Respondent and Response Behavior in Online Panel Surveys

Inauguraldissertation zur Erlangung des akademischen Grades eines Doktors der Sozialwissenschaften der Universität Mannheim

Vorgelegt von

Tobias Rettig

Dekan der Fakultät für Sozialwissenschaften
Prof. Dr. Michael Diehl

Betreuerinnen
Prof. Annelies G. Blom, PhD
Assistant-Prof. dr. Vera Toepoel

Gutachter
PD Dr. Tobias Gummer
Prof. Dr. Florian Keusch

Tag der Disputation
29.07.2022

Acknowledgements

I have received a great deal of support while working on this dissertation. Accordingly, I have many people to thank for their invaluable help. I want to thank my supervisor Annelies G. Blom for being an excellent mentor, for her great support and guidance during all stages of this dissertation, for the helpful feedback and the opportunities she has provided to me, and not least for her role in keeping this dissertation on track. I want to further thank Vera Toepoel for her additional supervision and guidance, and Tobias Gummer and Florian Keusch for reviewing this dissertation. The time and effort my supervisors and reviewers expended for this dissertation is deeply appreciated.

I want to further thank my coauthors Annelies G. Blom, Jan Karem Höhne, and Bella Struminskaya for their ideas, helpful comments, constructive feedback, and their overall contribution to our papers. Their input and expertise have benefitted this dissertation greatly and I am immensely grateful for their help.

I would like to further express my appreciation for all current and former colleagues at the German Internet Panel and the SFB 884 “Political Economy of Reforms”, chiefly Julian Axenfeld, Christian Bruch, Carina Cornesse, Barbara Felderer, Sabine Friedel, Marisabel Gonzalez Ocanto, Jessica Herzing, Marina Jesse, Uli Krieger, Eva Lübke, Marina Ungefucht, and Alexander Wenz, for always providing the positive and encouraging work environment that made the completion of this dissertation possible, and for the helpful discussions during our shared lunch breaks.

Finally, I want to thank my family, especially Manuela Rettig and Jörn Schroeder, Andrea, Inge and Lennart Rettig, and Brigitte Schroeder for their continued support and encouragement during all stages of my life. I want to especially thank Wiebke Wohltmann for her love and support, and for the times we spent working on our respective dissertations sitting back to back in our shared home office.

Table of Contents

1. General Introduction 1
    1.1 The cognitive response process 2
    1.2 Respondent inattention 4
    1.3 Memory effects 10
    1.4 Overview of this dissertation 12
        Paper I: Investigating Respondent Attention to Experimental Text Lengths 13
        Paper II: Memory Effects as a Source of Bias in Repeated Survey Measurement 14
        Paper III: Memory Effects: A Comparison Across Question Types 15
        Paper IV: Memory Effects in Online Panel Surveys: Investigating Respondents' Ability to Recall Responses from a Previous Panel Wave 17
    References 19
2. Investigating Respondent Attention to Experimental Text Lengths 25
    Abstract 25
    Keywords 25
    Acknowledgements 25
    2.1 Introduction 27
        2.1.1 Indicators of respondent attention 28
        2.1.2 Treating inattentive respondents 30
        2.1.3 Prior findings on respondent attention 31
    2.2 Research questions 33
    2.3 Data and method 34
        2.3.1 Experimental design 35
        2.3.2 Sample description 37
    2.4 Results 38
        2.4.1 Passing the attention check 38
        2.4.2 Response time as an indicator of attention 43
    2.5 Summary 46
    2.6 Conclusion 49
    References 52
    Appendix 56
3. Memory Effects as a Source of Bias in Repeated Survey Measurement 59
    Abstract 59
    Keywords 59

3.1 Introduction 60
3.2 Conceptualizing memory effects 61
3.2.1 Memory effects in repeated measurements of exactly the same question 63
3.2.2 Memory effects in a sequence of similar or related items 65
3.2.2.1 Memory effects in item batteries for latent constructs 67
3.2.3 The potential impact of memory effects 67
3.3 The state of the literature on memory effects in surveys 68
3.3.1 Literature on the cognitive response process 69
3.3.2 Literature on respondent memory 70
3.3.3 Literature on dependent interviewing 72
3.3.4 Literature on question order effects 73
3.4 Conclusion 76
References 79
4. Memory Effects: A Comparison Across Question Types 83
Abstract 83
Keywords 83
Acknowledgements 83
4.1 Introduction 84
4.2 Background and hypotheses 87
4.3 Methods 92
4.3.1 Study design 92
4.3.2 Data 96
4.3.3 Analytical strategy 97
4.4 Results 99
4.5 Discussion and conclusion 104
References 108
Appendix 112
5. Memory Effects in Online Panel Surveys: Investigating Respondents' Ability to Recall Responses from a Previous Panel Wave 115
Abstract 115
Keywords 116
Acknowledgements 116
5.1 Introduction 117
5.2 Background 119

5.3 Hypotheses 121
5.4 Data and method 126
5.4.1 Experimental design 126
5.4.2 Sample and variables 129
5.4.3 Analytical strategy 131
5.5 Results 133
5.6 Discussion and conclusion 142
References 147
Appendix 152
6. Conclusion 157
Paper I: Investigating Respondent Attention to Experimental Text Lengths 157
Paper II: Memory Effects as a Source of Bias in Repeated Survey Measurement 158
Paper III: Memory Effects: A Comparison Across Question Types 159
Paper IV: Memory Effects in Online Panel Surveys: Investigating Respondents' Ability to Recall Responses from a Previous Panel Wave 160
6.1 Overall research contribution and future prospects 161
References 165

List of Tables

2.1 Logistic regression models of passing the attention check 40
2.2 Pearson's correlations of logarithmized response time with passing the attention check by text length 44
2.3 Linear regression models of logarithmized response time across respondents who passed and failed the attention check 45
A2.1 Distributions of variables of interest across experimental groups and χ^2-tests for differences 56
A2.2 Model of passing the attention check with only text length and smartphone use as predictors 57
4.1 Wording and response scales of the test questions 92
4.2 Key indicators on alleged recall, correct recall, and recall certainty by question type 99
4.3 Key indicators on alleged recall, correct recall, and recall certainty by question type, extreme answers only 101
4.4 Regression models of alleged recall, recall certainty, and correct recall 103
A4.1 Original questions from ESS Round 8 112
A4.2 Wording and response scales of the follow-up questions 113
A4.3 Chi-squared tests of differences across experimental groups 113
A4.4 Chi-squared tests and one-way ANOVA for differences across question types 113
A4.5 Chi-squared tests and one-way ANOVA for differences between extreme and non-extreme answers 114
A4.6 Chi-squared tests and one-way ANOVA for differences across question types (extreme answers only) 114
5.1 Claimed recall, correct recall, weighted kappa, and mean recall certainty by question type 134
5.2 Logistic and linear regression models of claimed recall, recall certainty, and correct recall 137
5.3 Summary of results regarding our hypotheses 142
A5.1 English translations of the test questions and follow-up questions 152
A5.2 Coding scheme for education and age 153
A5.3 *T*-tests for differences in claimed recall and correct recall across question types 154
A5.4 *T*-tests for differences in correct recall between extreme and non-extreme responses by question types 154

A5.5 Logistic regression models of correct recall without claimed recall and recall certainty as additional predictors and separately by claimed recall 155

List of Figures

1.1 Model of the cognitive response process (adapted from Groves et al., 2009, fig. 7.1; Tourangeau et al., 2000) 3
1.2 The cognitive response process under weak or strong satisficing and inattention 6
2.1 The four versions of the attention check with the logo (1) and instruction (2)....36
2.2 Passing rates for the attention check across the four text length conditions....39
2.3 Density curves of response times in the attention check 43
3.1 Illustration of the cognitive response process by Tourangeau et al. (2000) extended by memory effects 63
3.2 Models of specific memory effects in the cognitive response process 65
4.1 Experimental design....95
5.1 Illustration of the experimental design....128
5.2 Differences between original response and recollection by claimed recall 135
5.3 Coefficient plot for the regression models of claimed recall, recall certainty, and correct recall 141

1. General Introduction

Surveys have been an important tool to gain insights for researchers across a variety of fields for many decades at this point (see, e.g., Groves et al., 2009): From medical surveys gathering patient data, to market researchers studying consumer satisfaction, polls predicting the outcomes of elections, and public opinion research informing politicians about the popularity of policies, to name only a few examples[1]. Across these various applications, survey researchers assume that respondents provide their "true" answers. We expect respondents to carefully consider each question and give a truthful response that accurately reflects how they feel about the topic at hand. However, a lot can go wrong during this process: Respondents may be distracted, not think carefully about the questions or misunderstand them, or they may misremember some information and thus give an inaccurate answer. The aforementioned are only some examples of the possible causes for measurement error; a situation in which the answer respondents give in a survey deviates from their "real" answer in some way (see, e.g., Groves et al., 2009, Chapter 2).

Whenever researchers draw any inference based on survey data, they need to be certain that the responses in their dataset accurately capture the information which they sought to measure from respondents with each question. Their response should accurately reflect respondents' true feelings toward an object or issue, their behaviors, or other factual information about them (Sudman & Bradburn, 1974). Any violation of this principle may lead to responses that do not truly capture the information researchers sought to measure with a given item (i.e., measurement error; Groves et al., 2009; Lynn, 2009) and based on these, researchers may draw inaccurate or incorrect conclusions. Researchers have long been acutely aware of this

[1] For a recent example of the impact surveys can make, we need to look no further than to the advent of the global COVID-19 pandemic. During a time when peoples' lives drastically changed on short notice, studies such as the Mannheim Corona Study (MCS; Blom et al., 2020) helped inform policy makers about public perceptions of the pandemic and the measures to combat it, as well as the impact these had on peoples' health and well-being.

problem and, consequently, a wealth of survey-methodological literature has been devoted to identifying, conceptualizing, measuring, preventing (or at least minimizing), and correcting for measurement error over the past decades. The present dissertation adds to the body of survey literature on measurement error by expanding both the current conceptual understanding and the empirical evidence on two potential sources of error in the response process: Respondent inattention and memory effects.

This chapter first gives an overview over the ideal cognitive response process respondents undergo to reach an optimal, unbiased response and how measurement error may arise from errors or interferences during this process in general. Subsequently, I look at the role of inattention and respondents' memory of previous responses as possible sources of measurement error specifically. This chapter furthermore provides an overview over the four papers which comprise this cumulative dissertation.

1.1 The cognitive response process

In their influential model of the cognitive response process, Tourangeau et al. (2000) propose four basic steps respondents undergo when answering a survey question (see also Cannell et al., 1981; Groves et al., 2009, Chapter 7): First, respondents read and comprehend the question. In this step, they seek to understand the meaning and focus of the question to identify which information it is asking from them. Second, respondents search their memory and retrieve the relevant information that is needed to answer the question. Third, based on the information respondents have retrieved in the previous step, they form a judgement about the issue at hand. Finally, in the fourth step, respondents select the appropriate response (i.e., the response option that most accurately represents their judgement; see Figure 1.1).

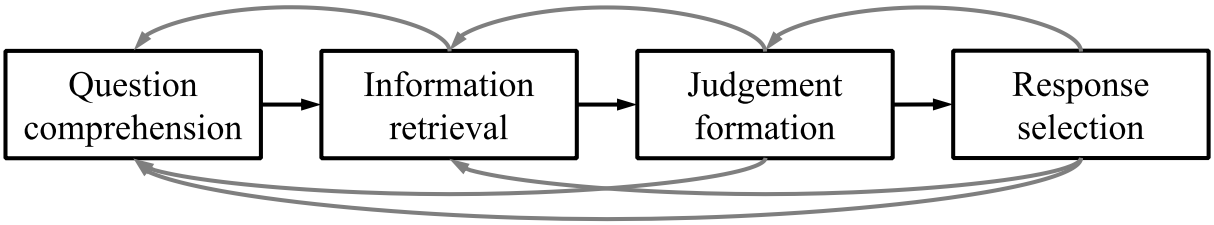

Figure 1.1. Model of the cognitive response process (adapted from Groves et al., 2009, fig. 7.1; Tourangeau et al., 2000).

These four steps are not necessarily completely distinct without any overlap, nor always undergone exactly once or in the presented order (Groves et al., 2009; Tourangeau et al., 2000). Respondents may, for example, already start forming a judgement while they retrieve additional information. They may also find that they cannot form a satisfactory judgement based on the information they have retrieved and thus return to seeking for additional retrieval cues in the question itself to retrieve more relevant information from memory. As Strack and Martin (1987) argue, respondents may also retrieve an existing judgement during the information retrieval step. Consequently, respondents would go from retrieving this judgement from memory to selecting a fitting response and thus skip forming an entirely new judgement as there is already one present.

During any stage of this ideal response process, errors or interferences may occur (Groves et al., 2009; Tourangeau, 1987). To name a few examples, during the question comprehension, respondents may misunderstand what exactly the question is asking and thus misidentify which information to retrieve (Bradburn & Sudman, 1991; Hippler & Schwarz, 1987), leading to a response that may accurately reflect respondents' views, but not on the exact topic researchers were interested in measuring. During the information retrieval stage, respondents may be unable to accurately recall all relevant information and thus form their response based on incomplete or inaccurate information (i.e., recall error; see, e.g., Eisenhower et al., 1991). Judgements may be influenced by factors that are not necessarily part of respondents' true opinions, such as their current mood (N. Schwarz & Strack, 1999) or their recent judgements

on other items (i.e., context effects; Schuman et al., 1983; Schuman & Presser, 1981; Tourangeau & Rasinski, 1988). Finally, respondents may select a response option that does not accurately reflect their judgement, for example because they failed to identify the most appropriate response option or because none of the available options fits them exactly (Bradburn & Sudman, 1991). Respondents may also make adjustments to their response for a number of reasons, such as presenting themselves more favorably, more consistently, or to appear more conforming to social norms (i.e., social desirability bias; Krosnick, 1999; Nederhof, 1985; Strack & Martin, 1987; Tourangeau, 1987; Tourangeau & Rasinski, 1988).

