple regressions from answers to question text). These are then addressed in a cognitive interview that is conducted after respondents have completed the survey. If eye tracking were to be used with concurrent probing, participants might produce eye movements that they would not normally make when they complete an online questionnaire on their own (Pernice & Nielsen, 2009). For example, unusual eye movements might be caused by participants looking away from the screen when describing something to the interviewer or by fixating on certain areas of the screen while describing their thought processes regarding that question. Unusual eye movements would be especially disadvantageous if the data were also evaluated quantitatively after the interview. Concurrent probing might also make participants more aware of the fact that their eye movements are being tracked. Therefore, when conducting cognitive

interviews in combination with eye tracking, it is reasonable to apply retrospective rather than concurrent probing.

In general, retrospective probing has the advantage that it does not interrupt the flow of answering an entire questionnaire and, thus, creates a more realistic field setting. However, retrospective probing also has some drawbacks, because participants may have forgotten key information or the information about their problems may no longer be accessible when they are finally asked to answer the probing questions (Willis, 2005). A potential solution to aid the participants' memory could be the use of a gaze video cue, a technique that has already been employed in usability research in combination with thinking-aloud (e.g., Ball, Eger, Stevens, & Dodd, 2006; Elling, Lentz, & DeJong, 2011; Hansen, 1991; Hyrskykari et al., 2008) as well as in field research with mobile eye tracking (Eghbar-Azar & Widlok, 2013). When using retrospective probing in conjunction with a gaze video, participants are presented with a replay of their eye movements during the cognitive interview. In the video replay, the eye movements appear as red dots that represent where participants were looking when answering the questions. The longer a participant looks at something, the larger the red dot becomes. Thus, it is possible for the participant to see how he or she read and answered the question. This video stimulus is supposed to serve as a visual cue that may better enable respondents to remember their thoughts while answering the questions by reviewing their eye movements.

On the negative side, showing participants a gaze video replay may increase the risk of false alarms, that is, identifying a problem that is not actually present (Conrad & Blair, 2009). When confronted with their own eye movements, participants might come up with a post hoc explanation for their behavior to meet what they think is expected of them, instead of just reporting their thinking.

In this study, we compare gaze video cued retrospective probing with retrospective probing without any cues within the framework of identifying problematic survey questions. Three research questions will be addressed:

Research question 1: Do both techniques differ in terms of the number of problems identified?

Research question 2: Do both techniques differ in the types of problems identified?

Research question 3: Do both techniques differ in the way they stimulate participants when commenting on their behavior?

3.4 Methods

3.4.1 Design

To answer our research questions, we used a randomized between-subject design with two conditions (gaze-replay video yes/no). All participants (n = 42) were seated in front of the eye tracker and, after a short explanation of the eye tracker and a standard calibration procedure, the participants completed the online questionnaire while their eye movements were recorded and their response behavior was monitored by a cognitive interviewer sitting in a different room. The interviewer used a coding scheme (described in section 3.4.5) to document any peculiar reading pattern that was observed. Following completion of the online survey, a cognitive interview was conducted. Each cognitive interview was videotaped. During the cognitive interview, participants in the retrospective probing condition (n = 21) received a paper version of the questionnaire with screenshots of the questions, to remind them of their initial thoughts, whereas participants in the gaze video cued retrospective probing condition (n = 21) were shown a video of their recorded eye movements while filling in the online questionnaire. In addition, respondents in both conditions were asked a set of probing questions about the questions under scrutiny.

3.4.2 Participants

This experiment was part of a larger study conducted in October and November 2012 in the pretest laboratory at GESIS – Leibniz Institute for the Social Sciences in Mannheim, Germany (see section 3.4.6 for detailed information). For this experiment, 33 participants were recruited from the respondent pool maintained by the institute, as well as by word of mouth. For their participation in the whole study, which took about one and a half hours, participants received a compensation of €30.

Additionally, nine colleagues and student assistants working primarily in non-scientific departments of the institute participated in the study for free, so that a total of 42 subjects participated in the experiment. Participants came separately to the pretest laboratory at GESIS for individual sessions. Table 3.1 shows some demographic characteristics of the participants.

Table 3.1. Demographic characteristics of participants (%).

Gender		Age		Years of schooling		Computer Usage	
Female	52%	18-34	60%	9 years or less	19%	(Almost) Daily	91%
Male	48%	35-54	33%	10 years	10%	Weekly	2%
		55+	7%	12 years or more	71%	Seldom or never	7%

3.4.3 The questionnaire

The questionnaire included 6 closed-ended items that were adapted from the International Social Survey Programme (ISSP 2003, 2004) and the European Social Survey (ESS, round 1, 2002; round 5, 2010). The language of the questionnaire was German. The official English translations of the questions provided by the survey organizers are available in Appendix A. The questions included two question formats: four single-choice questions and one grid question with 2 items. One of the questions asked about respondents' behavior, the other five about respondents' attitudes. 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.

3.4.4 Eye-tracking equipment

We used a Tobii T120 eye-tracking system together with the Tobii Studio 3.2.1 software to record the participants' eye movements. The Tobii T120 is a remote eye tracker embedded in a 17" TFT monitor (resolution 1280 x 1024) with two binocular infrared cameras placed underneath the computer screen. This system is particularly suitable when stimuli can be presented on a screen and provides unobtrusive

recording of respondents' eye movements and permits head movements within a scale of 30 x 22 x 30 cm. Eye movements were recorded at a sampling rate of 120 Hz, meaning that 120 gaze data points per second were collected for each eye. The Tobii Studio software allows the interviewer to play back a video recording of the original recording, with or without eye movements; in our case, a video of the respondents' eye movements recorded during completion of the online questionnaire. The software also includes an automatic retrospective think-aloud recording function that allows the interviewer to video and audio record the participants' comments and reactions while showing a playback from the previously recorded task. Finally, the software includes features that enable the interviewer to adjust playback speed, start or pause playing, rewind or fast forward the video. This allows the interviewer to control the recording, for example, to pause if the participant needs more time to respond, or to repeat a video sequence.

3.4.5 Interview protocol and interviewer instructions

The interview protocol included prescripted, general probing questions for all 6 items, such as "*Could you please explain your answer a little further?*", "*What were you thinking when answering the question?*", "*How easy or difficult was it for you to come up with your answer?*", and "*Why did you find it (rather/very) difficult?*". The use of prescripted probing questions ensured a relatively standardized application of the protocol between the different interviewers. The use of general probing (in contrast to specific probing) questions has the advantage that they do not influence the answer process of the respondent. Furthermore, general probes induce the participant to elaborate in a narrative way, which helps to collect information on how and why respondents answered the question as they did (Willson & Miller, 2014).

The interviewers were instructed to probe only those questions for which peculiar reading patterns were observed during the eye-tracking session. To document if a peculiarity occurred, interviewers were provided with a coding scheme for peculiar reading patterns: They had to check a box if they observed one of the following five behaviors: (1) long or repeated fixations on a word, (2) rereadings of specific words or text passages, (3) regressions from answers to question text, (4)

correction of the chosen response category, and (5) skipping a question. In addition, it was possible to check a box if an "other", not specified peculiarity occurred and to describe the corresponding behavior. If one or more of the behaviors described previously 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 particular behavior.

Participants in the gaze video cued retrospective probing condition were given the following instruction: "*I am now going to show you a recording of your eye movements during/while answering question x. The red dots that you are going to see in the replay show how you read and answered the question and represent where you were looking. The longer you were looking at something, the larger the red dot becomes. After you have watched the replay, I would like you to tell me how you came up with your answer and what you were thinking when answering the question.*"

3.4.6 Procedure

The experiment reported in this article was part of a larger study with several unrelated experiments. The entire study took about one and a half hours and consisted of three parts. In the first part, participants completed an online questionnaire while their eye movements were tracked. The entire questionnaire included 58 questions. In the second part, a cognitive interview was conducted (cf. Neuert & Lenzner, 2015). In the third part, participants completed another online questionnaire that consisted of different small experiments unrelated to this study (cf. Lenzner, Kaczmirek, & Galesic, 2014). The experiment reported in this paper refers to the last six questions of the online questionnaire (part one of the study), which were discussed at the end of the subsequent cognitive interview (part two of the study).

The interviews in both conditions were conducted by five interviewers (three researchers and two student assistants) who had all previously conducted cognitive interviews. Individual interviewers each conducted between three to five interviews

in each condition. The average survey completion time for the six questions was approximately 2.5 minutes (154 seconds). In terms of time required for conducting the cognitive interviews in both conditions, we found that administering retrospective probing in conjunction with a gaze video cue required close to 373 seconds, whereas the pure retrospective probing interviews took approximately 331 seconds.