1.2 Respondent inattention

In addition to potential errors in the response process, researchers have found that respondents can display different response behaviors. Not all respondents are necessarily motivated to undergo all the cognitive steps as thoroughly as would ideally be required to reach an optimal response for every single question (Krosnick, 1991, 1999). Instead of this ideal processing of each question to report an optimized answer (aptly called “optimizing”; Krosnick, 1991) respondents may undergo a more superficial response process to reach a response that is “good enough” while expending less time and cognitive effort (i.e., satisficing; Krosnick, 1991). This may be done in an effort to reduce the burden that thoroughly answering the survey puts on respondents (for a comprehensive discussion on response burden, see, e.g., Yan et al., 2020). Instead of carefully considering each question, respondents may settle for a superficial understanding of its meaning before retrieving the necessary amount of information to answer. Furthermore, some literature suggests that respondents do not usually keep searching their memory for every piece of information that is relevant to a given question, but instead stop once they have retrieved enough information to form a satisfactory judgement (Tourangeau et al., 1989; Tourangeau & Rasinski, 1988). This subset of the information from respondents’ memory is in turn likely to be biased towards that which

comes to the respondent's mind first (i.e., the most accessible information; Strack & Martin, 1987; Tourangeau & Rasinski, 1988). In selecting a response, respondents may pick the first appropriate option that seems reasonably fitting to them and subsequently stop searching the remaining response options for one that may fit their actual judgement more accurately (Couper et al., 2004; Krosnick & Alwin, 1987).

The less thorough response process outlined above, in which respondents perform all steps of the cognitive response process but do so only superficially is classified by Krosnick (1991) as "weak satisficing". In contrast, "strong satisficing" describes a process in which respondents skip some of these steps altogether by going from a superficial understanding of the question to selecting a response that seems like a reasonable answer to that question but is completely divorced from any actual views respondents hold[2].

Going beyond what Krosnick describes as strong satisficing, inattentive respondents in self-administered survey modes (such as online surveys) have the ability to forego reading the question altogether and skip to selecting any of the response options arbitrarily (Anduiza & Galais, 2017). Not having read the question at all (or at least not thoroughly enough to fully grasp its meaning), respondents bypass undergoing even a superficial comprehension of the question, and thus fully omit the first step of the cognitive response process. As respondents who have not read or comprehended the question do not know which information is requested from them, the retrieval of relevant information, formation of a judgement based on this information, and selection of an appropriate response are impossible (Tourangeau et al., 2000). As with strong satisficing, respondents who are inattentive (sometimes also referred to as random, careless or insufficient effort responding; see, e.g. Curran, 2016; Maniaci & Rogge, 2014) therefore also omit the steps of retrieving information and forming a judgement.

[2] While optimizing, weak- and strong satisficing could be viewed as different categories (or ideal types) of response behavior, it should be noted that Krosnick (1991) describes optimizing and strong satisficing as the endpoints of a continuum of a more or less thorough question processing instead.

Figure 1.2 illustrates this response process for inattentive respondents in comparison to the ideal cognitive response process, as well as weak and strong satisficing.

As inattentive respondents undergo an even less thorough response process (in which no actual processing of the question is performed at all), Anduiza and Galais (2017) describe inattentive responding as a "yet stronger type of satisficing" (p. 499) than strong satisficing. One might argue, however, that satisficing as proposed by Krosnick (1991) at the minimum entails a surface-level processing of the question itself and the selection of a "good enough" response. Respondents who arbitrarily select a response option without reading the question are, however, not necessarily expending this effort to maintain the appearance of giving reasonable responses. This response behavior may thus conceptually no longer qualify as satisficing. Regardless, the result of strong satisficing and inattentive responding for researchers is effectively the same: They receive a response that in no way reflects the real information they sought to measure from respondents.

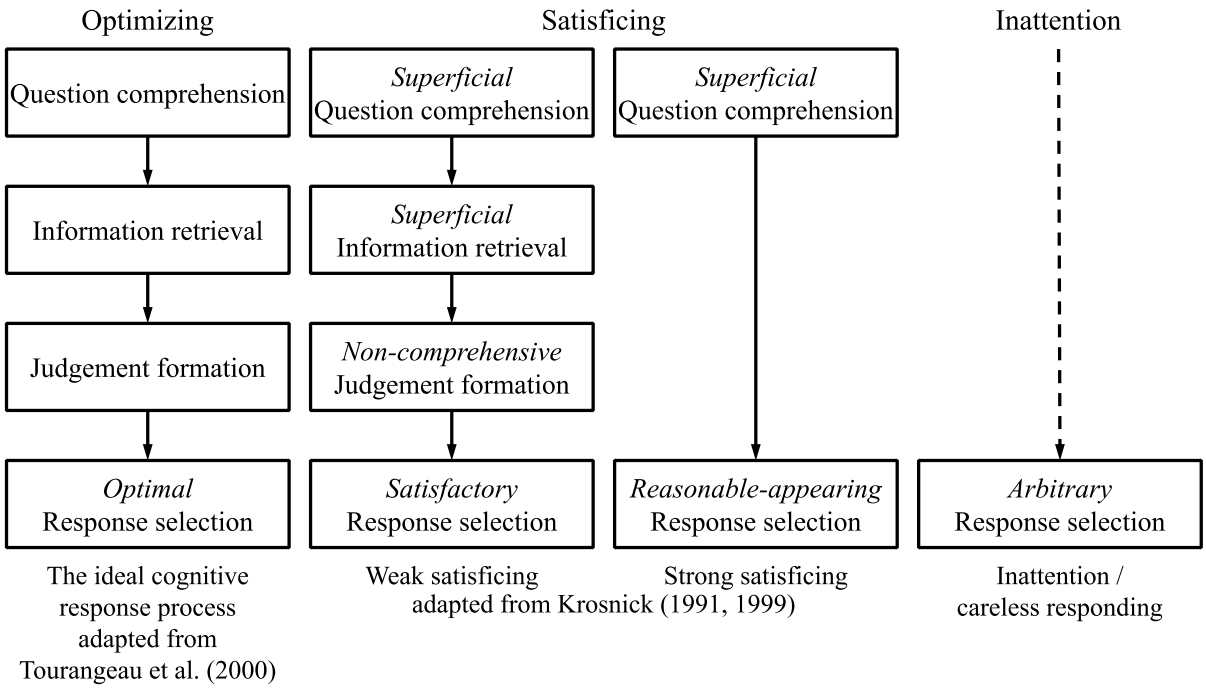

Figure 1.2. The cognitive response process under weak or strong satisficing and inattention.

However, while the data which inattentive respondents provide inaccurately reflect their "true" answers, the measurement error resulting from inattentive responding has traditionally been assumed to be a random phenomenon (see Curran, 2016). Data from inattentive respondents were thus presumed to potentially weaken statistical power or mask relationships between variables, but not systematically introduce any biases into analyses of the data. More recent research has shown, however, that inattentive respondents can produce nonrandom response patterns which may distort or produce correlations between variables that are not present for attentive respondents (Curran, 2016; Huang et al., 2015; Meade & Craig, 2012)[3].

In response to this type of response behavior, researchers have devised a number of methods for detecting (and subsequently excluding) inattentive respondents (for a review see, e.g., Curran, 2016). Especially in online surveys, where the completion time (i.e., the time respondents spent to answer the survey) is easily measured, short completion times are commonly used as an indicator of respondents who rushed through the survey without paying adequate attention (i.e. speeding; Zhang & Conrad, 2014). This practice relies on two basic assumptions: First, that a reduction of the time and effort spent on answering the survey is a key motivation for giving careless responses and second, that reading, processing, and thoughtfully answering questions takes respondents a certain minimum amount of time (Curran, 2016). In other words, researchers use "speeding" through the questionnaire as an indicator of inattention (and/or satisficing) because it can be assumed to be both typical for inattentive respondents and atypical for attentive respondents. However, a major drawback of this approach is that in practice, selecting a cutoff value (i.e., deciding when a fast response time becomes "too fast") has proven difficult (Curran, 2016; Niessen et al., 2016).

[3] Consequently, as Curran (2016) notes, this shift in researchers' understanding of inattentive responses has been reflected in a shift in terminology away from random responding and towards terms such as careless or insufficient effort responding over the past decade.

Other approaches to identify satisficing and inattentive respondents in a dataset include scanning their responses for patterns that may be indicative of low effort or inattention during the interview. Such patterns may, for example, include selecting the same scale point for a series of questions with little or no variation (i.e., straightlining; see, e.g., Schonlau & Toepoel, 2015). Researchers may also look for patterns of inconsistent responses, such as strong agreement with incompatible items (e.g., a positive and a negative version of otherwise identical statements) or, on the contrary, widely differing answers to very similar (or repetitions of the same) items (see Curran, 2016). These indirect attention measurements have the advantage that they can be applied to "regular" survey questions. Thus, they do not require researchers to use some of their (often very limited) questionnaire space on more specialized attention measurements and can often be applied to existing survey data ex post. These attention indicators have, however, also been noted to flag different respondents for inattention depending on which indicator is used (Curran, 2016; Meade & Craig, 2012) and their effectiveness at reliably identifying careless respondents has been called into question (Leiner, 2019).

To observe respondents' attention more directly, researchers have also developed specialized items and tasks (often referred to as "attention checks") that are designed to elicit a "correct" response from attentive respondents. These include items which directly ask respondents to pick a specific response option as well as nonsensical (or "bogus") items (e.g., respondents should not report that they have ever suffered a fatal heart attack because this is an implausible response; Paolacci et al., 2010). Such items provide a straightforward way of differentiating attentive (correct) from inattentive (incorrect) responses. However, correct responses may in some cases be a result of inattentive respondents picking the correct option by chance. Incorrect responses may in turn be caused by an error in the question comprehension or response selection (e.g., selecting "agree" when instructed to select "completely agree" because respondents missed the word "completely" or picking the fatal

heart attack because they misinterpreted the word “fatal”) and thus not necessarily a clear sign of inattention (see Curran & Hauser, 2019 for a discussion).

Going one step further, Oppenheimer et al. (2009) proposed the inclusion of a manipulation check into the instructions accompanying a survey question and asking respondents to perform an action that would not be expected when answering normally (e.g., responding “I read the instructions” to an open question instead of giving a substantive response). Such a task disrupts the normal procedure respondents perform when selecting their response, which inattentive respondents (and strong satisficers) emulate and, thus, virtually eliminates the possibility that respondents may pass this type of attention check due to chance.

Due to the reactive nature of these direct attention checks (i.e., at least the respondents who understood and passed the attention check will become aware that their attention is being monitored), some researchers have expressed concern that respondents may react negatively or subsequently be influenced in their response behavior to later items in the survey (e.g., Curran, 2016; Oppenheimer et al., 2009). Empirical investigations into possible influences of attention checks on the response behavior to subsequent items have yielded somewhat mixed results but mostly found little to no evidence of any undesired effects on response behavior (Berinsky et al., 2014; Breitsohl & Steidelmüller, 2018; Gummer et al., 2021; Hauser et al., 2016; a notable exception can be found in Hauser & Schwarz, 2015).

While including inattentive respondents may adversely affect data quality and the conclusions researchers draw based on these data (see Huang et al., 2015), removing inattentive respondents reduces the overall sample size and can introduce bias into the composition of the remaining sample (Aronow et al., 2019). In addition, as outlined above, researchers have employed a variety of different and imperfect methods for detecting inattentive respondents, which have yielded different results regarding which and how many respondents are classified as inattentive (Curran, 2016; Meade & Craig, 2012). More research is therefore needed to

provide survey researchers with further insight into the prevalence of inattentive responding, correlates of inattention, which method(s) to use for identification, how to treat inattentive respondents, and, crucially, how to optimize their surveys to prevent respondents from becoming inattentive in the first place. The present dissertation contributes to this research on respondent attention.

1.3 Memory effects

Respondents who are answering a survey attentively may experience a different type of interference in their cognitive response process specific to repeated survey measurements: Memory effects (see van Meurs & Saris, 1990). When respondents are repeatedly asked the same (or substantially similar) survey questions, their first step in processing a question may not only entail reading and comprehending it, but also recognizing that the question has been asked before. Subsequently, if their previous response is still present in respondents' memory, it may be one piece of the information they retrieve and use to form a judgement. This second judgement would in turn not be a new and independent judgement but one that was evaluated against (and adjusted to) respondents' previous judgement.

One potential danger of such an evaluation against an earlier answer may be that respondents give artificially time-consistent responses out of a desire to present themselves as being consistent (Alwin, 2007; Polit, 2014; Tourangeau et al., 2000). Alternatively, respondents may satisfice by treating their previous response as an already existing satisfactory judgement and opt to simply reiterate their old response instead of forming a new judgement[4]. Either of these avenues may produce artificially elevated consistency in respondents' answers (Alwin, 2007), which may in turn lead researchers to underestimate change of the underlying

[4] These two proposed ways in which later responses may be influenced by previous responses are discussed in detail in Chapter 3.

information between measurement repetitions (or, conversely, overestimate its stability). This type of measurement error induced by respondents' memory of their previous responses (i.e., memory effects; van Meurs & Saris, 1990) has seen a renewed interest from researchers in recent years.

A defining feature of longitudinal (panel) surveys is that the same respondents are surveyed repeatedly at different points in time (Lynn, 2009). This gives longitudinal designs the ability to observe change over time on the individual level by taking repeated measurements of the same information, often in regular intervals, which is considered one of their key advantages (Lynn, 2009; Lynn & Lugtig, 2017). The potential impact of memory effects on the responses to measurement repetitions in longitudinal settings has come into focus amidst recent trends towards quicker and more frequent (online) data collection (e.g., Blom et al., 2020). As information tends to be forgotten over time (see, e.g., Tourangeau et al., 2000), more frequent repetitions of survey questions after a shorter in-between time make it more likely that respondents can remember their previous response at the time a question is repeated (and thus experience memory effects).

In addition to longitudinal surveys measuring change over time, repetitions of the same questions – often after relatively short time intervals – are also commonly used to evaluate the effectiveness of experimental treatments by comparing repeated measurements taken before and after the treatment is administered (i.e., pretest-posttest designs; Campbell & Stanley, 1966) or to evaluate the quality of a survey measurement by using indicators such as its test-retest reliability (see, e.g., Saris & Gallhofer, 2014).

Although several authors have acknowledged the potential problem that respondents who remember their previous responses pose in repeated survey measurements (e.g., Alwin, 2007, 2011; Moser & Kalton, 1972; Polit, 2014; Saris & Gallhofer, 2014), few studies have empirically examined memory effects in practice. A notable early example is the study by van

Meurs and Saris (1990) and, expanding upon their design, researchers have taken renewed interest in investigating memory effects with a few studies in recent years (Höhne, 2021; Revilla & Höhne, 2021; H. Schwarz et al., 2020). However, the literature currently still has considerable gaps both in the conceptual understanding of how remembering a previous response interferes with respondents' cognitive response process and the empirical evidence on memory effects. The present dissertation adds to the understanding and empirical evidence on memory effects by providing a review of the state of the literature on memory effects, developing a conceptual framework of memory effects in the cognitive response process, and adding to the empirical evidence on respondents remembering their previous responses.

1.4 Overview of this dissertation

The present dissertation examines two distinct sources of measurement error in survey research, specifically in the context of online panel surveys – inattention and memory effects – through the lens of their interference with respondents' cognitive response process. Overall, the dissertation consists of four papers with each of them focusing on a different aspect of these two concepts. The first paper is an empirical study of respondent attention by way of a randomized survey experiment on respondents' attention to text stimuli of differing lengths. The second paper provides a theoretical conceptualization of memory effects and how these interfere with respondents' cognitive response processes, as well as a literature review of the existing research on memory effects and adjacent concepts from the wider survey-methodological literature. The third paper is an empirical investigation into respondents' ability to remember their previous responses to different questions short-term (within one survey wave). Finally, the fourth paper expands upon the design of the third paper and investigates respondents' ability to remember their previous responses longer-term (across multiple survey waves). Data collection for all studies in this dissertation was conducted in

the German Internet Panel (GIP), a probability-based online panel of the German population living in private households aged 16 to 75 years at the time of recruitment (Blom et al., 2015).

Paper I: Investigating Respondent Attention to Experimental Text Lengths

In this paper, I investigate whether online panel respondents read experimental treatment texts depending on their length. The paper provides a literature review of previous studies on respondent attention and their findings on the prevalence and correlates of inattention. This literature review reveals that these studies have employed various different methods to identify inattentive respondents and their findings on the proportion of inattentive respondents in a survey and predictors of inattention have been mixed. In addition to the inconclusive evidence on respondent attention, nearly no such research has been conducted in probability-based samples so far, which merits further research on this issue.

The paper investigates respondents' attention to experimental treatments by implementing an instruction manipulation check into the text preceding a survey question, which instructed respondents to click a logo to proceed instead of answering the question. The length of the surrounding treatment text was experimentally varied across respondents and ranged from only the instruction itself to placing the instruction within four paragraphs of text. Based on this experiment, the paper investigates the prevalence of inattention, its relationship with text length, correlates of inattention, and compares the results of the attention experiment to response time as a proxy for inattention.

A key result of this study is that the length of a text and whether respondents read it are closely related. Most respondents passed the attention check if they received the shortest text condition, while most respondents who received the longest text failed. This reveals both potential to optimize surveys for respondent attention by avoiding lengthy stimuli, but also problems for any studies that compare the effectiveness of different treatments when these have substantially different lengths. The study further finds that inattention is correlated with

respondents' age, gender, education level, and level of panel experience. Removing inattentive respondents from analyses is thus likely to result in a biased sample.

Another key finding of this study is that while response time is moderately to strongly correlated with attention (particularly for longer texts), it is not suitable as a standalone attention indicator. Setting a certain minimum response time as a cutoff value to mark fast respondents as inattentive bears considerable risk of systematically misidentifying fast readers as inattentive and non-speeding inattentive respondents as attentive.

Finally, this paper outlines the need for further research to guide researchers in deciding how to treat inattentive respondents. In particular, more research is needed to determine whether and how inattentive respondents may be successfully treated with direct interventions to encourage attentive participation in the survey instead of passively removing them from analyses. Further research is also needed to investigate whether inattention at one point in time is a good indication that respondents will continue to answer future surveys without the required care and attention (and thus, whether their removal from panel studies should be considered).

Paper II: Memory Effects as a Source of Bias in Repeated Survey Measurement

In the second paper, I develop a conceptual framework of how memory effects may interfere with respondents' cognitive response process. I propose that while comprehending the question, respondents may also recognize it and subsequently retrieve their previous response during the information retrieval stage. I then propose two possible avenues for memory effects interfering with the response process: (1) respondents use their previous response as a basis for reflection during the judgement formation in an attempt to form a response that is consistent with their previous response (called the "memory consistency" model), and (2) respondents treat their previous response as an existing judgement and thus forego forming a

new judgement and instead elect to simply reiterate the previous response (called the "memory satisficing" model).

The paper includes a literature review of the existing research on respondent memory and outlines gaps in the current understanding of memory effects. It further conceptually integrates memory effects into the wider literature on the cognitive response process and adjacent response effects. In particular, I argue that context effects (or question order effects) can be viewed as a form of memory effects, as respondents' memory of answering substantially similar questions and their previous responses is causing the interference.

The key insights from the literature review are that some empirical evidence indicates that a majority of respondents can remember their previous responses within one survey, and that even a repeated measurement after two weeks may not be free from memory effects. However, the existing knowledge on memory effects is insufficient to guide researchers in establishing time intervals for repeated survey measurements which can be assumed to be free from measurement error due to respondents remembering their previous responses. Such guidance has become especially crucial for applied survey researchers amidst recent trends towards faster and more frequent data collection, especially in the context of online panels. In addition, further research is needed to determine how commonly respondents use their previous response as a basis for reflection or a way to satisfice.

Paper III: Memory Effects: A Comparison Across Question Types

In the third paper, I investigate the ability of online survey respondents to remember their previous responses to different types of questions within one survey. I examine the results of a randomized survey experiment which allows an investigation into the extent to which respondents were able to correctly recall their responses within one panel wave (after about 20 minutes). I further investigate how respondents' memory of their previous responses differs across questions dealing with their beliefs, attitudes, or behaviors, extreme and non-

extreme responses, experienced and freshly recruited respondents, and across socio-demographic groups.

A key finding of the study is that respondents claim to remember their response and are able to correctly repeat it in most cases. In addition, responses to the three question types are remembered at different rates, extreme responses are more likely to be remembered, and remembering previous responses is correlated with respondents' age, gender, and education level. These results have several implications for research designs which incorporate repeated measurements after a short time, such as pretest-posttest or test-retest designs. In such studies, respondents are likely to remember their previous responses and thus at risk of experiencing memory effects which introduce measurement error into their responses. In addition, I find that these effects are likely to systematically vary across different questions and across (socio-demographic) groups of respondents, posing a problem for comparisons across these groups. Differences in the effectiveness of an experimental treatment across socio-demographic groups in a pretest-posttest design may, for example, be an artifact of responses to the posttest measurement that are affected by the responses to the pretest measurement at different rates.

The paper also outlines remaining gaps in the literature on memory effects that merit further research. While the evidence suggests that respondents are likely to remember their responses within one survey, further research is needed to examine how much this presence of their previous response in respondents' memory affects their later responses in practice. In addition, this study indicates that a time interval of 20 minutes is insufficient to allow respondents to forget their previous responses. However, further research is required to guide researchers in establishing time intervals for repeated survey measurements which can be assumed to be free from memory effects. Finally, the paper points out that nearly all research on memory effects thus far has been conducted in self-administered web surveys. Additional research is thus needed to examine how these results translate to other survey modes (e.g.,

interviews without a visual presentation of the response options, which may serve as a recall cue).

Paper IV: Memory Effects in Online Panel Surveys: Investigating Respondents' Ability to Recall Responses from a Previous Panel Wave

In the final paper of this dissertation, I investigate to what extent online panel respondents remember their responses from a previous panel wave. Expanding upon the experiment described in the third paper, this study extends the investigation of memory effects across different question types to a longitudinal context. In particular, I investigate whether respondents remember their responses to the same questions after a time interval of 4 months (2 panel waves later), potential differences across question types, and correlates of remembering a previous response including socio-demographics, giving an extreme response, panel experience, as wells as respondents' self-reported response burden and survey enjoyment. I further investigate how far off from their original answer respondents were if they failed to correctly remember it.

I find some evidence that after four months, some respondents are still able to remember their responses from a previous panel wave correctly (i.e., not all cases in which respondents are able to repeat their original response are explained by a stable opinion or chance). However, this group is relatively small, and respondents are unable to remember their response in most cases. Respondents who give an incorrect recollection of their previous response are most commonly off by a single scale point. I further find that responses to different types of questions are remembered at different rates, extreme responses are more likely remembered, and that remembering a previous response is correlated with gender. However, I find no relation between remembering a response from a previous panel wave and age, education level, panel experience, response burden, or survey enjoyment.

The primary conclusion of this study is that while I find some evidence for memory effects even after 4 months, the number of affected respondents is likely to be small and not systematically different from unaffected respondents. I therefore conclude that for repeated survey measurements after a time of four or more months, measurement error due to memory effects may be negligible. However, the paper also outlines the need for further research on memory effects in survey measurements with more than two repetitions, as the repeated presentation and answering of the same questions may lead respondents to remember these questions and their answers to them more clearly or form a “routine” response they give every time.

References

Alwin, D. F. (2007). *Margins of Error. A Study of Reliability in Survey Measurement*. John Wiley & Sons.

Alwin, D. F. (2011). Evaluating the Reliability and Validity of Survey Interview Data Using the MTMM Approach. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), *Question Evaluation Methods* (pp. 265–295). John Wiley and Sons.

Anduiza, E., & Galais, C. (2017). Answering Without Reading: IMCs and Strong Satisficing in Online Surveys. *International Journal of Public Opinion Research*, *29*(3), 497–519. https://doi.org/10.1093/ijpor/edw007

Aronow, P. M., Baron, J., & Pinson, L. (2019). A Note on Dropping Experimental Subjects who Fail a Manipulation Check. *Political Analysis*, *27*(4), 572–589. https://doi.org/10.1017/pan.2019.5

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-administered Surveys. *American Journal of Political Science*, *58*(3), 739–753. https://doi.org/10.1111/ajps.12081

Blom, A. G., Cornesse, C., Friedel, S., Krieger, U., Fikel, M., Rettig, T., Wenz, A., Juhl, S., Lehrer, R., Möhring, K., Naumann, E., & Reifenscheid, M. (2020). High-Frequency and High-Quality Survey Data Collection: The Mannheim Corona Study. *Survey Research Methods*, *14*(2), 171–178. https://doi.org/10.18148/srm/2020.v14i2.7735

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting Up an Online Panel Representative of the General Population: The German Internet Panel. *Field Methods*, *27*(4), 391–408. https://doi.org/10.1177/1525822X15574494

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 Error in Surveys* (pp. 29–40). John Wiley & Sons. https://doi.org/10.1002/9781118150382

Breitsohl, H., & Steidelmüller, C. (2018). The Impact of Insufficient Effort Responding Detection Methods on Substantive Responses: Results from an Experiment Testing Parameter Invariance. *Applied Psychology*, *67*(2), 284–308. https://doi.org/10.1111/apps.12121

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin Company.

Cannell, C. F., Miller, P. V, & Oksenberg, L. (1981). Research on Interviewing Techniques. *Sociological Methodology*, *12*, 389–437.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Crawford, S. D. (2004). What They See Is What We Get. Response Options for Web Surveys. *Social Science Computer Review*, *22*(1), 111–127. https://doi.org/10.1177/0894439303256555

Curran, P. G. (2016). Methods for the Detection of Carelessly Invalid Responses in Survey Data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Curran, P. G., & Hauser, K. A. (2019). I'm Paid Biweekly, Just Not by Leprechauns: Evaluating Valid-but-incorrect Response Rates to Attention Check Items. *Journal of Research in Personality*, *82*, 103849. https://doi.org/10.1016/j.jrp.2019.103849

Eisenhower, D., Mathiowetz, N. A., & Morganstein, D. (1991). Recall Error: Sources and Bias Reduction Techniques. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Error in Surveys* (pp. 127–144). John Wiley & Sons. https://doi.org/10.1002/9781118150382

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (2nd ed.). John Wiley & Sons.

Gummer, T., Roßmann, J., & Silber, H. (2021). Using Instructed Response Items as Attention Checks in Web Surveys: Properties and Implementation. *Sociological Methods and Research*, *50*(1), 238–264. https://doi.org/10.1177/0049124118769083

Hauser, D. J., & Schwarz, N. (2015). It's a Trap! Instructional Manipulation Checks Prompt Systematic Thinking on "Tricky" Tasks. *SAGE Open*, *5*(2). https://doi.org/10.1177/2158244015584617

Hauser, D. J., Sunderrajan, A., Natarajan, M., & Schwarz, N. (2016). Prior Exposure to Instructional Manipulation Checks does not Attenuate Survey Context Effects Driven by Satisficing or Gricean Norms. *Methods, Data, Analyses*, *10*(2), 195–220. https://doi.org/10.12758/mda.2016.008

Hippler, H.-J., & Schwarz, N. (1987). Response Effects in Surveys. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social Information Processing and Survey Methodology. Recent Research in Psychology*. (pp. 102–122). Springer.

Höhne, J. K. (2021). New Insights on Respondents' Recall Ability and Memory Effects When Repeatedly Measuring Political Efficacy. *Quality and Quantity*. https://doi.org/10.1007/s11135-021-01219-2

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient Effort Responding: Examining an Insidious Confound in Survey Data. *Journal of Applied Psychology*, *100*(3), 828–845. https://doi.org/10.1037/a0038510

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology*, *50*, 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

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. https://doi.org/10.1086/269029

Leiner, D. J. (2019). Too Fast, Too Straight, Too Weird: Non-Reactive Indicators for Meaningless Data in Internet Surveys. *Survey Research Methods*, *13*(3), 229–248. https://doi.org/10.18148/srm/2019.v13i3.7403

Lynn, P. (2009). Methods for Longitudinal Surveys. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (pp. 1–19). John Wiley & Sons. https://doi.org/10.1002/9780470743874.ch1

Lynn, P., & Lugtig, P. J. (2017). Total Survey Error for Longitudinal Surveys. In P. P. Biemer, E. D. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total Survey Error in Practice* (pp. 279–298). John Wiley & Sons. https://doi.org/10.1002/9781119041702.ch13

Maniaci, M. R., & Rogge, R. D. (2014). Caring about Carelessness: Participant Inattention and Its Effects on Research. *Journal of Research in Personality*, *48*(1), 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Meade, A. W., & Craig, S. B. (2012). Identifying Careless Responses in Survey Data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Moser, C. A., & Kalton, G. (1972). *Survey Methods in Social Investigation* (2nd ed.). Heinemann.