3.5 Results

In the analysis described subsequently, we compared gaze video cued retrospective probing and retrospective probing both quantitatively, that is, in terms of the total number of problems identified (including recurrences of the same problem) and the number of unique problems identified, and qualitatively, that is, in terms of the types of problems identified and the types of comments given by respondents. First, we examined the total number of problems identified in each condition. Subsequently, we categorized the types of problems and examined the number of unique problems. Finally, we categorized the types of comments given by respondents.

3.5.1 Number of problems

To identify problems, the first author reviewed all videotapes of the cognitive interviews and gave each questionnaire item, for each interview, a dichotomous score that reflected whether a problem was identified in the question (1) or not (0). Those sections of the cognitive interviews that contained a context relevant for understanding potential problems were transcribed. Afterward, a student assistant reviewed and coded all interviews, to estimate interrater reliability. Agreement between these two raters was 93% and Cohen's Kappa (1960) was found to be .84, which is "almost perfect", according to Landis and Koch's (1977, p.165) criteria. The number of problems that resulted from this analysis contained all detected problems for all participants, which means that problems can occur repeatedly for specific questions, because several participants might have encountered the same problem.

Table 3.2 shows the total number of problems identified in each condition and the distribution of these problems per question. A comparison of the total number of problems across conditions revealed that the combination of a gaze video with retrospective probing did not identify significantly more problems (n = 44) than retrospective probing (n = 41; χ^2 = 1.38, df = 1, p = .160). In both conditions, most problems were identified in Question 5 (23 problems) and in Question 1.2 (19 problems), whereas only one participant in each condition experienced a problem when answering Question 1.1.

Table 3.2. Number of problems identified, by condition.

	Total number of problems	Number of problems in					
		Q1.1	Q1.2	Q2	Q3	Q4	Q5
Retrospective probing (no cue)	41	1 (2%)	10 (24%)	7 (17%)	8 (20%)	5 (12%)	10 (24%)
Gaze video cued retrospective probing (video cue)	44	1 (2%)	9 (21%)	5 (11%)	9 (21%)	7 (16%)	13 (30%)

3.5.2 Types of problems

In our next analysis step, we evaluated whether both techniques identified different types of problems. For each item that was perceived as problematic, we reviewed the transcripts of the interviews and coded them into problem types, using a problem classification scheme adopted from various existing schemes (DeMaio & Landreth, 2004; Lessler & Forsyth, 1996; Presser & Blair, 1994; Rothgeb, Willis, & Forsyth, 2001).

The problem classification scheme included a total of 30 problem codes that were grouped according to the four stages of the survey response process (comprehension, retrieval, judgment, response selection; Tourangeau 1984; Tourangeau, Rips, & Rasinski, 2000; see Appendix B, section 3.9). Individual items could be assigned to multiple problem codes. Problem types were also coded by a student assistant, resulting in an agreement of 79% and a Kappa of .74 (classified as "substantial" reliability by Landis & Koch, 1977). The types of problems discovered

in the questions came from three of the four stages of the survey response process: comprehension difficulties, judgmental issues, and response selection. Problems with information retrieval were not detected (see Table 3.3).

Table 3.3. Types of problems identified, by condition.

	Total number of problems	Types of problems			
		Compre-hension	Retrieval	Judgment	Response Selection
Retrospective probing (no cue)	41	36 (88%)	0 (0%)	3 (7%)	2 (5%)
Gaze video cued retrospective probing (video cue)	44	40 (91%)	0 (0%)	4 (9%)	0 (0%)

In both conditions, the highest proportion of problems was classified as comprehension problems. Two types of problems from the "response selection" category were detected in the retrospective probing condition, but problems with response selection were not found in the gaze video cue retrospective probing condition. Again, no statistically significant difference between the two conditions, with regard to the types of problems identified, was found ($\chi^2 = 2.25$, df = 2, $p = .325$).

Besides the general productivity of each technique, it is important to establish how many unique problems each technique identified. We therefore also looked at the number of unique problems detected in each condition (Table 3.4). We classified a problem as unique if it occurred at least once per question (irrespective of how many participants had experienced the same problem). When comparing the total number of unique problems across conditions, we found that gaze video cued retrospective probing identified significantly less unique problems ($n = 14$) than retrospective probing ($n = 20$; $\chi^2 = 5.56$, df = 1, $p = .037$).

Although, in Question 1.1, one problem in the retrospective probing condition and one in the gaze video cued retrospective probing condition were detected, retrospective probing identified one (Questions 2-5) or even two unique problems (Question 1.2) more than gaze video cued retrospective probing had detected in all other questions. Whereas, in Question 1.2, the problem that the term "civil

disobedience" was unknown to some respondents (Code 4, see Table 3.5) was identified in both conditions, two other problems were identified exclusively in the retrospective probing condition. In this condition, the question was also found to be vague and unclear (Code 1) and to have a complex syntactical structure (Code 11). Altogether, three unique problems were detected in Question 2. Even though two problem types, namely that the question was vague and unclear (Code 1) and that it contained a complex topic (Code 2), were identified in both conditions, the more specific problem – the question contained undefined terms (United Nations; intervene) – was only detected in the retrospective probing condition. A summary of the number and types of problems identified per question and condition is presented in Table 3.5.

Table 3.4. Number of unique problems identified, by condition.

	Total number of problems	Number of problems in					
		Q1.1	Q1.2	Q2	Q3	Q4	Q5
Retrospective probing (no cue)	20	1 (5%)	3 (15%)	3 (15%)	4 (20%)	5 (25%)	4 (20%)
Gaze video cued retrospective probing (video cue)	14	1 (7%)	1 (7%)	2 (14%)	3 (21%)	4 (29%)	3 (21%)

In both conditions, the highest proportion of unique problems was classified as "vague or unclear question" (25% retrospective probing and 29% gaze video cued retrospective probing), or as containing "undefined or vague terms" (20% retrospective probing and 21% gaze video cued retrospective probing). Four types of unique problems were detected exclusively in the pure retrospective probing condition: Only respondents in this condition referred to the error codes "knowledge may not exist" (Question 4), "erroneous or inappropriate assumption" (Question 3), "response categories missing" (Question 5), and "no formally adequate answer" (Question 4).

Table 3.5. Number and types of unique problems identified, by condition.

Questions		Number of unique problems	Types of problems (Code)	Frequency
Q1.1	no video	1	Undefined/vague term [opportunities to participate in public decision-making] (4)	1
	video cued	1	Undefined/vague term [opportunities to participate in public decision-making] (4)	1
Q1.2	no video	3	Undefined/vague term [civil disobedience] (4)	8
			Vague/unclear question (1)	1
			Complex or awkward syntax (11)	1
	video-cued	1	Undefined/vague term [civil disobedience] (4)	9
Q2	no video	3	Vague/unclear question (1)	1
			Complex topic (2)	2
			Undefined/vague term [United Nations; intervene] (4)	4
	video cued	2	Vague/unclear question (1)	1
			Complex topic (2)	4
Q3	no video	4	Vague/unclear question (1)	2
			Complex or awkward syntax (11)	3
			Potentially sensitive or desirability bias (21)	1
			Erroneous/inappropriate assumption (12)	2
	video cued	3	Vague/unclear question (1)	2
			Complex or awkward syntax (11)	6
			Potentially sensitive or desirability bias (21)	1
Q4	no video	5	Vague/unclear question (1)	1
			Complex topic (2)	1
			Undefined/vague term [direct influence] (4)	1
			Knowledge may not exist (5)	1
			No formally adequate answer (28)	1
	video cued	4	Vague/unclear question (1)	2
			Complex topic (2)	1
			Undefined/vague term [direct influence] (4)	3
			Boundary lines (6)	1
Q5	no video	4	Vague/unclear question (1)	1
			Boundary lines (6)	6
			Complex estimation (20)	2
			Response categories missing (27)	1
	video cued	3	Vague/unclear question (1)	4
			Boundary lines (6)	6
			Complex estimation (20)	3

3.5.3 Classification of problems

To examine whether the different cues stimulate the participants in different ways when commenting on their behavior, we classified participants' comments into three categories, according to the coding scheme of verbalizations suggested by Hansen (1991), which was slightly altered for our purposes (see Table 3.6). Instead of speaking of "manipulative operations" that describe an action in a usability test (Hansen, 1991), we used the term "behavioral" to code comments that express exclusively an action, for example "I have read the question and answered it". "Cognitive" comments are defined as interpretations, assessments, and expectations of the respondents (e.g., "I have never heard the term [x] before."). Our third category is a combination of both, where "cognitive and behavioral" comments are associated with each other, for example "I wasn't sure about the term [x] and that is why I read the question several times." For the classification of comments, we coded all those sections of the cognitive interviews that contained a relevant context for understanding whether a problem existed or not. A total of 95 comments (48 in the retrospective probing condition and 47 in the gaze video cued retrospective probing condition, see Table 3.6) were coded by the first author and a student assistant, respectively. Interrater reliability between both coders was found to be Kappa = .78, which is generally classified as "substantial" reliability (Landis & Koch, 1977, p.165) and agreement was found to be 87%. Only one code was assigned to each comment. The results are shown in Table 3.6.