Nederhof, A. J. (1985). Methods of Coping With Social Desirability Bias: A Review. *European Journal of Social Psychology*, *15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting Careless Respondents in Web-based Questionnaires: Which Method to Use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Polit, D. F. (2014). Getting Serious About Test-Retest Reliability: A Critique of Retest Research and Some Recommendations. *Quality of Life Research*, *23*, 1713–1720. https://doi.org/10.1007/s11136-014-0632-9

Revilla, M., & Höhne, J. K. (2021). Repeatedly Measuring Political Interest: Can we Reduce Respondent' Recall Ability and Memory Effects in Surveys Using Memory Interference Tasks? *International Journal of Public Opinion Research*, *33*(3), 678–689. https://doi.org/10.1093/ijpor/edaa035

Saris, W. E., & Gallhofer, I. N. (2014). *Design, Evaluation, and Analysis of Questionnaires for Survey Research* (2nd ed.). John Wiley and Sons.

Schonlau, M., & Toepoel, V. (2015). Straightlining in Web Survey Panels over Time. *Survey Research Methods*, *9*(2), 125–137. https://doi.org/10.18148/srm/2015.v9i2.6128

Schuman, H., Kalton, G., & Ludwig, J. (1983). Context and Contiguity in Survey Questionnaires. *Public Opinion Quarterly*, *47*(1), 112–115. https://doi.org/10.1086/268771

Schuman, H., & Presser, S. (1981). *Questions and Answers in Attitude Surveys. Experiments on Question Form, Wording, and Context*. Academic Press.

Schwarz, H., Revilla, M., & Weber, W. (2020). Memory Effects in Repeated Survey Questions. Reviving the Empirical Investigation of the Independent Measurements Assumption. *Survey Research Methods*, *14*(3), 325–344. https://doi.org/10.18148/srm/2020.v14i3.7579

Schwarz, N., & Strack, F. (1999). Reports of Subjective Well-Being: Judgmental Processes and Their Methodological Implications. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-Being: The Foundations of Hedonic Psychology* (pp. 61–84). Russel Sage Foundation.

Strack, F., & Martin, L. L. (1987). Thinking, Judging, and Communicating: A Process Account of Context Effects in Attitude Surveys. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social Information Processing and Survey Methodology. Recent Research in Psychology*. (pp. 123–148). Springer.

Sudman, S., & Bradburn, N. M. (1974). *Response Effects in Surveys. A Review and Synthesis*. Aldine Publishing Company.

Tourangeau, R. (1987). Attitude Measurement: A Cognitive Perspective. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social Information Processing and Survey Methodology. Recent Research in Psychology*. (pp. 149–162). Springer.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin*, *103*(3), 299–314. https://doi.org/10.1037/0033-2909.103.3.299

Tourangeau, R., Rasinski, K. A., Bradburn, N., & D'Andrade, R. (1989). Belief Accessibility and Context Effects in Attitude Measurement. *Journal of Experimental Social Psychology, 25*(5), 401–421. https://doi.org/10.1016/0022-1031(89)90030-9

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

van Meurs, A., & Saris, W. E. (1990). Memory Effects in MTMM Studies. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait Multimethod Studies* (pp. 134–146). North Holland.

Yan, T., Fricker, S., & Tsai, S. (2020). Response Burden: What Is It and What Predicts It? In P. C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G. B. Willis, & A. Wilmot (Eds.), *Advances in Questionnaire Design, Development, Evaluation and Testing* (pp. 193–212). John Wiley & Sons. https://doi.org/10.1002/9781119263685.ch8

Zhang, C., & Conrad, F. G. (2014). Speeding in Web Surveys: The Tendency To Answer Very Fast and Its Association With Straightlining. *Survey Research Methods*, *8*(2), 127–135.

2. Investigating Respondent Attention to Experimental Text Lengths[5]

Abstract

Whether respondents pay adequate attention to a questionnaire has long been of concern to survey researchers. In this study, we measure respondents' attention with an instruction manipulation check. We investigate which respondents read question texts of experimentally varied lengths and which become inattentive. Our study was implemented in a probability-based online panel of the general population and thus carries external validity. We find that respondent attention is closely linked to text length. Individual response speed is strongly correlated with respondent attention, but a fixed cutoff time is unsuitable as a standalone attention indicator. Differing levels of attention are also correlated with respondents' age, gender, education, panel experience, and the device used to complete the survey. Removal of inattentive respondents is thus likely to result in a biased remaining sample. Instead, questions should be curtailed to encourage respondents of different backgrounds and abilities to read them attentively and provide optimized answers.

Keywords

Respondent attention, instruction manipulation check, online survey, experimental treatment, text length

Acknowledgements

The authors gratefully acknowledge the opportunity to include the experiment presented in this paper in wave 38 of the German Internet Panel (GIP, DOI: 10.4232/1.1339, Blom et al., 2019). GIP data are available to the scientific community from the GIP data archive as

[5] This chapter is joint work with Annelies G. Blom. It is at the time of writing under peer review: Rettig, T. & Blom, A.G. (under review). Investigating Respondent Attention to Experimental Text Lengths.

Scientific Use Files at

https://dbk.gesis.org/dbksearch/GDesc2.asp?no=0109&tab=&ll=10&notabs=1&db=E.

In addition, our analyses used wave 38 paradata (device and response time). These variables are available at the GIP On-Site Data Access (ODA) facilities. A study description of the GIP can be found in Blom et al. (2015).

The authors also gratefully acknowledge the fruitful discussions with Dr. Jan Karem Höhne during the development stage of the project.

2.1 Introduction

Whether survey respondents answer a question with as much effort, care, and attention as the task at hand requires, has long been a concern among survey researchers (see, for example, Krosnick, 1991). Ideally, at each question respondents undergo a four-step cognitive response process (Strack & Martin, 1987; Tourangeau et al., 2000): Comprehending the question and understanding which information is being asked for, retrieving relevant information from memory, forming a judgement based on the retrieved information, and selecting the appropriate response. As Krosnick (1991) argues, not all respondents necessarily have the motivation to fully perform this cognitive process for each item of a survey and may instead use shortcuts to reach an acceptable response with lower cognitive effort (i.e., they "satisfice"). Krosnick distinguishes weak and strong satisficing. In weak satisficing, respondents undergo all four steps of the response process less thoroughly than would be required to reach an optimal response. In strong satisficing, however, respondents may omit retrieving information as well as forming a judgement. Instead, they skip directly from a superficial comprehension of the question to selecting a response that will appear reasonable while not reflecting any actual view that the respondent holds (Krosnick, 1991).

Self-administered surveys, like online surveys, allow for an even more extreme form of satisficing, because respondents can omit the first step of the cognitive process by selecting a response option without having read the question, instructions, text stimuli, or the response options (Anduiza & Galais, 2017; Shamon & Berning, 2020). Consequently, this disruption of the first step makes it impossible for respondents to undergo the other three steps. Without knowing which information a question is asking for, respondents cannot retrieve any relevant information, form a judgement based on this information, or select a response that corresponds to this judgement.

Clearly, such skipping behavior cannot result in a meaningful response. Consequently, researchers have devised both direct and indirect measurements to detect survey respondents who pay insufficient attention (overviews can be found in Curran, 2016 and Huang et al., 2012). The routine use of one or more such measurements is a common suggestion in the literature (Abbey & Meloy, 2017; Curran, 2016; Meade & Craig, 2012; Shamon & Berning, 2020).

2.1.1 Indicators of respondent attention

A popular ad hoc indicator of attention, especially when no other dedicated attention indicator is available, is to look at the time it took respondents to fill out a questionnaire and exclude those with unreasonably fast response times, i.e. speeders (Aust et al., 2013; Curran, 2016; Meade & Craig, 2012). This practice follows the idea that a serious effort to read and carefully answer all items in a survey should take respondents a certain minimum amount of time and that saving time (and effort) is a primary motive of satisficing (Curran, 2016). However, while fast response times are a good indication of inattention, a cutoff value to distinguish attentive from inattentive respondents has proven difficult to determine in practice (Aust et al., 2013; Curran, 2016; Meade & Craig, 2012; Niessen et al., 2016).

Direct measurements of assessing respondent attention by using specialized items specifically created for this purpose include instructed response items, so-called “bogus” items, and instruction manipulation checks. Instructed response items directly ask respondents to pick a specific response option (such as “If you read this, check ‘completely disagree’”), making the instructed response the “correct” option to pick for attentive respondents and flagging the selection of any other response option as failing to pay attention (Curran, 2016; Huang et al., 2012; Meade & Craig, 2012). Bogus items work similar, but instead of asking respondents to select a specific option, they have only one obvious correct response. For example, a

respondent who agrees with the statement "I was born on the 30th of February" has apparently not read the statement (Curran, 2016; Meade & Craig, 2012; Paolacci et al., 2010).

Whereas in theory both measures lead to right or wrong answers that allow distinguishing attentive from inattentive respondents, they can also produce both false negatives (inattentive respondents picking the correct option by chance) and false positives, especially when respondents overthink the meaning of the item or take an impossible statement as hyperbole rather than literally (Curran & Hauser, 2019). Furthermore, in many surveys, respondents are explicitly assured that researchers are interested in their opinions and that there are no right or wrong answers. The sudden introduction of items with right and wrong answers runs counter to this assertion. Finally, as Goodman et al. (2013) noted, unusual questions and response options bear the risk of drawing attention and thus alerting respondents who were just skimming over the questionnaire.

In contrast, instruction manipulation checks are inserted in the instructions preceding the actual survey question. The checks instruct respondents to perform an action that would not normally be expected of them, such as clicking on a logo or writing "I read the instructions" into the answer field (Hauser et al., 2016; Hauser & Schwarz, 2016; Oppenheimer et al., 2009). Asking for an unusual action in this way eliminates the possibility of respondents randomly passing the attention check without having read the instructions. One might retort that skipping such instruction texts is a common practice among respondents who feel that they do not need any additional help to answer the question. Thus, it would not necessarily mean that the response itself was not thought through and honest (Maniaci & Rogge 2014). However, questionnaire designers usually consider instructions a vital part of the question and, thus, a full response necessitates reading, comprehending and carefully considering all information provided (see, e.g., Tourangeau et al., 2000).

2.1.2 Treating inattentive respondents

Respondents who answer a questionnaire with insufficient attention have been shown to provide lower quality data (Aust et al., 2013; Gummer et al., 2021; Huang et al., 2012; Peer et al., 2014), introduce error into the data (Meade & Craig, 2012), reduce statistical power (Aust et al., 2013), and, as a consequence, mask effects present in attentive respondents (Goodman et al., 2013; Maniaci & Rogge, 2014). Consequently, it is now common practice to remove inattentive respondents from analyses (Aronow et al., 2019; Berinsky et al., 2014; Meade & Craig, 2012; Paolacci et al., 2010), especially when post-hoc identification methods are used after the end of fieldwork. Such removal can improve data quality (Aust et al., 2013; Shamon & Berning, 2020), improve scale measurement (Huang et al., 2012), and increase statistical power (Maniaci & Rogge, 2014). However, removing inattentive respondents may also introduce or worsen existing biases in the sample composition (Anduiza & Galais, 2017; Aronow et al., 2019; Berinsky et al., 2014; Oppenheimer et al., 2009).

Attention checks also give researchers the option to intervene during the interview. Respondents who failed an attention check may either be allowed to continue without intervention or be made aware of their failure to pass an attention test. Most researchers opt for the passive approach of letting inattentive respondents continue (and later remove them). A few studies have experimented with making those who failed an attention check retry to pass (Hauser et al., 2016; Hauser & Schwarz, 2015; Oppenheimer et al., 2009). Researchers' reservations with the latter approach may stem from concerns of angering respondents who failed the attention check by making them feel caught (Aust et al., 2013; Curran, 2016; Gummer et al., 2021; Hauser et al., 2016; Peer et al., 2014). However, hard evidence supporting or alleviating this concern is sparse beyond Oppenheimer et al. (2009) noting that they did not receive any complaints.

Some authors suggest adding a warning about the importance of good data quality and the presence of attention tests throughout the survey. This procedure seems to increase attention but may also increase social desirability bias, decrease attitudes towards the survey, and the effects of such a warning may also depend on its wording (Clifford & Jerit, 2015; Meade & Craig, 2012). Furthermore, it is unclear whether the observed increase in attention is caused by respondents paying more attention to the survey questions or whether such a warning just puts them on alert to look out for attention checks.

2.1.3 Prior findings on respondent attention

The practice of removing respondents who were identified as not paying sufficient attention has sparked two main concerns about its effects: on the sample size and on the sample composition when certain groups of respondents are removed more frequently than others.

The size of the problem varies widely across studies, samples, types of attention tests, and cognitive effort required to pass these tests (Anduiza & Galais, 2017; Hauser & Schwarz, 2016; Mancosu et al., 2019; Meade & Craig, 2012). In some settings, a large majority of respondents passed the attention checks and only a small percentage was flagged for inattention (Gummer et al., 2021; Hauser et al., 2016; Hauser & Schwarz, 2015; Johnson, 2005; Paolacci et al., 2010). In contrast, other studies have reported higher rates of attention check failures up to the point where a majority of respondents was flagged as inattentive (Hauser & Schwarz, 2016; Liu & Wronski, 2018; Shamon & Berning, 2020). Ex ante predicting how many respondents will be identified as inattentive in a given survey is thus difficult.

The effect on the sample composition of removing inattentives equally varies across studies. Whereas some report that male respondents tend to be less attentive than female respondents (Berinsky et al., 2014; Maniaci & Rogge, 2014), others find no difference in attention between genders (Gummer et al., 2021; Mancosu et al., 2019; Oppenheimer et al., 2009).

There seems to be some evidence that younger respondents are less attentive than older respondents (Anduiza & Galais, 2017; Berinsky et al., 2014; Gummer et al., 2021; Maniaci & Rogge, 2014), but again, not all studies find this effect (Mancosu et al., 2019; Oppenheimer et al., 2009). Furthermore, lower attention may also correlate with lower education (Anduiza & Galais, 2017; Gummer et al., 2021; Mancosu et al., 2019; Maniaci & Rogge, 2014), although Berinsky et al. (2014) barely find any difference. Gummer et al. (2021) report slightly higher attention from respondents with prior survey experience, which the authors attribute to higher participation ability. This does, however, seem to contradict other research indicating that more experienced panel respondents tend to respond less carefully and engage in more satisficing behavior (Schonlau & Toepoel, 2015; Toepoel et al., 2008). Finally, Gummer et al. (2021) report no difference in attention across the devices respondents used to complete the survey.

It should be noted that these results are not only very mixed, they also near exclusively stem from studies that were conducted in nonprobability samples or samples of specific population subgroups, such as respondents from commercial access panels or student samples. This is notable for two reasons: First, both data quality and sample accuracy tend to be lower in nonprobability samples and can vary widely and unpredictably across different nonprobability samples (Brüggen et al., 2016; Cornesse et al., 2020; Cornesse & Blom, 2020; Pasek & Krosnick, 2020). This may explain the mixed findings on respondent attention described above. Second, probability-based recruitments of respondents are generally associated with considerably higher costs and efforts (Baker et al., 2010; Brüggen et al., 2016; Pasek & Krosnick, 2020; Sakshaug et al., 2019; Wiśniowski et al., 2020). For researchers aiming to infer to the general population and thus working with a high-quality probability-based sample, removing inattentive respondents may thus be prohibitively expensive. Both the loss of any number of respondents due to their exclusion following attention checks and the resulting introduction of bias into the sample are undesirable outcomes. In addition, the concerns about

losing respondents who were angered by attention checks outlined above may contribute to the scarcity of attention research in probability-based samples.

2.2 Research questions

We have revealed several gaps in the literature that our study aims to fill. So far, direct attention measurements have generally been applied to individual questions or their instruction text. We extend this approach by introducing an instruction manipulation check into a treatment text to accurately measure the prevalence of inattention.

Q1: Which proportion of the population skips reading a treatment text in an online survey?

In addition, an experimental variation of the text length allows us to investigate whether and to what extent respondent attention to the text treatment is dependent on its length.

Q2: How much does respondent attention vary with text length?

We investigate, which sociodemographic groups (gender, age, education, and occupation) are at risk of being disproportionately excluded from analyses if they were removed following a failed attention check.

Q3: Which sociodemographic characteristics correlate with respondent inattention?

In addition, we investigate the relationship between respondent attention and participation characteristics, in particular their level of panel experience, the device used to complete the survey, and how early in the field and at which time of the day respondents participated.

Q4: Are other participation indicators correlated with skipping a treatment text?

Furthermore, our experimental design in combination with a wealth of background characteristics and paradata also enables us to observe whether an increase in text length disproportionately causes inattention among specific groups of respondents. We therefore investigate potential interactions between text length and respondent characteristics.

Q5: Does an increase in text length disproportionately cause inattention among certain groups of respondents?

Finally, very short completion times are commonly used as an indicator that respondents did not answer a survey with the required care and attention. We, therefore, investigate how this indirect detection method compares to our direct attention measurement. More specifically, we investigate the relationship between the time respondents spend on the survey page that contains the treatment text and the result of the attention check. To be a useful measure of respondent attention, response time should both correlate strongly with the result of the attention check and allow researchers to set a sensible cutoff that distinguishes between attentive and inattentive respondents.

Q6: Is response time a suitable predictor of respondent attention?

The context of a probability-based online panel of the general population allows us to generalize our findings beyond this specific online survey.

2.3 Data and method

To answer our research questions, we implemented a survey experiment in the November 2018 wave of the German Internet Panel (GIP; Blom et al., 2019). The GIP is a probability-based online panel of the general population of Germany (see Blom et al., 2015, 2017). GIP

panelists are surveyed online bimonthly with a questionnaire of 20 to 25 minutes. Recruitments into the panel were conducted in 2012, 2014, and 2018 among the general population aged 16 to 75 years living in private households in Germany at the time of recruitment. The November 2018 wave constitutes wave 38 of the GIP, but was the first regular wave for the newly recruited 2018 sample. Placing our experiment in this particular wave thus allows us to make comparisons between freshly recruited panelists and experienced panelists from the 2012 and 2014 recruitments.

2.3.1 Experimental design

Our attention check experiment was inserted in the middle of the questionnaire within a question on vote choice in a hypothetical referendum on Germany remaining in or leaving the European Union. As similar topics are frequently surveyed in the GIP, this question was unlikely to draw any unusual attention. Within the text leading up to the vote choice question, we added an instruction asking respondents not to answer the question presented, but instead to click on the GIP logo in the top left corner of the page. This logo is shown on every page of each GIP questionnaire but usually has no function. Thus, we can be very certain that no respondent randomly clicked on the logo without having read our prompt to do so.

To maximize compliance and avoid angering respondents, the instruction was carefully worded, such as to not give respondents the feeling that they are being monitored. The instruction thus read:

> "It is not always easy to stay concentrated throughout. To improve our study, we would like to know whether our participants carefully read the texts. Please do not answer the question below on Germany's membership in the EU. Instead, to continue please click on the logo of 'Society in Change' in the top left corner. This way, you show us that you have read this text." (survey fielded in German, own translation).

To investigate respondent attention to texts of different lengths, we designed four versions of the substantive question. The first version only displayed the instruction itself right above the survey question (see Figure 2.1, panel 1). In the other three versions, the instruction was placed within a text with one, two, or four paragraphs (see Figure 2.1, panels 2–4). To minimize the chance of respondents becoming alerted to the instruction while skimming over the text or after reading it partially, the instruction was always placed before the final sentence of the last paragraph of each text. Respondents were randomly allocated to one of the attention check treatments. The length of the texts in the four versions ranged from 58 words to 347 words (not counting the question and response options).

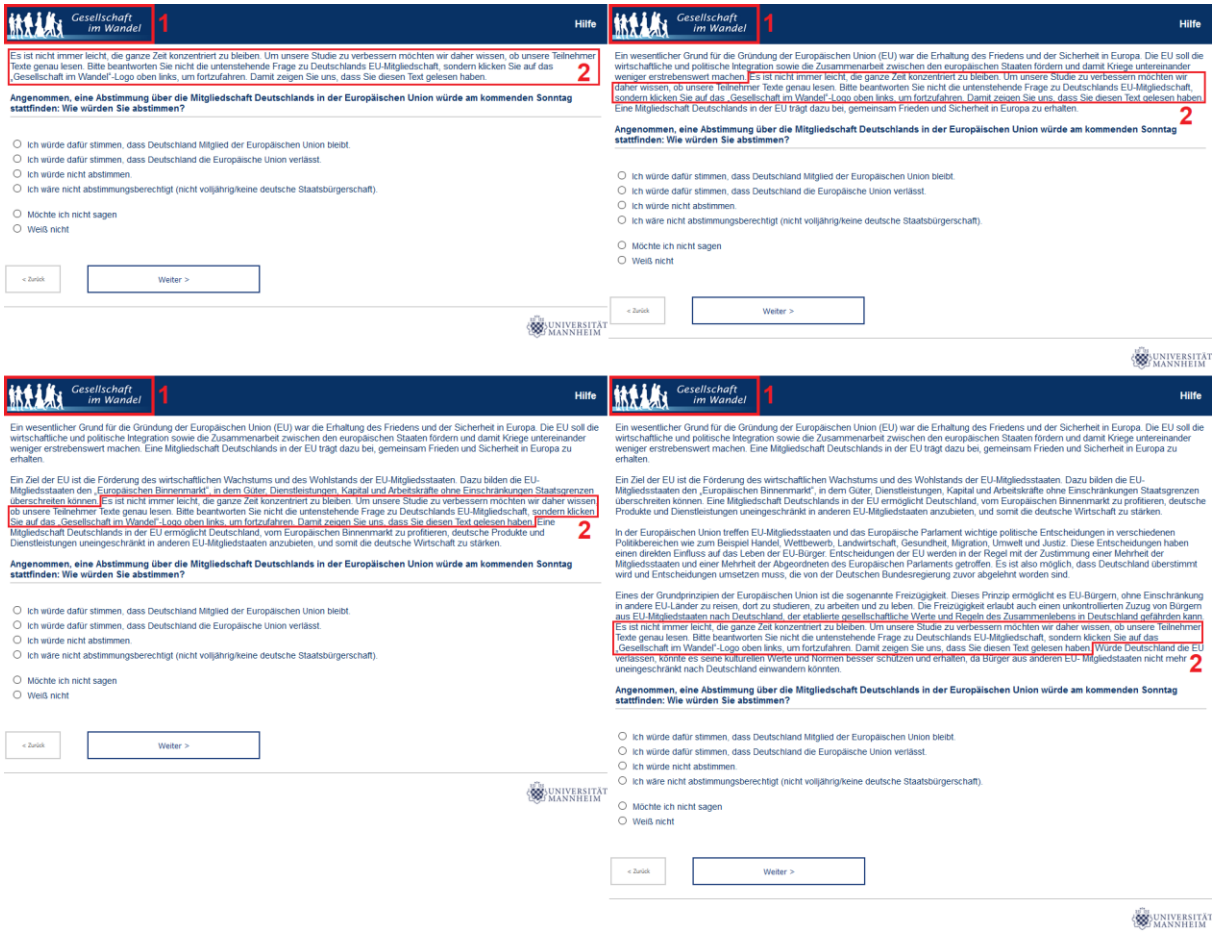

Figure 2.1. The four versions of the attention check with the logo (1) and instruction (2).

Note. Red boxes and numbers were not displayed on the original questionnaire.
[1] Respondents were instructed to click on this logo to continue.
[2] The position of the instruction text on the question page.

Both clicking on the GIP logo (and thus passing the attention check) and clicking on the continue button (and thus failing the attention check) took respondents to the next survey page. Respondents continued with the survey independently of whether they had passed or failed the attention check. Those who failed were not made aware of this. There was no "back" button on the subsequent page.

2.3.2 Sample description

In total, 4,294 respondents participated in the November 2018 wave of the GIP. Out of these, 3.1% (134) were not included in our experiment because the necessary JavaScript was not enabled on their device, 0.9% (38) broke off the survey before answering the attention check, and 1.9% (83) were missing other information used in our analyses. We also excluded 0.3% (13 respondents) who remained on the experiment's questionnaire page for an excessively long time (>20 minutes). Therefore, 4,026 respondents were included in our experiment and provided information on all variables used in our analyses.

Our sample had a median age of 51 years and 48.2% were female. 15.5% had low education ("Hauptschule" or no degree), 28.9% medium-low ("Realschule" or equivalent), 22.4% medium-high ("Fachhochschulreife", "Abitur" or equivalent with no college degree) and 33.2% high education (college degree or higher). 46.5% worked full-time, 17.4% part-time, and 8.9% were unemployed or not in the labor force, 19.6% were retired, and 7.7% were in education or voluntary (civil or military) service. The newly recruited 2018 respondents accounted for 43.9% of the sample. 21.7% of respondents used smartphones to complete the survey.

The questionnaire was open for completion from November 1 to November 30, 2018. 34.7% of the sample participated by the end of day 3, 55.9% by the end of the first week, and 76.5% by the end of the second week. Finally, 25.6% participated in the morning (6am–noon),

38.7% in the afternoon (noon–6pm), 28.4% in the evening (6–10pm), and 7.3% at night (10pm–6am).

A random 25.3% of respondents (1,018) received the instruction only version of the attention check, 25.1% (1,009) received one paragraph of text, 24.8% (999) two paragraphs, and 24.8% (1,000) four paragraphs. The random assignment to treatment delivered a uniform distribution of respondents with different backgrounds across treatment groups (no significant differences in gender, education, occupational status, panel experience or smartphone use). There were also no differences in response date or time of day. Due to chance, younger respondents are slightly overrepresented in the one paragraph condition and slightly underrepresented in the instruction only condition (see Appendix Table A2.1 for distributions of respondents across experimental groups and χ^2-tests).

2.4 Results

2.4.1 Passing the attention check

Overall, 60.8% of respondents complied with the instruction to click on the logo and thus passed the attention check. This passing rate varied greatly across experimental groups, ranging from 79.4% for the instruction only condition and falling with each increase in text length to 41.1% for the four paragraphs condition (Figure 2.2). A logistic regression model of the association between text length and likelihood of passing the attention check (Model 1 in Table 2.1) confirms this significant correlation between longer text conditions and lower likelihood of respondents passing the attention check. Computing separate models for each of the text lengths as the respective reference category (not shown) revealed significant differences across comparisons of all pairs of text lengths. This indicates that each increase in text length significantly decreases the proportion of respondents passing the attention check.