Table 3.6. Class of comments, by condition.

	Total number of comments	Types of comments		
		Behavioral	Cognitive	Behavioral - cognitive
Retrospective probing (no cue)	48	2 (4%)	31 (65%)	15 (31%)
Gaze video cued retrospective probing (video cue)	47	5 (11%)	25 (53%)	17 (36%)
Total	95	7 (7%)	56 (59%)	32 (34%)

With respect to the types of comments, gaze video cued retrospective probing stimulated the participants to produce slightly more "behavioral" comments (11% vs. 4%) and to produce less "cognitive" comments than when no cue was used (53% vs. 65%), meaning that participants were commenting more on what they were doing and less on what they were thinking when answering questions.

The gaze video cued retrospective probing condition also stimulated the participants to produce slightly more "behavioral and cognitive" comments (36% vs. 31%) in which the participants linked their behavior with what they were thinking at the time. Overall, the highest proportion of comments was classified as "cognitive" in both conditions.

In order to evaluate how well the technique of gaze video cued probing worked, we took brief notes after reviewing each cognitive interview in the gaze video cued probing condition and categorized participants into three groups: technique worked well, moderately well, or not at all. For almost half of the participants (n = 9), seeing a replay of their own eye movements worked well and they were able to associate what they were seeing with what they had been thinking. For a further eight participants, the technique worked moderately well. However, in this group, after a period of adaptation, the technique worked increasingly better towards the end of the interview. The remaining four participants had problems with the task and were either simply looking at their eye movements or were describing what they were seeing, but not referring to the question.

3.6 Discussion and conclusion

The goal of this experiment was to compare retrospective probing, in conjunction with a gaze video replay, with retrospective probing without any cue when testing survey questions in pretesting studies supported by eye tracking. Results show that the combination of retrospective probing with a gaze video cue and the pure retrospective probing did not differ significantly in terms of their quantitative output (i.e., total number of problems identified). However, gaze video cued retrospective probing identified significantly fewer unique problems and fewer types of problems.

Hence, we do not find evidence that eye movement replay serves as an extra cue that enables participants to better remember what they were thinking when answering the questions. However, due to the relatively small sample size of this study, our conclusions have to be considered with caution and we encourage further methodological investigations to confirm or reject our results.

A potential explanation for why the gaze video cue did not produce better results than pure retrospective probing might be that the eye movements not only supported participants in remembering their initial thoughts, but also distracted them. For most participants, seeing their own eye movements was a new experience. Although we explained to them what they would see, we observed that it was often difficult for participants to interpret the replay of their eye movements. The categorization of the comments made by the participants revealed that gaze video cued retrospective probing stimulated the participants to produce slightly more "behavioral" comments and to produce fewer "cognitive" comments than when no cue was used. Seeing a replay of their own eye movements might have stimulated the participants simply to describe what they were doing instead of what they were thinking while answering the questions. In line with this argument, by exclusively describing what they were seeing, the participants might not have provided the interviewers with enough information to diagnose whether a problem existed and, if so, what caused the problem. In addition, we were concerned that the gaze video cue might increase the risk of false alarms, because participants could be tempted to provide post hoc explanations for their viewing behavior. However, our findings do not indicate that showing a gaze replay increased the risk of false alarms. Even though gaze video cued retrospective probing identified slightly more problems than pure retrospective probing, both techniques did not differ in the types of identified problems and retrospective probing identified even more unique problems than video cued retrospective probing.

Our results are limited by a number of factors that encourage additional studies. First, the cognitive interviewing protocol was prescripted and relatively structured, so that interviewers were not encouraged to probe spontaneously. Furthermore, we asked exclusively general probing questions and did not use

specific probes (specially designed to address response processes within the four-stage cognitive model). In cognitive interviews, interviewers typically probe participants' responses in a more flexible manner and it might be worth examining whether more specific questions that are based on the observed eye movements have a positive effect on respondents remembering what they thought while seeing their eye movements. Maybe we would have identified more, or other, problems if interviewers had been given more flexibility, which is a general strength of cognitive interviewing as a pretesting method. Additionally, the experiment reported in this article was conducted only for the last six questions of a longer questionnaire and participants answered probing questions for the other questions without seeing a video of their eye movements in a previous part of the cognitive interview. By the time, the gaze video recording was shown, some respondents might have got used to to the previously applied probing style and seeing the video recording of their eye movements in addition might have caused confusion. Furthermore, the benefit of the eye movement replay might have been stronger if participants had been given more time to habituate to the recording. Hence, it may be worth investigating whether training respondents in interpreting their eye movements for a few minutes before starting the actual interview and using the gaze video cue earlier in the cognitive interview could render the technique more useful.

Another limitation of our study is that we used relatively short survey questions. It is possible that the technique is not, or less, suitable for short survey questions or short texts in general. The added value of showing participants a video of their eye movements might be greater when websites or more complex question designs, such as those used in business surveys, are tested; these require an enhanced interaction with an online questionnaire or website (e.g., questions with lookup databases, question navigation with tabs). We encourage future research on questions in which more complex designs are used. For those questions, it might also be worth to compare whether seeing a replay of the answer process without the gaze overlay might decrease participants confusion which could thus be more effective than seeing a video replay with a gaze overlay when identifying question problems. A final limitation is that no concurrent techniques such as thinking aloud or concurrent

probing techniques were used in this experiment. Future research could investigate whether combining the gaze video cue with thinking aloud or concurrent probing might be more appropriate than combining it with retrospective verbal probing.

With regard to the practical implications of this study, our findings suggest that using a gaze video replay in combination with retrospective probing is not worth the effort when pretesting short survey questions, because gaze video cued retrospective probing identified significantly less unique problems and less types of problems than pure retrospective probing. Moreover, the application of a gaze video replay is more time consuming than simple verbal probing and some participants clearly had difficulties in interpreting their own eye movements, which might have distracted them from reporting problems they had actually experienced when answering the questions. We therefore do not recommend the use of gaze video cued retrospective probing in eye tracking supported pretesting studies unless there is a special interest in usability and questionnaire navigation that should be discussed with participants.

3.7 References

Ball, L., Eger, N., Stevens, R., & Dodd, J. (2006). Applying the post-experience eye-tracked protocol (PEEP) method in usability testing. *Interfaces, 67,* 15-19.

Beatty, P. C. (2004). The dynamics of cognitive interviewing. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer, *Methods for testing and evaluating survey questionnaires* (pp. 45-66). Hoboken, NJ: Wiley.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71(2),* 287-311. doi:10.1093/poq/nfm006

Blair, J., & Conrad, F.G. (2011). Sample size for cognitive interview pretesting. *Public Opinion Quarterly, 75(4),* 636-658. doi: 10.1093/poq/nfr035

Campanelli, P. (2008). Testing survey questions. In E. D. De Leeuw, J. J. Hox, D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 176-200). New York/London: Erlbaum/Taylor & Francis.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20(1),* 37-46.

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

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). Hoboken, NJ: Wiley. doi: 10.1002/0471654728.ch5

Eghbal-Azar, K., & Widlok, T. (2013). Potentials and limitations of mobile eye tracking in visitor studies: Evidence from field research at two museum exhibitions in Germany. *Social Science Computer Review, 31(1),* 103-118. doi: 10.1177/0894439312453565

Elling, S., Lentz, L., & De Jong, M. (2011). *Retrospective think-aloud method: Using eye movements as an extra cue for participants' verbalizations*. Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems (pp. 1161-1170). New York, NY: ACM Press.

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: John Wiley.

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.

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

Hansen, J. P. (1991). The use of eye mark recordings to support verbal retrospection in software testing. *Acta Psychologica 76*, 31-49.

Hyrskykari, A., Ovaska, S., Majaranta, P., Räihä, K.-J., & Lehtinen, M. (2008). Gaze path stimulation in retrospective think aloud. *Journal of Eye Movement Research 2*, 1-18.

ISSP (2003). International Social Survey Programme 2003: National Identity II (ISSP 2003). GESIS Data Archive, Cologne, Germany, ZA3910. Source Questionnaire.

ISSP (2004). International Social Survey Programme 2004: Citizenship (ISSP 2004). GESIS Data Archive, Cologne, Germany, ZA3950. Source Questionnaire.

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. (2014). Left feels right: a usability study on the position of answer boxes in web surveys. *Social Science Computer Review, 32(6),* 743-764. doi: 10.1177/0894439313517532

Lessler, J. T., & Forsyth, B. H. (1996). A coding system for appraising questionnaires. In N. Schwarz, & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 259-291). San Francisco, CA: Jossey-Bass.

Neuert, C., & Lenzner, T. (2015). Incorporating eye tracking into cognitive interviewing to pretest survey questions. *International Journal of Social Research Methodology* (online first).

Pernice, K. & Nielsen, J. (2009). *How to conduct eyetracking studies*. Fremont, CA: Nielsen Norman Group. Retrieved May 11, 2014, http://www.nngroup.com/reports/how-to-conduct-eyetracking-studies/

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results. *Sociological methodology*, *24(1)*, 73-104.

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: National Academy Press.

Rothgeb, J., Willis, G., & Forsyth, B. (2001, May). *Questionnaire pretesting methods: Do different techniques and different organizations produce similar results?* Paper presented at the annual meeting of the American Association for Public Opinion Research, Montreal. Retrieved June 28, 2012, from https://www.census.gov/srd/papers/pdf/rsm2005-02.pdf

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, Germany: Federal Statistical Office.

Tries, S., Nebel, S. & Blanke, K. (2012). *How to provide high data quality in online-questionnaires: Setting guidelines in design*. Paper presented at the European Conference on Quality in Official Statistics, Athens, Greece, May 29 –June 1, 2012. Retrieved November 19, 2014, from

http://www.q2012.gr/articlefiles/sessions/34.1_Tries_On%20line%20questionnaires%20setting%20guidelines%20in%20design.pdf

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Willis, G. B. & Miller, K. (2011). Cross-cultural cognitive interviewing: Seeking comparability and enhancing understanding. *Field Methods 23(4),* 331-341.

Willson, S., & Miller, K. (2014). Data collection. In K. Miller, S. Willson, V. Chepp, & J. L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 15-33). Hoboken, NJ: John Wiley.

3.8 APPENDIX A. Questions

Question 1

Wie wichtig oder unwichtig sind für Sie folgende Rechte in einer Demokratie? *Bitte in jeder Zeile eine Antwort auswählen.*

Q1.1 Dass man den Menschen Möglichkeiten gibt, an politischen Entscheidungen teilzuhaben.

Q1.2 Dass Bürger die Möglichkeit des zivilen Ungehorsams gegenüber Regierungsentscheidungen haben.

Antwortoptionen:

Überhaupt nicht wichtig 1 – 2 – 3 – 4 – 5 – 6 – 7 Sehr wichtig

English translation:

There are different opinions about people's rights in a democracy. On a scale of 1 to 7, where 1 is not at all important and 7 is very important, how important is it: *Please tick one box on each line.*

Q1.1 That people be given more opportunities to participate in public decision-making.

Q1.2 That citizens may engage in acts of civil disobedience when they oppose government actions.

Answer options:

Not at all important 1 – 2 – 3 – 4 – 5 – 6 – 7 Very important

Question 2

Welche dieser zwei Aussagen kommt Ihrer Ansicht am nächsten?

Wenn ein Land die Menschenrechte ernsthaft verletzt, sollten die Vereinten Nationen eingreifen.

Selbst wenn die Menschenrechte ernsthaft verletzt werden, muss die Souveränität eines Landes respektiert werden, und die Vereinten Nationen sollten nicht eingreifen.

Weiß nicht, was die Vereinten Nationen sind

English translation:

Which of these two statements comes closer to your view?

If a country seriously violates human rights, the United Nations should intervene.

Even if human rights are seriously violated the country's sovereignty must be respected, and the United Nations should not intervene.

Don't know what the United Nations is.

Question 3

Inwieweit stimmen Sie der folgenden Aussage zu oder nicht zu?

Ich bin oft weniger stolz auf Deutschland, als ich es gerne wäre.

Antwortoptionen:

Stimme voll und ganz zu - Stimme zu - Weder noch - Stimme nicht zu - Stimme überhaupt nicht zu

English translation:

How much do you agree or disagree with the following statements?

I am often less proud of Germany than I would like to be.

Answer options:

Agree strongly - Agree - Neither agree nor disagree - Disagree - Disagree strongly

Question 4

Was würden Sie sagen: In welchem Ausmaß ermöglicht das politische System Deutschlands Menschen wie Ihnen direkten Einfluss auf die Politik auszuüben?

Antwortoptionen:

Überhaupt nicht - In sehr geringen Ausmaß – Ein wenig – In hohem Ausmaß – In sehr hohem Ausmaß

English translation:

And how much would you say that the political system in Germany allows people like you to have a direct influence on politics?

Answer options:

Not at all - Very little - Some - A lot - A great deal

Question 5

Abgesehen davon, was Sie für Ihre Familie, an Ihrem Arbeitsplatz oder in Vereinen, Verbänden oder Organisationen tun, wie oft helfen Sie anderen Menschen – wenn überhaupt?

Antwortoptionen:

Täglich – Mehrmals in der Woche - Einmal in der Woche - Mehrmals im Monat – Einmal im Monat - Seltener - Nie

English translation:

Not counting anything you do for your family, in your work, or within voluntary organisations, how often, if at all, do you actively provide help for other people?

Answer options:

Every day - Several times a week - Once a week - Several times a month - Once a month - Less often - Never

3.9 APPENDIX B. Classification scheme

Comprehension	Retrieval
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. Objectively wrong answer, question is misunderstood *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 *Reference period* 16. Reference periods are missing or undefined 17. Reference period carried over from earlier question	*Retrieval from memory* 18. High detail required or information unavailable 19. Long recall or reference period
Judgment	**Response Selection**
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 *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	

4 HOW DO RESPONDENTS PROCESS FORCED-CHOICE VS. CHECK-ALL-THAT-APPLY QUESTIONS? EVIDENCE FROM EYE TRACKING[24]

4.1 Abstract

Recent research has shown that the check-all-that-apply (CATA) and forced-choice (FC) question formats do not produce comparable results. The cognitive processes underlying respondents' answers to both types of question formats still require clarification. The present study contributes to filling this gap by using eye-tracking data. In a between-subject lab experiment (*n*=84), respondents answered two questions formatted either as CATA or as FC questions. Both question formats are compared by analyzing the amount of attention paid to the questions and the cognitive effort (operationalized by response latencies, fixation times, and fixation counts) respondents spend while answering the questions. Differences in cognitive effort are not found in the factual question. In the opinion question, the overall cognitive effort is higher in the FC than in the CATA format. The findings indicate that higher endorsement in FC questions cannot only be explained by the specific format. Other possible causes for these differences are discussed.

4.2 Introduction

Both check-all-that-apply (CATA) question formats as well as forced-choice (FC) question formats are commonly used in self-administered, visually presented surveys (paper-pencil/mail or web-based surveys; Thomas & Klein, 2006). In the CATA question format, respondents are presented with a list of response options and are asked to mark all that apply to them. Conversely, in the FC question format, the

[24] A version of this chapter is currently under review as:
Neuert, C. (under review). How do respondents process forced-choice vs. check-all-that-apply question formats? Evidence from eye tracking.
Parts of this chapter were presented at the 6th Conference of the European Survey Research Association, July 13-17, 2015, Reykjavik, Iceland, and at the 12th Conference of the European Sociological Association, August 25-28, 2015, Prague, Czech Republic.

response options are presented as a series of "yes/no" questions and the respondent explicitly indicates for each response option whether it applies or not. Recent experimental research has shown that both question formats do not produce comparable results (Nicolaas et al., 2015; Smyth et al., 2006): the mean number of response options marked with "yes" is higher in the FC format than the mean number of response options marked in the CATA format (Nicolaas et al., 2015; Rasinsky, Mingay, & Bradburn, 1994; Smyth et al., 2006, 2008; Thomas & Klein, 2006). However, the cognitive processes underlying respondents' answers to both types of question formats are unclear. Do FC formats lead respondents to read each answer option and to devote more thought to their responses? Are respondents who are presented with a multiple response list more likely to skim the list rather than read the items thoroughly, a behavior also known as "satisficing" (Krosnick, 1991)?