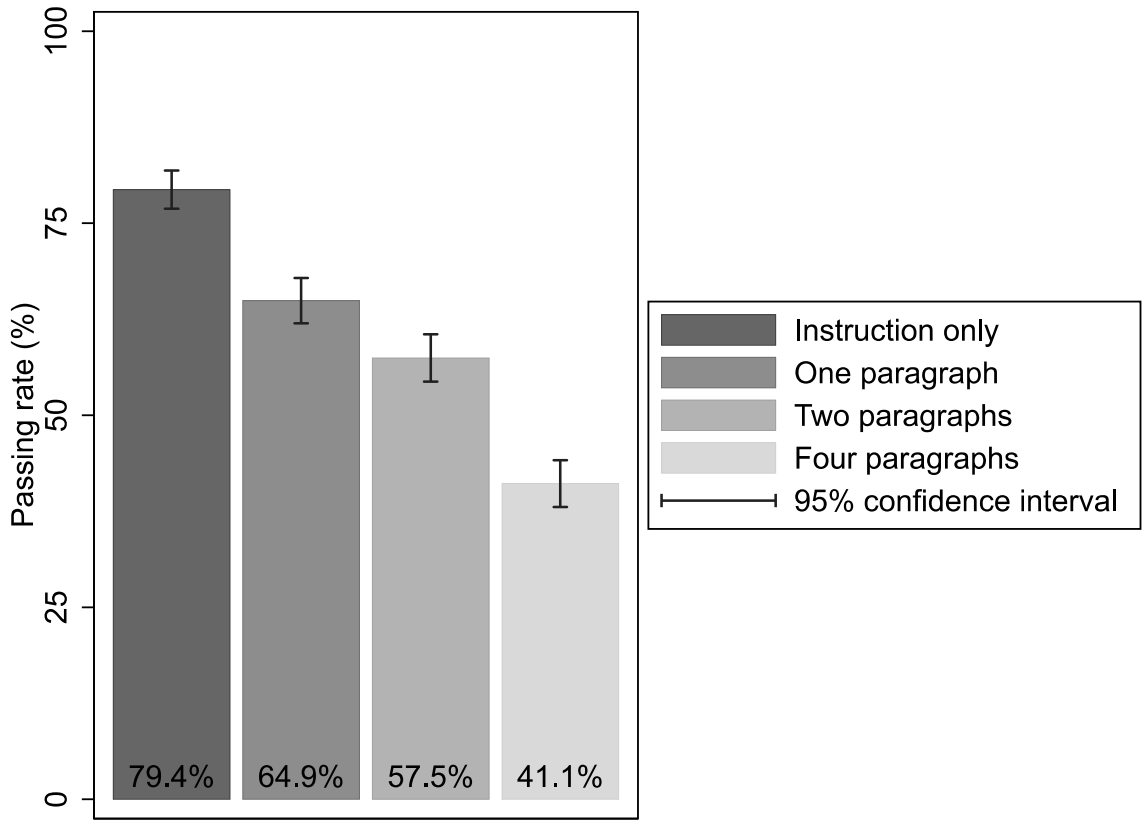

Figure 2.2. Passing rates for the attention check across the four text length conditions.

Note. Lower and upper boundaries of the displayed 95% confidence intervals:
instruction only: [76.9%, 81.9%]; one paragraph: [62.0%, 67.9%]; two paragraphs: [54.4%, 60.5%]; four paragraphs: [38.0%, 44.2%].

To further investigate which groups of respondents were more likely to pass the attention check (or which respondents were, in turn, more likely to skip reading the text), we computed three additional logistic regression models (Models 2–4 in Table 2.1). Model 2 additionally contains respondents' sociodemographics (age, gender, education, and occupation) as predictors of passing the attention check, as well as the text length as a control. Model 3 further expands the model with participation characteristics (panel experience, smartphone use, as well as the day and the time of day of participation). Finally, in Model 4 we add interaction effects of sociodemographics (age, gender, and education), panel experience and smartphone use with text length to investigate whether these groups are differently affected by increased inattention due to increased text length.

Table 2.1. Logistic regression models of passing the attention check.

	Model 1 OR	**Model 2** OR	**Model 3** OR	**Model 4** OR
Text length (ref.: Instruction only)				
One paragraph	.481***	.475***	.471***	.567
Two paragraphs	.351***	.343***	.341***	.558
Four paragraphs	.181***	.172***	.167***	.141***
Age (ref.: <44 years)				
44–58 years		1.466***	1.699***	1.880**
>58 years		1.368**	1.643***	1.441
Female		1.425***	1.383***	1.186
Education (ref.: low)				
Medium-low		1.289*	1.282*	1.535
Medium-high		1.807***	1.795***	2.604***
High		2.035***	2.127***	2.671***
Occupation (ref.: full-time work)				
Part-time work		1.192	1.192	1.203
Unemployed or not in labor force		1.192	1.181	1.199
Retired		1.032	1.034	1.040
In education or voluntary service		1.175	1.098	1.116
New recruitment sample			1.604***	1.612**
Smartphone respondent			1.532***	1.985**
Field time (ref.: day 1)				
Days 2–3			1.154	1.172
Days 4–7			1.420*	1.433*
Days 8–14			1.406*	1.426*
After day 14			1.138	1.164
Day time (ref.: morning)				
Afternoon			1.274**	1.263**
Evening			1.273*	1.285*
Night			1.398*	1.392*
Age * text length				
44–58 years * one paragraph				.827
44–58 years * two paragraphs				.891
44–58 years * four paragraphs				.944
>58 years * one paragraph				1.001
>58 years * two paragraph				1.078
>58 years * four paragraph				1.569
Female * text length				
One paragraph				1.207
Two paragraphs				1.093
Four paragraphs				1.353
Education * text length				
Medium-low * one paragraph				.707
Medium-low * two paragraphs				.693
Medium-low * four paragraphs				1.081
Medium-high * one paragraph				.642
Medium-high * two paragraphs				.480*
Medium-high * four paragraphs				.892

Table 2.1 (continued).

Education * text length				
High * one paragraph				.978
High * two paragraphs				.505*
High * four paragraphs				.965
New recruitment sample * text length				
One paragraph				1.286
Two paragraphs				.958
Four paragraphs				.839
Smartphone respondent * text length				
One paragraph				.563*
Two paragraphs				.860
Four paragraphs				.818
Constant	3.848***	1.618***	.745	.623
Observations	4,026	4,026	4,026	4,026
Pseudo-$R^2_{McKelvey \& Zavoina}$	.103	.137	.165	.178
AIC	5,069	4,984	4,914	4,929
BIC	5,094	5,072	5,059	5,225

Note. OR = odds ratios. * p < 0.05, ** p < 0.01, *** p < 0.001

First, we have a look at the sociodemographics of respondents who passed and who failed the attention check (Model 2). The youngest group is less likely to pass the attention check than the middle and older age groups. Changing the reference category (not shown) reveals that the middle and older groups do not significantly differ in their passing rates. Female respondents show a significantly higher chance to pass the attention check than male respondents. Attention also differs across levels of education: Respondents with medium-low, medium-high, and high education are significantly more likely to pass the attention check than those with a low level of education. Changing the reference category (not shown) reveals that respondents with medium-high and high education are also significantly more likely to pass than those with a medium-low level of education. However, respondents with medium-high and high education do not significantly differ. Finally, comparing respondents with different occupation status does not reveal any significant differences.

The effects observed in Model 2 remain stable when adding participation indicators (panel experience, smartphone use, and the day and time of participation) in Model 3. We find that new GIP recruits are significantly more likely to pass the attention check than experienced

respondents who have been panel members for four to six years. Curiously, respondents who completed the survey on a smartphone were significantly more likely to pass the attention check than those who used computers or tablets. It should be noted that in the displayed model we find this effect while adding (and thereby controlling for) respondents' age in the same model. That is, smartphone respondents are more attentive than non-smartphone respondents of the same age. However, smartphone use was not evenly distributed across age groups. In the younger age group (under 44 years old), which we found to be less attentive than the older age groups, 42.1% of respondents completed the survey on a smartphone. This proportion is only 14.9% in the middle- and 5.0% in the older age group respectively. A sensitivity analysis in a reduced model (question length and smartphone use as the only correlates) confirms the positive effect of smartphone completion on the passing rates (Appendix Table A2.2).

The point in time during the 30-day field period that respondents participated in the survey did not consistently significantly correlate with attention. However, both early and late responders seem to be less likely to pass the attention check than those in-between. Finally, we find that respondents who completed the questionnaire in the afternoon (from noon to before 6pm), in the evening (from 6pm to before 10pm), or at night (from 10pm to before 6am) were significantly more likely to pass the attention check than those who participated in the morning (from 6am to before noon). Selecting a different reference category did not reveal any further significant differences.

Adding interaction terms for age, gender, education, panel experience and smartphone use with text length (Model 4) did not reveal any consistent interaction effects. The main effects from previous models, however, remained substantially unchanged with all of them pointing in the same direction as before but some losing statistical significance. Computing individual models with only one of these interaction terms each (not shown) did not result in any

substantially different findings. In addition, the AIC and BIC values of these models did not indicate that the inclusion of any interaction terms resulted in an improved model fit over Model 3.

2.4.2 Response time as an indicator of attention

In lack of an attention test, researchers often resort to identifying inattentive respondents by setting a fixed response time cutoff. We evaluate this method by comparing the response times between respondents who passed and those who failed the attention check across the four text lengths. Figure 2.3 displays density curves for the response times of respondents who passed and those who failed the attention check separately for each of the four text lengths.

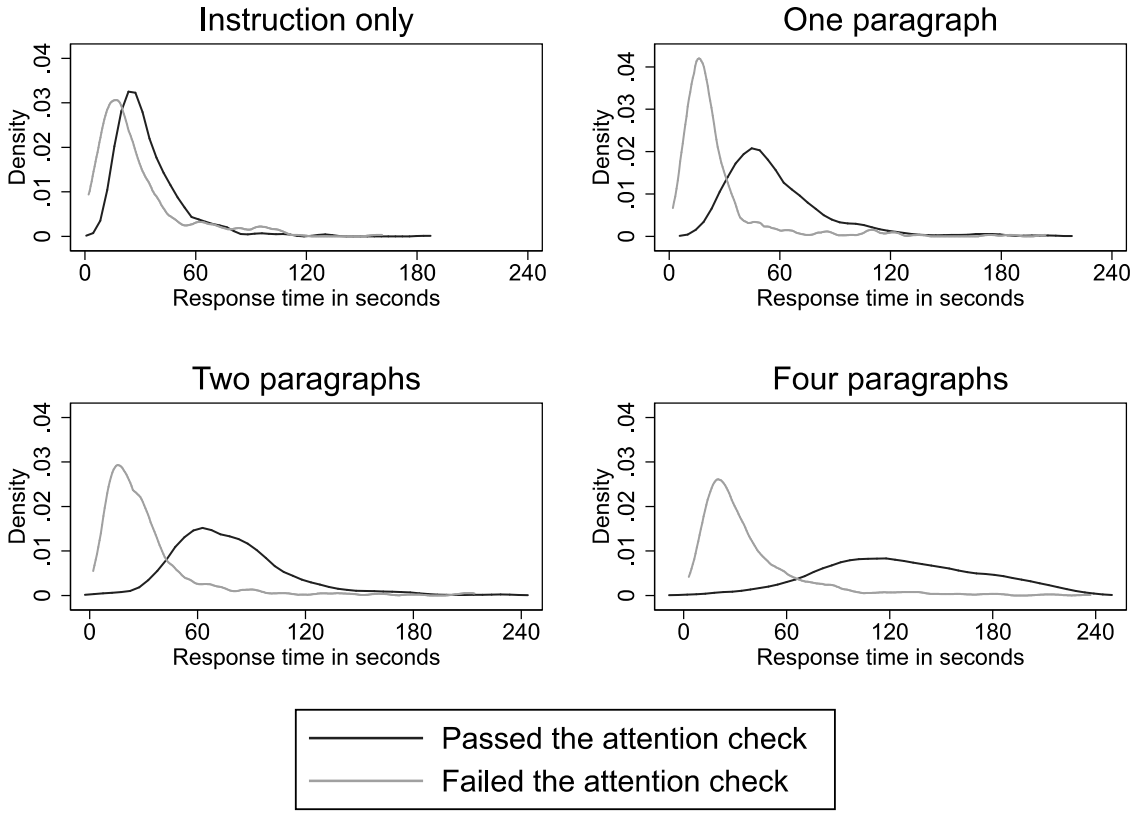

Figure 2.3. Density curves of response times in the attention check.

Note. Response times limited to 240 seconds (4 minutes) for display purposes.

With increasing text length, the mean response time of respondents who passed the attention check increases, as does the variance. As would be expected if failing the attention check equates to skipping the text, the mean and variance of the response times of respondents who failed the attention check are hardly affected by text length. We also see that there is considerable overlap of the two curves, especially in the shorter text conditions. Using a fixed response time cutoff to identify inattentive respondents would thus bear the risk of either not identifying a portion of the slower inattentive respondents, incorrectly identifying fast readers as inattentive, or both, especially in shorter texts.

This observation is confirmed by comparing the correlation of response time with the results from the attention check for each text length condition (Table 2.2)[6]. The correlation of response time with attention is stronger for longer text conditions. Response time and attention are only weakly correlated for the instruction only condition. However, for the longer texts, response time and attention are moderately to strongly correlated. For the longest text condition, whether respondents read the text accounts for about half of the variance in their (logarithmized) response times.

Table 2.2. Pearson's correlations of logarithmized response time with passing the attention check by text length.

	r	R^2	N
Overall	.468***	.219	4,026
Instruction only	.248***	.062	1,018
One paragraph	.633***	.401	1,009
Two paragraphs	.671***	.450	999
Four paragraphs	.709***	.502	1,000

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In addition, to investigate whether a misidentification of attentive or inattentive respondents by using a fixed time cutoff may disproportionately affect certain groups of respondents, we

[6] Given the skewed nature of the response time distribution, we use the logarithmized response time in all further analyses.

computed two linear regression models to identify correlates of response time for respondents who passed and those who failed the attention check separately while controlling for text length (Table 2.3).

Table 2.3. Linear regression models of logarithmized response time across respondents who passed and failed the attention check.

	Passed	**Failed**
	b	b
Text length (ref.: Instruction only)		
One paragraph	.574***	- .046
Two paragraphs	.926***	.156*
Four paragraphs	1.430***	.391***
Age (ref.: <44 years)		
44–58 years	.054*	.231***
>58 years	.152***	.380***
Female	- .030	- .057
Education (ref.: low)		
Medium-low	- .059	- .069
Medium-high	- .096**	- .027
High	- .157***	- .121*
New recruitment sample	.090***	.269***
Smartphone respondent	.025	.234***
Constant	- .693***	- 1.310***
Observations	2,448	1,578
$R^2_{Adj.}$	.531	.110

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Among respondents who passed the attention check, we find that younger respondents, respondents with medium-high or high education, and experienced respondents tended to have shorter response times and may thus be misidentified as inattentive if a response time cutoff is used as a proxy for attention. Among respondents who failed the attention check, the middle and older age groups, newly recruited respondents, and those completing the survey on a smartphone tended to have longer response times and may in turn be misidentified as attentive. Respondents with low education also had significantly longer response times than those with high education and may be misidentified as attentive (however, the medium-low and medium-high groups did not significantly differ from respondents with high or low education).