In this paper, I extend previous research by including eye tracking data to gain a better understanding of the cognitive processes underlying respondents' behavior when confronted with CATA and FC questions and to enhance our understanding of the differences in the response task. Understanding these differences is important for the interpretation of existing survey data and informed questionnaire design.

4.3 Previous research

In one of the earliest studies comparing CATA and FC questions, Rasinski et al. (1994) showed that, for the same three questions in a self-administered paper questionnaire, the mean number of response options marked with "yes" in the FC format is higher than the mean number of options selected in the CATA format. Smyth et al. (2006) extended this work to web surveys and also found that the FC format produced more "yes" responses than the CATA format. The higher endorsement in FC questions has been replicated across different types of questions (Smyth at al., 2006), countries and languages (Thomas & Klein, 2006), survey modes (web and telephone, Smyth et al., 2008; web, telephone and face-to-face, Nicolaas et al., 2015) and with variations of the classic "yes/no" wording (e.g., "Fan/Not a fan";

"Applies/Does not apply"; Smyth et al., 2006; Tsuchiya & Hirai, 2010). A literature review is provided by Callegaro et al. (2015). Theorizing about the reasons for these differences, Sudman & Bradburn (1982) argue that the response task and, consequently, respondents' strategies for answering are fundamentally different for the two formats. One possible explanation for respondents providing less "yes" responses in CATA formats is that this format might encourage a satisficing response strategy (Krosnick 1991, 1999; Krosnick & Alwin, 1987). In his work on satisficing, Krosnick (1991, 1999) argues that respondents may vary in how much cognitive effort they are willing or able to expend in answering questions. Respondents who are presented with a question in a CATA format are asked to choose only those options that apply. Such a response task may encourage respondents to minimize time and effort for answering the question by considering and selecting only the first response options (or the options that are most prominent, for other reasons) and then move on to the next question without paying sufficient attention to the remaining response options (Krosnick 1991, 1999; Krosnick & Alwin, 1987). In contrast, FC questions (with explicit "yes/no" categories) require respondents to consider each option and to provide an answer for each item individually. This may induce respondents to process every option more deeply before arriving at a decision; this should discourage a satisficing response strategy (Sudman, Bradburn, & Schwarz, 1996; Smyth et al., 2006).

In general, the time taken by respondents to provide an answer to a survey question is assumed to be a good indicator of the cognitive effort they invest in arriving at an answer or a judgment (Fazio, 1990). In their research, Smyth et al. (2006) used paradata to investigate the amount of time respondents spend on the respective question formats. They demonstrated that questions in a FC format took longer to answer than the same questions in a CATA format, indicating deeper processing of the questions in the former. They also found that respondents who spend less than the mean response time answering CATA questions are more likely to mark options affirmatively when they appear in the first three positions of the list rather than in the last three positions. Respondents who needed more than the mean

response time did not differ in the options marked affirmatively, compared to FC respondents.

Sudman & Bradburn (1982) also point to the difficulty in interpreting the responses themselves in these two types of question formats and propose avoiding the use of CATA questions. They argue that response options that are left blank in FC questions are more easily interpreted as missing data or undecided respondents. In contrast, the absence of a check in a CATA question might be due to several factors: (1) the option does not, in fact, apply to the respondent; (2) the respondent simply overlooked the option or did not notice it; (3) the respondent is neutral or undecided. Smyth et al. (2006) point to possible unintended consequences of an explicit "no" category for respondents who fall into the third category and are neutral or undecided. These respondents might be more likely to agree than to disagree; this is referred to as "acquiescence" or "agreeing response bias": the tendency to agree regardless of the content (Schuman & Presser, 1981).

However, FC formatted questions might also produce responses that are difficult to interpret if respondents do not complete the response task as requested. There are respondents who treat FC as CATA questions by ignoring the "no" category and check only within the "yes" category, which produces higher nonresponse in the data. Callegaro et al. (2015) remark that, when such response behavior occurs relatively often, the decision to exclude these cases from subsequent analysis is based on an ambiguous assumption, because this behavior may simply be an indication of not "yes". Smyth et al. (2006) hypothesize that the response pattern of treating a FC as a CATA question might differ between opinion questions and factual questions because the former require more consideration, whereas, for the latter, the information is typically readily available. Thus, answering behavior and factual questions may lead respondents to perform "quick clicking" what consequently, leads to an increased likelihood in missing or ignoring the "no" category. Another explanation is that respondents who mark the first response option affirmatively then continue to concentrate on the "yes" category and hardly even notice the "no" category (Smyth et al., 2005). To date, neither response format has

been shown to be more or less effective in reducing the effects of primacy or acquiescence (Schaeffer & Dykema, 2011).

My purpose in this study is to extend previous research in two important ways. First, I use eye-tracking data, which enables me to observe directly what respondents look at and what they do not look at while responding to questions. Second, tracking respondents' eye movements enables me to examine how much time respondents spend on each question format and allows me to analyze relatively direct measures of attention and cognitive effort, such as fixation times and fixation counts (Galesic & Yan, 2011). In contrast to previously used methods, such as response latencies, respondents' answers, or mouse movements, which could be described as relatively indirect data (Galesic et al., 2008), eye movement data can help to gain a deeper understanding about how the two question formats are processed. Although response latencies are good indicators for the overall cognitive effort involved in answering a question, eye tracking allows precise observation of participants' reading patterns and an examination of respondents' attention to specific parts of the question. Moreover, tracking of eye movements shows exactly where and for how long respondents look and whether they tend to skim lists of response options rather than read them thoroughly.

The link between eye movements and cognitive processing is based upon two common assumptions: the *immediacy assumption* and the *eye-mind assumption* (Just & Carpenter, 1980; Rayner, 1998). The immediacy assumption postulates that words or visual objects that are fixated by the eyes are processed immediately. The eye-mind assumption postulates 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: the time a respondent spends fixating an area of the question (screen) is (more or less) equal to the time this area is being processed (Staub & Rayner, 2007). Based on the findings reviewed above, I expect that the FC format will yield a higher mean number of marked options than the CATA format (Hypothesis 1). Furthermore, because the FC format is supposed to encourage respondents to consider each answer option

individually and because response times are generally assumed to reflect the cognitive effort that is required to answer a survey question, I expect the FC format to produce longer response times than the CATA format (Hypothesis 2). I also expect that this difference will be greater for the attitudinal than for the behavioral question, as most people have immediately accessible information on the behavioral question, whereas they have to form an opinion to answer attitudinal questions. Using question fixation times and counts as more direct indicators of respondents' attention and effort, I examine how much cognitive effort and attention is paid to different parts of the question. In general, I expect that more attention will be paid to questions in the FC format than in the CATA format (Hypothesis 3). In addition, because the CATA format might lead respondents to read and select only the first response options, I expect that the question format will affect the number of answer options actually read and considered (Hypothesis 4).

4.4 Method

4.4.1 Respondents and procedure

The eye-tracking experiment was conducted at the pretest laboratory of GESIS—Leibniz Institute for the Social Sciences in Mannheim, Germany in October and November 2012. A session took about one and a half hours and consisted of three parts. In the first part, participants completed an online questionnaire while their eye movements were tracked. In the second part, a cognitive interview was conducted (cf. Neuert & Lenzner, 2015). The present experiment was embedded in the third part, in which participants completed another online questionnaire containing several unrelated experiments (cf. Lenzner, Kaczmirek, & Galesic, 2014). A random number generator was used to assign one of the two question formats (CATA vs. FC) to each respondent when the 3-part experimental session started. Participants received €30 for participating in the entire study.

In total, 84 respondents participated in the eye-tracking experiment (41 in the CATA condition, and 43 in the FC condition). Respondents were between 17 and 76 years old (M = 36, standard deviation [SD] = 14.3) and 54% were female. Of the

respondents, 68% had received at least 12 years of schooling, 12% had received 10 years of schooling, and 20% had received 9 or less years of schooling. Most respondents were experienced computer and internet users who used computers and the internet on a daily basis (88% and 87%, respectively) and 81% had already participated in at least one web survey prior to this study. Technical difficulties prevented recording the eye movements of seven respondents. These recordings were excluded from the analysis of eye movements, leaving 77 respondents (38 in the CATA condition, and 39 in the FC condition).

In the experiment reported in this article, participants were seated in front of the eye tracker such that their eyes were approximately 60 cm from the screen. They were instructed to read at normal speed while responding to the questions. The experimenter initiated the standardized calibration procedure. After a successful calibration, the web questionnaire started and participants' eye movements were tracked. During questionnaire completion, the experimenter remained in the observer room, next to the laboratory, to assist in case of problems. The average web questionnaire completion time was approximately twelve minutes. In the first third of the questionnaire, all respondents received two questions on the role of government (taken from the ISSP 2006 Questionnaire; see Appendix A). These two questions were used to compute participants' average response time, average fixation time, and average fixation count; these were subsequently used as covariates in the analysis, to control for individual differences.

4.4.2 Questions

The experiment included two questions, one factual and one opinion question, presented in two different question formats: either in a CATA format or in a FC format. The first question (Q1) asks which of the listed technical equipment the respondent's household owns, with six answer options. The second question (Q2) asks about characteristics of a successful marriage with nine answer options (see section 4.9, Appendix B for screenshots). The questions were presented in German.

4.4.3 Eye-tracking equipment / Apparatus

Eye movement data were collected with the Tobii T120 Eye Tracker and were analyzed with Tobii Studio 3.2.1. The Tobii T120 is a remote eye-tracker embedded in a 17″ TFT monitor (resolution 1280 x 1024 pixels) with two binocular infrared cameras placed underneath the computer screen that provide unobtrusive recording of respondents' eye movements and permit head movements within a range of 30 x 22 x 30 cm. The sampling rate is 120 Hz, meaning that 120 gaze data points per second are collected for each eye. The web 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. This larger than usual display font size was used to maximize measurement precision and to enable identification of which response options respondents had actually read or not read.

4.5 Results

4.5.1 Number and percent of items marked

Table 4.1 shows the mean number and percentage of response options marked affirmatively in the CATA and the FC conditions. Overall, the CATA format yielded an average endorsement of 8.5 of the options (61.1%), while the FC format yielded an average of 9.7 (66.6%) endorsements (t = -2.624, p = .010). Individually, the differences are significant in only one of the two questions: The difference between the means in the factual question, "Which of the following does your household have?" was not significant (t = .310, p = .757) and none of the individual items were checked by a different percentage of respondents (Analysis shown in Appendix D). This is inconsistent with Hypothesis 1. For the second question, about the characteristics of a successful marriage, an average of 3.6 answer options were marked in the CATA format, whereas, in the FC version, the average number of options endorsed was significantly higher, namely 4.9 (t = -3.350, p = .001). Seven of nine response options were marked affirmatively by a greater percentage of respondents in the FC format, whereby four were significantly greater (see Appendix D, section 4.11). The results for the attitudinal question support Hypothesis 1.

Table 4.1. Mean number and percentage of items marked in the CATA and FC conditions.

	CATA		FC		Difference		One-sided *t*-test	
Questions	No.	%	No.	%	No.	%	*t*	*p*
Q1: household (6)	4.90	81.7	4.81	80.2	.09	1.5	.310	.757
Q2: Characteristics successful marriage (9)	3.63	40.4	4.86	52.9	1.23	12.5	-3.350	.001
Overall means	8.54	61.1	9.67	66.6	1.13	5.5	-2.624	.010

Note: Parentheses contain the number of response options offered for each question.

4.5.2 Cognitive effort and attention

As an indicator of respondents' cognitive effort, I used response latencies and the eye-tracking metrics question fixation time and question fixation count.

Response latencies. Response latencies were measured from the time when the page was loaded to when the respondent clicked the "submit" button to receive the next question. Table 4.2 shows the mean response latencies in the two conditions. Due to the skewed distribution of response times (Yan & Tourangeau, 2008), log transformed response latencies are reported as well (cf. Fazio, 1990).

For Q1, there was no significant difference in response latencies between conditions, whereas in Q2, respondents spent more time on answering the questions in the FC format than when answering them in the CATA format. To control for inter-individual differences in respondents' reading speed, an analysis of covariance was conducted with the mean log-transformed response latency as the dependent variable and respondents' baseline speed as a covariate. There was no significant difference in response latencies for Q1 ($F_{(1,81)} = .07$, n.s.). When answering Q2, respondents spent more time on the FC format than respondents on the CATA formatted question did ($F_{(1,81)} = 42.1$, $p = .00$). Thus, Hypothesis 2 is supported only for Q2.

Table 4.2. Means of response latencies in the CATA and FC conditions.

Questions	Raw response latencies (in sec.)			Log-transformed response latencies		
	CATA	FC	$F_{(1, 81)}$	CATA	FC	$F_{(1, 81)}$
Q1	14.01 (.79)	14.58 (.78)	.26	4.13 (.02)	4.14 (.02)	.07
Q2	25.52 (1.55)	38.32 (1.51)	34.84**	4.39 (.02)	4.56 (.02)	42.09**

Note: Standard errors in parentheses. Reported are estimated marginal means after controlling for respondents' baseline speed (covariate). *$p < .05$, ** $p < .01$

Fixation times and counts. To examine how much cognitive effort is invested in different parts of the question across conditions, I analyzed eye-tracking data on the basis of three predefined areas of the screen, so-called area of interests (AOI). The three AOIs were defined, covering the question text (Qtext), the answer options (Aboxes), and the whole question (Qtotal) (see Appendix C). Within these AOIs, *fixation time* and *fixation count* were considered as measures of the respondent's level of attention and amount of cognitive processing. *Fixation time* is the time spent looking at a target AOI. The interpretation of fixation time and fixation counts can be quite different, depending on the context, and it is still controversial whether long fixation times and high numbers of fixations are due to a more conscientious response style and greater interest or to difficulties in comprehending or encoding information (Jacob & Karn, 2003; Poole & Ball, 2005, Lenzner et al., 2011). Because, aside from the question format, neither the question text nor the formulation of the response options differed across conditions, longer fixation duration within a specific, predefined part of the question (target AOI) is interpreted as deeper processing and as an indicator of a higher degree of attention within this specific part. *Fixation count* is the total number of fixations within a target AOI. In the present experiment, fixation count is interpreted as a measure of attention to the respective part of the question across conditions. This means that AOIs that were fixated more frequently in one format received more attention than in the other question format (Ehmke & Wilson, 2007; Jacob & Karn, 2003; Poole & Ball, 2005).

Table 4.3 shows the mean fixation times and counts in the two conditions and the test statistics of ANCOVAS with baseline response time and fixation count as

covariates. Regarding Q1, there were no significant differences in fixation times and fixation counts for the whole question (Qtotal), for the question text (Qtext) or the area of the response options. Thus, Hypothesis 3 cannot be confirmed for Q1. This finding is not surprising, given that neither the number of response options checked nor the response latencies differed across question formats in this question. Nevertheless, it indicates that merely the differences in question format are not the reason for a longer response time *per se*.

Table 4.3. Mean fixation times and fixation counts in the CATA and FC conditions.

	Mean fixation time (in sec.)			Mean fixation count (n)		
AOIs	CATA	FC	$F_{(1, 77)}$	CATA	FC	$F_{(1, 77)}$
Q1						
question text	1.83	1.54	1.935	9.93	8.64	1.589
answer options	8.42	9.60	1.453	30.97	35.67	3.336
question total	10.69	11.63	.790	42.43	47.07	2.547
Q2						
question text	3.68	4.39	3.904	20.70	24.71	3.942
answer options	13.94	25.18	43.603**	62.86	95.96	42.972**
items only	10.81	10.62	.043	53.00	55.28	.355
question total	17.95	30.09	39.987**	84.55	23.39	36.210**

Note: Reported are estimated marginal means after controlling for the covariates respondents' reading rate and respondents' fixation rate, respectively.
*$p < .05$, ** $p < .01$