2.5 Summary

This paper addresses respondent attention to survey questions in a probability sample of the general population. Six research questions guide our analyses. We discuss the results of each in turn:

Q1: Which proportion of the population skips reading a treatment text in an online survey?

To answer this question, we implemented an attention experiment in the middle of a 20–25 minutes long survey wave of a probability-based online panel. In the text leading up to a survey question, we instructed respondents to click the survey's logo instead of answering the question and experimentally varied the length of this text. We recorded whether respondents complied with this instruction to measure their attention to the text. We found that overall, 60.8% of respondents read the text.

Q2: How much does respondent attention vary with text length?

With increasing text length, the proportion of respondents who passed the attention check decreased. Respondents who received the instruction only passed the check most frequently with 79.4%, whereas 64.9%, 57.5%, and 41.1% of respondents who received one, two, and four paragraphs, respectively, passed the check.

These findings have several implications for applied survey research. The association of respondent attention and text length reveals potential for optimizing respondent attention by keeping texts as short as possible. When designing questionnaires, the use of excessively long question texts should thus be avoided, or else a majority of respondents will likely not read the treatment carefully. This may be of particular importance in experiments that test different question stimuli against each other. The effect of the stimuli may be confounded with the

proportion of respondents who read the respective text. As a consequence, a detected treatment effect may only be an artefact of the text length.

Q3: Which sociodemographic characteristics correlate with respondent inattention?

We find that higher rates of failing the attention check in our experiment were associated with younger, male, and lower educated respondents. We find no difference concerning respondents' occupation.

The common practice of excluding inattentive respondents from analyses will therefore likely result in a biased remaining sample. Measures to optimize for respondent attention should thus be preferred over a post-hoc identification and exclusion of respondents. In addition, researchers should be cautious about making comparisons of treatment effects across sociodemographic groups, as the effectiveness of treatments across sociodemographic groups may be confounded with different rates of reading and thus receiving the treatment.

Q4: Are other participation indicators correlated with skipping a treatment text?

In line with previous findings, we find that more experienced respondents are less attentive than newly recruited, inexperienced respondents. In addition, we find that respondents who completed the survey using smartphones were less likely to skip reading a treatment text than those using computers or tablets. This is notable, as concerns that smartphone respondents may provide lower quality data are commonly voiced in the literature (see, e.g., Lugtig & Toepoel, 2016; Struminskaya et al., 2015). Furthermore, while we may think of early responders as more motivated to participate in the survey, we find that both those who participated early during field period (day 1 or days 2–3) and early in the day (i.e., in the morning) were less attentive. This indicates that some early responders may be more

motivated to have the survey “off their desk” instead of taking the time to participate when it is most convenient for them. Late responders who participated after more than two weeks are, however, also associated with lower attention. Survey practitioners should thus consider up to which point leaving the survey open for participants and sending out additional reminders are worthwhile, as fewer and less attentive respondents seem to participate late in the field period.

Q5: Does an increase in text length disproportionately cause inattention among certain groups of respondents?

While we found differences in the general attentiveness across certain groups of respondents, we did not find any consistent interaction effects with text length. Longer texts will generally be read by considerably fewer respondents and some groups of respondents are more likely to skip reading a text than others, but we did not find evidence that these group differences are more pronounced for longer texts. This is good news, because it allows a universal treatment of inattentive respondents.

Q6: Is response time a suitable predictor of respondent attention?

Our findings are twofold: The time respondents spent on the survey page containing the treatment text was moderately to strongly correlated with whether they read it, particularly for the longer text conditions. In this sense, response time is a good predictor of respondent attention.

However, we also find that using a hard response time cutoff value as a proxy for respondent attention bears considerable risk of misidentifying slower inattentive respondents as attentive (false negatives) and fast readers as inattentive (false positives). In addition, we found evidence that such misidentifications would disproportionately affect specific respondent

groups. For instance, we found that younger respondents tended to generally have shorter response times than older respondents (both among respondents who passed and those who failed the attention check). In addition, younger respondents were also more likely to fail the attention check. Removing respondents based on their response time may thus result in an age bias in the remaining sample that is driven by three factors: Genuinely lower attentiveness among younger respondents, the disproportionate misidentification of younger respondents as inattentive, and the disproportionate misidentification of older respondents as attentive. In addition, we found evidence that other biases may be masked. Respondents with high education were more likely to pass the attention check but also tended to have faster response times and were thus at greater risk of being misidentified as inattentive. The result may be a sample that looks unbiased regarding respondents' education levels at the cost of a disproportionate exclusion of attentive respondents with high education. While an exceedingly fast response time may be an indication of respondents who answer a survey without the required care and attention, we advise against the use of a fixed cutoff time as the sole indicator of respondent attention.

2.6 Conclusion

Whenever inferences are made from survey data, researchers need to be certain that respondents paid adequate attention to the questions that they have been asked. By means of an instruction manipulation check experiment our study investigates the prevalence of inattention, the relationship between inattention and text length, sociodemographic and other participation indicator correlates of inattention, as well as the suitability of response times as a proxy for attention. Our study was implemented in a probability-based online panel of the general population and thus carries external validity.

We find that respondent attention to text stimuli decreases with increasing text length. Furthermore, individual response speed is moderately to strongly correlated with respondent attention. However, the use of a fixed cutoff time to separate attentive from inattentive respondents bears a considerable risk of creating both false positives and false negatives. In addition, we find that such misidentification is likely to disproportionately affect respondents of different age groups, education levels, and levels of panel experience. A fixed response time cutoff is therefore not suitable as a standalone indicator of respondent attention. Differing levels of attention are also correlated with respondents' age, gender, education, panel experience, and the device used to complete the survey. An exclusion of inattentive respondents is therefore likely to result in a biased remaining sample. Instead, researchers should optimize questions and treatment texts to encourage respondents of different backgrounds and abilities to read them carefully and provide optimized responses.

In addition to new insights on the topic of respondent attention and practical suggestions for both applied survey research and survey management, our study also points towards areas that require further research. We conducted one specific experiment in a very specific way. Variations on this experiment would give insights on the generalizability of our findings across different settings. For example, we placed the experiment in the middle of a 20–25 minute questionnaire within a question on vote choice in a hypothetical referendum on Germany remaining in or leaving the European Union. Our data thus do not allow us to investigate whether attention levels would have been different earlier or later in the same survey or for a different topic. Moreover, a systematic evaluation of respondents' reactions to the attention check and interventions on inattentive respondents were beyond the scope of our study. Concerns about angering respondents seem to be the key reason for why such interventions are barely explored and used in practice. Therefore, evidence on whether these concerns are justified is sparse, despite its great value in guiding researchers' decision on whether to implement an attention check or not.

Furthermore, the literature on respondent attention is currently divided on whether inattention at one point in time is a good predictor of whether respondents will be inattentive at a later point (Anduiza & Galais, 2017; Berinsky et al., 2014; Gummer et al., 2021). Our study cannot contribute to this important debate, as our experiment has only been implemented once so far. Much value may be generated from insights on how to treat respondents in longitudinal studies, because the decision to remove a long-term panelist from the survey is particularly costly here.

We thus encourage researchers and survey practitioners to implement attention checks in their surveys to gather more evidence on how to best prevent, detect, and handle inattentive respondents.

References

Abbey, J. D., & Meloy, M. G. (2017). Attention by Design: Using Attention Checks to Detect Inattentive Respondents and Improve Data Quality. *Journal of Operations Management*, *53–56*(1), 63–70. https://doi.org/10.1016/j.jom.2017.06.001

Anduiza, E., & Galais, C. (2017). Answering Without Reading: IMCs and Strong Satisficing in Online Surveys. *International Journal of Public Opinion Research*, *29*(3), 497–519. https://doi.org/10.1093/ijpor/edw007

Aronow, P. M., Baron, J., & Pinson, L. (2019). A Note on Dropping Experimental Subjects who Fail a Manipulation Check. *Political Analysis*, *27*(4), 572–589. https://doi.org/10.1017/pan.2019.5

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness Checks Are Useful to Improve Data Validity in Online Research. *Behavior Research Methods*, *45*(2), 527–535. https://doi.org/10.3758/s13428-012-0265-2

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., & Lavrakas, P. J. (2010). AAPOR Report on Online Panels. *Public Opinion Quarterly*, *74*(4), 711–781. https://doi.org/10.1093/poq/nfq048

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-administered Surveys. *American Journal of Political Science*, *58*(3), 739–753. https://doi.org/10.1111/ajps.12081

Blom, A. G., Fikel, M., Friedel, S., Höhne, J. K., Krieger, U., Rettig, T., Wenz, A., & SFB 884 ´Political Economy of Reforms´ Universität Mannheim. (2019). *German Internet Panel, Wave 38 (November 2018)*. GESIS Data Archive, Cologne. ZA6958 Data file Version 1.0.0. https://doi.org/10.4232/1.13391

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting Up an Online Panel Representative of the General Population: The German Internet Panel. *Field Methods*, *27*(4), 391–408. https://doi.org/10.1177/1525822X15574494

Blom, A. G., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U., & Bossert, D. (2017). Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence From the German Internet Panel. *Social Science Computer Review*, *35*(4), 498–520. https://doi.org/10.1177/0894439316651584

Brüggen, E., Van den Brakel, J., & Krosnick, J. A. (2016). Establishing the Accuracy of Online Panels for Survey. *Statistics Netherlands*, *11*(04), 43.

Clifford, S., & Jerit, J. (2015). Do Attempts to Improve Respondent Attention Increase Social Desirability Bias? *Public Opinion Quarterly*, *79*(3), 790–802. https://doi.org/10.1093/poq/nfv027

Cornesse, C., & Blom, A. G. (2020). Response Quality in Nonprobability and Probability-based Online Panels. *Sociological Methods and Research*. https://doi.org/10.1177/0049124120914940

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36. https://doi.org/10.1093/jssam/smz041

Curran, P. G. (2016). Methods for the Detection of Carelessly Invalid Responses in Survey Data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Curran, P. G., & Hauser, K. A. (2019). I'm Paid Biweekly, Just Not by Leprechauns: Evaluating Valid-but-incorrect Response Rates to Attention Check Items. *Journal of Research in Personality*, *82*, 103849. https://doi.org/10.1016/j.jrp.2019.103849

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. https://doi.org/10.1002/bdm.1753

Gummer, T., Roßmann, J., & Silber, H. (2021). Using Instructed Response Items as Attention Checks in Web Surveys: Properties and Implementation. *Sociological Methods and Research*, *50*(1), 238–264. https://doi.org/10.1177/0049124118769083

Hauser, D. J., & Schwarz, N. (2015). It's a Trap! Instructional Manipulation Checks Prompt Systematic Thinking on "Tricky" Tasks. *SAGE Open*, *5*(2). https://doi.org/10.1177/2158244015584617

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks than Do Subject Pool Participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Hauser, D. J., Sunderrajan, A., Natarajan, M., & Schwarz, N. (2016). Prior Exposure to Instructional Manipulation Checks does not Attenuate Survey Context Effects Driven by Satisficing or Gricean Norms. *Methods, Data, Analyses*, *10*(2), 195–220. https://doi.org/10.12758/mda.2016.008

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and Deterring Insufficient Effort Responding to Surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Johnson, J. A. (2005). Ascertaining the Validity of Individual Protocols from Web-based Personality Inventories. *Journal of Research in Personality*, *39*(1 SPEC. ISS.), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Liu, M., & Wronski, L. (2018). Trap Questions in Online Surveys: Results from Three Web Survey Experiments. *International Journal of Market Research*, *60*(1), 32–49. https://doi.org/10.1177/1470785317744856

Lugtig, P., & Toepoel, V. (2016). The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey: Effects on Survey Measurement Error. *Social Science Computer Review*, *34*(1), 78–94. https://doi.org/10.1177/0894439315574248

Mancosu, M., Ladini, R., & Vezzoni, C. (2019). 'Short is Better'. Evaluating the Attentiveness of Online Respondents Through Screener Questions in a Real Survey Environment. *BMS Bulletin of Sociological Methodology/ Bulletin de Methodologie Sociologique*, *141*(1), 30–45. https://doi.org/10.1177/0759106318812788

Maniaci, M. R., & Rogge, R. D. (2014). Caring about Carelessness: Participant Inattention and Its Effects on Research. *Journal of Research in Personality*, *48*(1), 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Meade, A. W., & Craig, S. B. (2012). Identifying Careless Responses in Survey Data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting Careless Respondents in Web-based Questionnaires: Which Method to Use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Pasek, J., & Krosnick, J. A. (2020). Relations Between Variables and Trends over Time in RDD Telephone and Nonprobability Sample Internet Surveys. *Journal of Survey Statistics and Methodology*, *8*(1), 37–61. https://doi.org/10.1093/jssam/smz059

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*(4), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

Sakshaug, J. W., Wiśniowski, A., Ruiz, D. A. P., & Blom, A. G. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics*, *35*(3), 653–681. https://doi.org/10.2478/jos-2019-0027

Schonlau, M., & Toepoel, V. (2015). Straightlining in Web Survey Panels over Time. *Survey Research Methods*, *9*(2), 125–137. https://doi.org/10.18148/srm/2015.v9i2.6128

Shamon, H., & Berning, C. (2020). Attention Check Items and Instructions in Online Surveys with Incentivized and Non-Incentivized Samples: Boon or Bane for Data Quality? *Survey Research Methods*, *14*(1), 55–77. https://doi.org/10.18148/srm/2020.v14i1.7374

Strack, F., & Martin, L. L. (1987). Thinking, Judging, and Communicating: A Process Account of Context Effects in Attitude Surveys. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social Information Processing and Survey Methodology. Recent Research in Psychology.* (pp. 123–148). Springer.

Struminskaya, B., Weyandt, K., & Bosnjak, M. (2015). The Effects of Questionnaire Completion Using Mobile Devices on Data Quality. Evidence from a Probability-based General Population Panel. *Methods, Data, Analyses*, *9*(2), 261–292. https://doi.org/10.4232/1.12245.