For Q2, statistically significant effects were found for both the fixation times ($F_{(1, 77)} = 39.9$, $p = .01$) and the fixation counts on the whole question ($F_{(1, 77)} = 36.2$, $p = .01$), with longer times and more fixation counts in the FC format, compared to the CATA format. Concerning the area of the answer options, the results show significantly longer fixation times ($F_{(1, 77)} = 43.6$, $p = .01$) and more fixation counts ($F_{(1, 77)} = 42.9$, $p = .01$) for the FC format, compared to the CATA question format. No significant differences were found for fixation times and fixation counts on the question text. Given that more response options were selected in the FC format and that the area of answer options required longer fixation times and more fixation counts, the question

about how long respondents stayed in the area of the pure items arose and whether respondents differed in the time required for processing the pure items. Therefore, an additional AOI covering the items without the answer boxes (items only) was defined. For this area, no significant differences in fixation times or counts between the two formats were found (see Table 4.3). This finding indicates that there is not a difference in processing or understanding the response options per se but in the time it takes to provide an answer. It is possible that the longer period of time spent on FC questions might be due to mechanical response steps, because FC questions require respondents to move the mouse and to click either "yes" or "no" for each item, while CATA grids demand a click on only one row (Thomas & Klein, 2006). Therefore, the time spent fixating on the items (items only) was subtracted from the total time spent looking at the area of the answer options (Aboxes). The difference in the time spent on the answer options without the items in the FC format is four times greater than that in the check-all format (3.14 vs. 14.56). To examine the amount of attention given to the area of answer options without considering the items, the number of all fixations within the items-only area (items only) was additionally subtracted from the area of the answer options (Aboxes). The results show four times more fixation counts in the FC format, compared to the CATA question format (9.86 vs. 40.67; $F_{(1, 77)} = 111.83$, $p = .00$). Although some of this additional time and attention was undoubtedly spent on mechanical response steps that are not required on the CATA format, such as clicking the "no" category, the extent of the time differences indicates that respondents spent more time on the FC format for cognitive reasons and not because of mechanical demands.

4.5.3 Number of options read

Because eye movements allow for a precise observation of respondents' reading patterns, and also to gain a better understanding of the response behavior, the eye-tracking videos were reviewed and coded, in addition to the analysis of the eye movement data. The videos were coded by counting the number of response options that had actually been read, in order to examine whether respondents actually read each response option or whether they satisficed and considered only the first options

in the list. All videos were coded by one coder, and a randomly selected subset of 25% of the total number of videos was independently coded by a second coder, to estimate reliability. The intercoder agreement between these two raters was 93% for both questions. The rare discrepancies were discussed until consensus was reached.

The reading patterns showed no significant difference between the mean number of response options read in the two conditions, both in Q1 (t = -.037, p = .97) and in Q2 (t = -1.42, p = .16). For Q1, the vast majority of respondents (97%) in both conditions read all six answer options. In both question formats, one respondent did not read the last answer option. For Q2, four respondents in the CATA format (10%) and one in the FC format (3%) did not read all of the nine response options. In total, three respondents ignored the last response option (two in CATA, one in FC). The other two respondents missed or overlooked the first and the second answer option of the check-all list, respectively. Contrary to Hypothesis 4, the observation of respondents' reading behavior did not reveal differences in the number of options read across question formats.

4.5.4 Item nonresponse in the forced-choice format

In this section, I analyze whether respondents completed the response task as requested or whether they treated FC-formatted questions as CATA questions by marking only within the "yes" category. An example of such a response behavior is depicted in Figure 4.1.

For Q1, only 2% of respondents appeared to treat the FC question as a CATA question, whereas, for Q2, the percentage of respondents who treated the FC questions in a CATA manner was 12%. Of these respondents, all marked at least the first response option affirmatively and, except for one, all respondents did not fixate on the "no" category. This indicates that they simply did not notice it. This explanation is also supported by response behavior that was observed when reviewing the eye movement videos. Two further respondents treated Q1 in a CATA manner by clicking only within the "yes" category until they noticed the "no" category in the last response option and marked only the last one with "no" immediately before clicking the "next" button.

Figure 4.1. Example of a respondent treating a FC question as a CATA question.

4.6 Discussion and conclusion

The study examined the cognitive processes underlying respondents' behavior when confronted with CATA and FC questions. For the factual question (Q1), which asked about technical equipment in the household, higher means of response options marked in the FC question format were not detected, compared to the CATA format. However, because there was also no difference in respondents' cognitive effort and attention respondents needed to answer the question, this result shows that FC questions do not automatically require more time to be answered. A possible explanation could be that the more stable or easily accessible information related to factual questions is less likely to produce response format effects. An evaluation of this hypothesis is an interesting avenue for future research.

For the opinion question (Q2), prior findings of higher endorsement in FC questions were replicated and Hypothesis 1 was confirmed. Regarding Hypotheses 2

and 3, the overall cognitive effort and attention (operationalized by response latencies, mean fixation time and mean fixation count) was higher in the FC than in the CATA format. Using eye tracking, it was possible to further analyze these differences in responding. Across formats there was no difference in the time required for processing the response options, indicating that there is no difference in processing or understanding the response options *per se*, but in the time it takes to provide an answer. Further analysis showed that some of the additional time was indeed spent on mechanical steps (such as moving the mouse, clicking either "yes" or "no") rather than on deeper cognitive processing. However, the extent of the time difference suggests that respondents spent more time on the FC format for cognitive reasons. The findings indicate that higher endorsement in FC questions cannot be explained by the different format but is due to the differences in the response tasks respondents have to perform when answering questions in both formats. Finally, the observations of the eye movements show that only few respondents did not consider the last response options, which does not indicate that one format is more likely to evoke to a satisficing response strategy.

There are several avenues for further research. First, in a fully crossed experimental design that varies type of question and number of response options, it would be possible to compare the format effect regarding question type by using different numbers of answer options, for example six, nine, and twelve. Second, respondents in the current experiment may differ from respondents in real field survey settings because they completed the online questionnaire in a laboratory and received payment for participation. This might have led to a more conscientious response style that might have prevented respondents from showing more satisficing response behavior, such as skimming lists of options. It would be interesting to examine whether the results generalize to the field, for example, with mobile eye-tracking technology. Third, it appears worthwhile to combine eye-tracking and cognitive interviewing techniques in order to further explore possible causes for differing answer distributions, depending on the question format, and to gain a better understanding of the underlying response task. Employing cognitive interviewing

techniques might also contribute to a better understanding of the reasons why response options are left blank, depending on format.

4.7 References

Callegaro, M., Murakami, M. H., Tepman, Z., & Henderson, V. (2015). Yes-no answers versus check-all in self-administered modes. *International Journal of Market Research*, *57*(2), 203-223.

Ehmke, C., & Wilson, S. (2007). Identifying web usability problems from eye-tracking data. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1* (pp. 119-128). Lanchester, UK: British Computer Society.

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. *Research methods in personality and social psychology*, *11*, 74-97.

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

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*(5), 892-913.

ISSP (2006). International Social Survey Programme 2004: Role of Government IV (ISSP 2006). GESIS Data Archive, Cologne, Germany, ZA4700. Source Questionnaire.

Jacob, R. J. K., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: ready to deliver the promises. In J. Hyönä, R. Radach & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 573-605). Amsterdam: Elsevier.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354.

Krosnick, J. A. (1999) Survey research. *Annual Review of Psychology*, *50*(1), 537–567.

Krosnick, J. A. (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*(2), 201-219.

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*(6), 743-764. doi:10.1177/0894439313517532

Lenzner, T., Kaczmirek, L., & Galesic, M. (2011). Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension. *International Journal of Public Opinion Research, 23*(3), 361-373. doi: 10.1093/ijpor/edq053

Neuert, C. E., & Lenzner, T. (2015). Incorporating eye tracking into cognitive interviewing to pretest survey questions. *International Journal of Social Research Methodology* online first.

Nicolaas, G., Campanelli, P., Hope, S., Jäckle, A. & Lynn, P. (2015). Revisiting "yes/no" versus "check all that apply": Results from a mixed modes experiment. *Survey Research Methods*, *9*(3), 189-204. doi: 10.18148/srm/2015.v9i3.6151

Poole, A., & Ball, L. J. (2005). Eye tracking in human-computer interaction and usability research: current status and future. In C. Ghaoui (Ed.), *Encyclopedia of human computer interaction*. Pennsylvania: Idea Group, Inc.

Rasinski, K. A., Mingay, D. & Bradburn, N. M. (1994). Do respondents really 'mark all that apply' on self-administered questions*? Public Opinion Quarterly 58*(3), 400-408.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124* (3), 372-422.

Schaeffer, N. C., & Dykema, J. (2011). Questions for surveys. Current trends and future directions. *Public Opinion Quarterly*, *75*(5), 909-961.

Schuman, H., & Presser, S. (1981*). Questions and answers in attitude surveys: experiments on question form, wording, and context*. New York: Academic Press.

Smyth, J. D., Christian, L. M. & Dillman, D. A. (2008). Does 'yes or no' on the telephone mean the same as 'check-all-that-apply' on the web? *Public Opinion Quarterly, 72*(1), 103–113.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly, 70*(1), 66-77.

Smyth, J. D., Dillman, D. A., Christian, L. M. & Stern, M. J. (2005). Comparing check-all and forced-choice question formats in web surveys. The role of satisficing, depth of processing, and acquiescence in explaining differences. Social and Economic Sciences Research Center, Pullman, Washington. Available online at: http://www.sesrc.wsu.edu/dillman/papers/2005/comparingcheckall.pdf (accessed 11 June 2015).

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. In: G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 327-342). Oxford, UK: Oxford University Press.

Sudman, S. & Bradburn, N. M. (1982). *Asking questions*. San Franciso, CA: Jossey-Bass.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Franciso, CA: Jossey-Bass.

Thomas, Randall K., & Jonathan D. Klein (2006). Merely Incidental? Effects of response format on self-reported behavior. *Journal of Official Statistics, 22*, 221-44.

Tsuchiya, T. & Hirai, Y. (2010). Elaborate item count questioning: why do people underreport in item count response? *Survey Research Methods*, *4*(3), 139-149.

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.

4.8 APPENDIX A. Additional questions to compute baseline speed, reading rate and fixation rate (covariate)

1. Was meinen Sie, wie erfolgreich ist zurzeit der Staat, wenn es darum geht mit Bedrohungen der inneren und äußeren Sicherheit Deutschlands umzugehen?

Antwortoptionen:

Sehr erfolgreich - ziemlich erfolgreich - weder noch - ziemlich erfolglos - äußerst erfolglos

English translation:

1. How successful do you think the government is nowadays in dealing with threats to Germany's security?

Answer options:

Very successful - Quite successful - Neither successful nor unsuccessful - Quite unsuccessful - Very unsuccessful

2. Und wie erfolgreich ist zurzeit der Staat, wenn es darum geht die Arbeitslosigkeit zu bekämpfen?

Antwortoptionen:

Sehr erfolgreich - ziemlich erfolgreich - weder noch - ziemlich erfolglos - äußerst erfolglos

English translation:

2. And how successful do you think the government is nowadays in fighting unemployment?

Answer options:

Very successful - Quite successful - Neither successful nor unsuccessful - Quite unsuccessful - Very unsuccessful

4.9 APPENDIX B. Screenshots and translations of questions

Q1: Question on technical equipment

Figure 4.2. Screenshot of question on technical equipment (Q1) in the CATA condition.

Figure 4.3. Screenshot of question on technical equipment (Q1) in the FC condition.

English translation of Q1:
Which of the following does your household have?

- washing machine
- dishwasher
- television
- DVD player
- personal computer
- landline phone

Q2: Question about characteristics of a successful marriage

Figure 4.4. Screenshot of question about characteristics of a successful marriage (Q2) in the CATA condition.

Figure 4.5. Screenshot of question about characteristics of a successful marriage (Q2) in the FC condition.

English translation of Q2:
Independently of whether you are married or not: which of the following is important for a successful marriage?

- Faithfulness
- An adequate income
- Being of the same social background
- Mutual respect and appreciation
- Shared religious beliefs
- Good housing
- Agreement on politics
- Sharing household chores
- Children

4.10 APPENDIX C. Areas of interest for the analysis of eye tracking data

Figure 4.6. Screenshot of Q2 in the CATA condition showing the areas of interest (AOIs) for the question text (Qtext), the answer options (Aboxes), and the whole question (Qtotal).

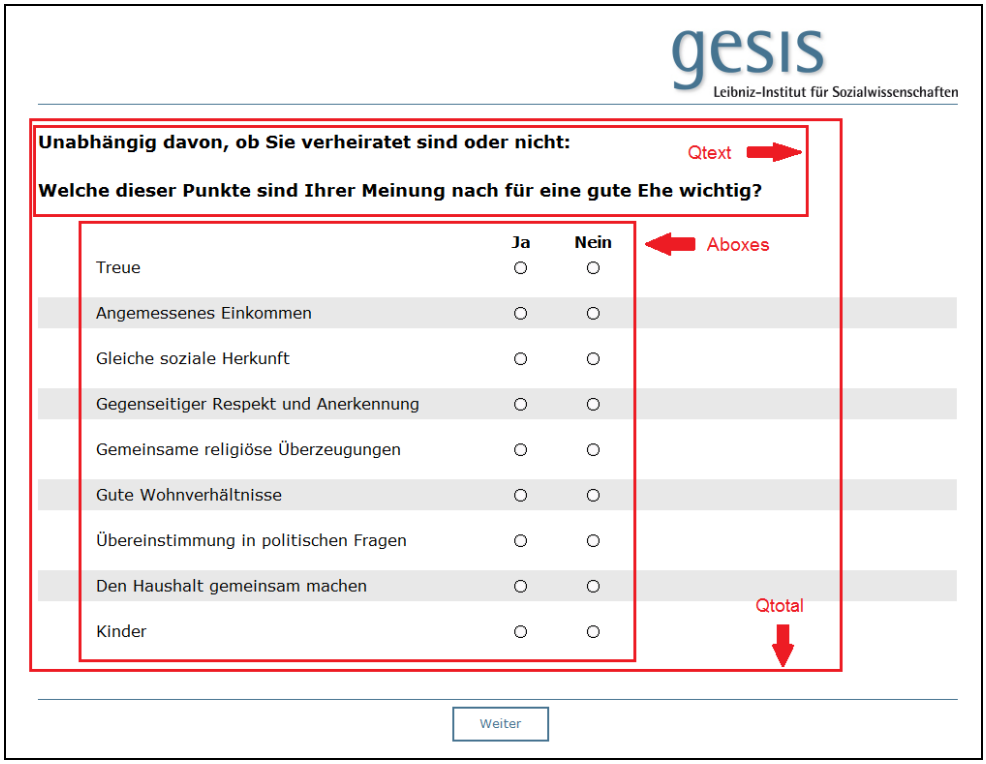

Figure 4.7. Screenshot of Q2 in the FC condition showing the areas of interest (AOIs) for the question text (Qtext), the answer options (Aboxes), and the whole question (Qtotal).

4.11 APPENDIX D. Percentage of items endorsed in the CATA and FC formats

Q1: Percentage of items endorsed in the CATA and FC formats.

	CATA	FC			CATA vs. FC	
Q1	Checked	Yes	No	Blank	χ^2	*p*
(n) = 84						
Washing machine	95.1	88.4	11.6	-	1.52	.263
Dishwasher	58.5	65.1	30.2	4.7	.39	.535
Television	90.2	81.4	16.3	2.3	1.34	.247
DVD player	85.4	79.1	18.6	2.3	.57	.451
Personal computer	87.8	95.3	-	4.7	1.56	.211
Landline phone	73.2	72.1	23.3	4.7	.01	.912
Mean (%)	81.7	80.2	16.7	3.1		
Mean (items=6)	4.90	4.81			t = .310, *p* = .757	

Q2: Percentage of items checked in the CATA and FC formats.

	CATA	FC			CATA vs. FC	
Q2	Checked	Yes	No	Blank	χ^2	*p*
(n) = 84						
Faithfulness	85.4	93.0	4.7	2.3	1.29	.257
An adequate income	22.0	44.2	46.5	9.3	4.67	.031*
Being of the same social background	7.3	27.9	55.8	16.3	6.07	.014*
Mutual respect and appreciation	100.0	100.0	-	-	-	-
Shared religious beliefs	9.8	10.7	62.8	16.3	2.00	.157
Good housing	31.7	58.1	34.9	7.0	5.92	.015*
Agreement on politics	12.2	11.6	74.4	14.0	.006	.936
Sharing household chores	58.5	86.0	14.0	-	7.99	.005*
Children	36.6	44.2	51.2	4.7	.503	.478
Mean (%)	40.4	52.9	38.3	7.8		
Mean (items = 9)	3.63	4.86			t = -3.350, *p* = .001	

5 APPENDIX

A. Eidesstattliche Erklärung

Hiermit erkläre ich, dass es sich bei der vorliegenden Dissertation mit dem Titel „Eye Tracking in Questionnaire Pretesting“ um mein eigenständig erstelltes Werk handelt. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtliche Zitate aus anderen Werken als solche kenntlich gemacht.

Mannheim, den 17.12.2015

Cornelia Neuert