Toepoel, V., Das, M., & Van Soest, A. (2008). Effects of Design in Web Surveys: Comparing Trained and Fresh Respondents. *Public Opinion Quarterly, 72*(5), 985–1007. https://doi.org/10.1093/poq/nfn060

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., & Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, *8*(1), 120–147. https://doi.org/10.1093/jssam/smz051

Appendix

Table A2.1. Distributions of variables of interest across experimental groups and χ^2-tests for differences.

		Experimental group				χ^2-test		
	Overall	Instruction	1 paragraph	2 paragraphs	4 paragraphs	χ^2	df	p
Age						15.494	6	.017
<44 years	1,457	331 (22.7%)	404 (27.7%)	366 (25.1%)	356 (24.4%)			
44–58 years	1,323	345 (26.1%)	305 (23.1%)	344 (26.0%)	329 (24.9%)			
>58 years	1,246	342 (27.4%)	300 (24.1%)	289 (23.2%)	315 (25.3%)			
Gender						4.059	3	.255
Female	1,941	476 (24.5%)	510 (26.3%)	467 (24.1%)	488 (25.1%)			
Male	2,085	542 (26.0%)	499 (23.9%)	532 (25.5%)	512 (24.6%)			
Education						12.061	9	.210
Low	623	156 (25.0%)	155 (24.9%)	168 (27.0%)	144 (23.1%)			
Medium-low	1,165	320 (27.5%)	285 (24.5%)	257 (22.1%)	303 (26.0%)			
Medium-high	900	211 (23.4%)	242 (26.9%)	227 (25.2%)	220 (24.4%)			
High	1,338	331 (24.7%)	327 (24.4%)	347 (25.9%)	333 (24.9%)			
Occupation						9.270	12	.680
Full-time	1,871	455 (24.3%)	477 (25.5%)	467 (25.0%)	472 (25.2%)			
Part-time	699	189 (27.0%)	179 (25.6%)	168 (24.0%)	163 (23.3%)			
Unemployed	358	93 (26.0%)	77 (21.5%)	91 (25.4%)	97 (27.1%)			
Retired	788	205 (26.0%)	186 (23.6%)	203 (25.8%)	194 (24.6%)			
Education / service	310	76 (24.5%)	90 (29.0%)	70 (22.6%)	74 (23.9%)			
Recruitment						.512	6	.998
2012	676	172 (25.4%)	172 (25.4%)	169 (25.0%)	163 (24.1%)			
2014	1,583	401 (25.3%)	391 (24.7%)	390 (24.6%)	401 (25.3%)			
2018	1,767	445 (25.2%)	446 (25.2%)	440 (24.9%)	436 (24.7%)			
Device						5.483	3	.140
Smartphone	872	223 (25.6%)	238 (27.3%)	193 (22.1%)	218 (25.0%)			
Computer	3,154	795 (25.2%)	771 (24.4%)	806 (25.6%)	782 (24.8%)			
Field time						.737	12	1.000
Day 1	293	73 (24.9%)	76 (25.9%)	71 (24.2%)	73 (24.9%)			
Days 2–3	1,102	273 (24.8%)	278 (25.2%)	276 (25.0%)	275 (25.0%)			
Days 4–7	855	221 (25.9%)	214 (25.0%)	207 (24.2%)	213 (24.9%)			
Days 8–14	828	211 (25.5%)	208 (25.1%)	206 (24.9%)	203 (24.5%)			
After day 14	948	240 (25.3%)	233 (24.6%)	239 (25.2%)	236 (24.9%)			
Daytime						4.435	9	.881
Morning	1,029	277 (26.9%)	256 (24.9%)	245 (23.8%)	251 (24.4%)			
Afternoon	1,559	382 (24.5%)	405 (26.0%)	382 (24.5%)	390 (25.0%)			
Evening	1,143	285 (24.9%)	281 (24.6%)	291 (25.5%)	286 (25.0%)			
Night	295	74 (25.1%)	67 (22.7%)	81 (27.5%)	73 (24.8%)			

Table A2.2. Model of passing the attention check with only text length and smartphone use as predictors.

	OR
Text length (ref.: Instruction only)	
One paragraph	.478***
Two paragraphs	.353***
Four paragraphs	.181***
Smartphone respondent	1.323***
Constant	3.633***
Observations	4,026
Pseudo-$R^2_{McKelvey \& Zavoina}$	.107

Note. OR = odds ratios. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3. Memory Effects as a Source of Bias in Repeated Survey Measurement[7]

Abstract

A key advantage of longitudinal data collections is the ability to measure change over time by repeatedly asking the same questions to the same respondents. Estimations based on such longitudinal data, as well as other designs that incorporate repetitions of the same questions, generally rely on the assumption that at each point of data collection, respondents answer the questions independently of their previous responses. This assumption implies that respondents either do not remember their previous responses, or that they at least do not use this information in forming their later responses. This is a strong assumption, given that data collections are becoming more and more frequent, giving respondents less time to forget earlier responses. If respondents do, however, remember both being asked the same question and their previous response, they may be influenced by this information. This form of bias is known as a memory effect. In this chapter, we conceptualize the potential role of respondents' memory when answering survey questions and propose a model of the cognitive response process that takes potential memory effects into account. This is supplemented with the literature on the cognitive response process, the sparse existing research on memory effects, as well as adjacent literature on dependent interviewing and question order effects. We conclude the chapter by identifying gaps in this literature and highlighting areas that require additional research to further our understanding of memory effects in longitudinal survey research.

Keywords

measurement error, repeated measurements, memory effects

[7] This chapter is joint work with Annelies G. Blom. The final version has been published: Rettig, T. & Blom, A. G. (2021). Memory Effects as a Source of Bias in Repeated Survey Measurement. In A. Cernat & J. W. Sakshaug (Eds.), *Measurement Error in Longitudinal Data* (pp. 3–18). Oxford University Press.

3.1 Introduction

In longitudinal survey research, the term "memory effect" refers to a specific type of measurement error that occurs when a response is influenced by the respondents' memory of other responses they have previously given. Memory effects may occur in two different scenarios: When respondents are repeatedly asked the same question and when respondents are asked a sequence of questions on a related topic.

In the first scenario, the response to repetitions of the same question may be influenced by respondents' memory of their previous response. This process may be relevant when researchers want to measure change over time in longitudinal studies (Lynn, 2009), in pretest-posttest experimental designs, where a treatment is evaluated by taking identical measurements before and after its administration (Dimitrov & Rumrill, 2003), and to evaluate the reliability, and thus the quality, of measurement instruments (Saris & Gallhofer, 2014). Respondents may use the memory of their previous response as a basis to evaluate their later response against, or even to simply repeat it. In longitudinal studies, this issue becomes even more pressing with the current trend towards more rapid data collection. Respondents are often no longer surveyed once a year, but more and more frequently (monthly, or even daily) and are thus given much less time to forget about previous responses (for an example of a daily study, see Blom et al., 2020).

In the second scenario, memories of previous responses may trigger memory effects in a later response when the topics of two questions are sufficiently related for respondents to evaluate their memory of the first question and answer as relevant to the second question. In survey questionnaires, it is quite common to ask several questions on the same topic such that respondents' answer to one question may be related to their answer to the next. In the extreme, we find this phenomenon when asking a battery of questions designed to measure an underlying latent construct, because the items contributing to the latent construct are related to

each other by design. Respondents may then use their memory of responses to previous items as a basis to evaluate later responses against. This phenomenon is closely related to what has been described as "context effects" or "question order effects" (see the literature review below), though we argue that memory of previous responses, rather than the mere order of questions, is the underlying mechanism here.

Memory effects as a potential source of bias are, therefore, a concern for any survey that incorporates repeated measurement sequences of the same question or more than one question on the same topic. In practice, this applies to virtually every survey. Memory effects are, however, especially problematic for longitudinal studies, as the ability to take repeated measurements from the same respondents is one of their main purposes. Despite their potential to bias results, memory effects have received relatively little attention in the survey literature to date, an important gap in the literature that this chapter aims to fill.

The chapter gives an overview of the ways responses may be influenced by respondents' memory of their previous responses to the same or related questions, as well as the existing literature on the cognitive response process, literature on memory effects, and adjacent literature on dependent interviewing and question order effects. Gaps in the existing literature and areas that require further research are subsequently identified and presented.

3.2 Conceptualizing memory effects

The American Psychological Association defines memory as "the ability to retain information or a representation of past experience, based on the mental processes of learning or encoding, retention across some interval of time, and retrieval or reactivation of the memory" (VandenBos, 2015). In other words, memory refers to information that a person received, stored, and is able to access at a later point in time.

When responding to a survey, the questions received and the response given will be stored in respondents' memory for some amount of time. This may pose a problem for measurements if previous questions and responses are still present in respondents' memory at the time the same or a similar question is repeated. In particular, measurement error may occur if, instead of going through a new, independent cognitive response process, respondents use their memory of the previous question in their processing of the later question. Then the participants' memory of a previous question and response may introduce measurement variance and bias.

Conceptually, memory effects can interfere in several ways with the cognitive response process as proposed by Tourangeau et al. (2000). On the left of Figure 3.1 we replicate Tourangeau et al.'s model; on the right we display at which points respondents' memory may interfere with this ideal process. The first step of their 4-step model encompasses question comprehension where respondents read, process, and seek to understand the focus of the question itself. This process may, however, trigger a secondary process of question recognition where respondents identify whether the same or similar questions were asked before. In the second step, respondents retrieve from their memory the relevant information they need to answer the question at hand. As a part of this process, respondents may retrieve their previous responses to the same or similar questions. In the third step, respondents form a judgement on the matter based on the retrieved information. This judgement may be influenced by the previous responses that it is evaluated against to form a more consistent overall picture. Alternatively, this step may even be skipped completely if respondents treat their previous response as an already existing and sufficient judgement on the matter. Finally, in the fourth step, respondents select a response that corresponds to their (now formed or simply reiterated) judgement.

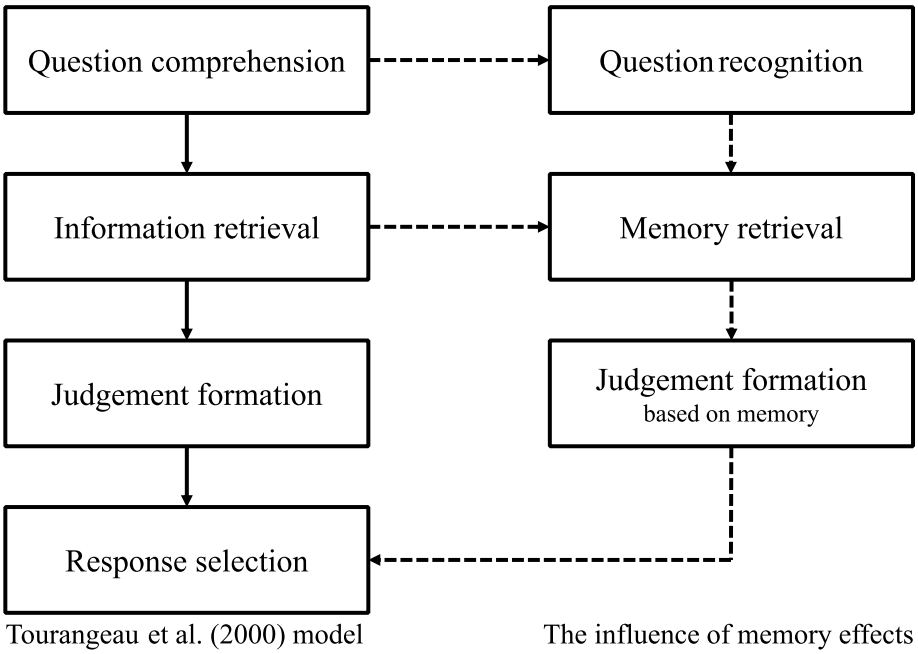

Figure 3.1. Illustration of the cognitive response process by Tourangeau et al. (2000) extended by memory effects.

Please note that, despite their semantic similarity, memory effects are distinctly different from recall errors. Recall error describes a situation in which respondents inaccurately or incompletely retrieve the information needed to answer a question (see Eisenhower et al., 1991; Tourangeau et al., 2000). Thus, recall errors occur when respondents are unable to accurately recall the information that researchers need them to recall, whereas memory effects occur when respondents remember previous questions and their response to these, i.e. they hold a memory of something that researchers do not want them to hold.

3.2.1 Memory effects in repeated measurements of exactly the same question

Memory effects in the strictest sense occur when respondents are repeatedly asked exactly the same question. When measurements are repeated, researchers expect respondents to answer the repetitions independently of their previous responses. Memory effects may, however, occur if respondents remember the previous time(s) the same question has been asked and are subsequently influenced by their memory of their previous response.

Once respondents have recognized the question as one that they have answered before and once they have retrieved their previous response from memory, they may use this information in one of two ways: either as an existing judgement on the matter that they can re-use, or as a basis for reflection on their current opinion. In the first case, as respondents have now retrieved a pre-formed judgement, they may elect to use a cognitive shortcut and reiterate their previous response. This may be done either without any reflection on their current stance at all or by superficially evaluating the previous response to be "close enough" to their current stance. This shortcut allows respondents to skip the formation of a new judgement altogether, thus reducing the cognitive burden of the response process, saving them time and effort, but it leads to a response that violates the statistical assumption of independent measurements. This cognitive shortcut has clear parallels with the phenomenon of satisficing, in which some respondents put in the minimum effort that is required to reach an acceptable response as easily as possible, rather than trying to reach the optimal response with higher effort (see, for example, Krosnick, 1991). In this chapter, we thus use the term "memory satisficing" to describe this type of memory effect (see Figure 3.2).

Another way in which respondents may use their previous response after retrieving it from memory is as a basis for forming a new judgement. In this case, respondents evaluate their current stance against their previous stance at the time the question was last presented to them. In doing so, respondents may strive for consistency in their responses over repetitions of the same question. In addition, in a longitudinal survey context, respondents may reflect upon whether and how their position has changed since their previous answer (i.e. "I have rated party X 7/10 last year, but I like them a bit more now"). Either of these options ultimately result in a more consistent overall set of responses and are thus part of the "memory consistency" model (Figure 3.2).

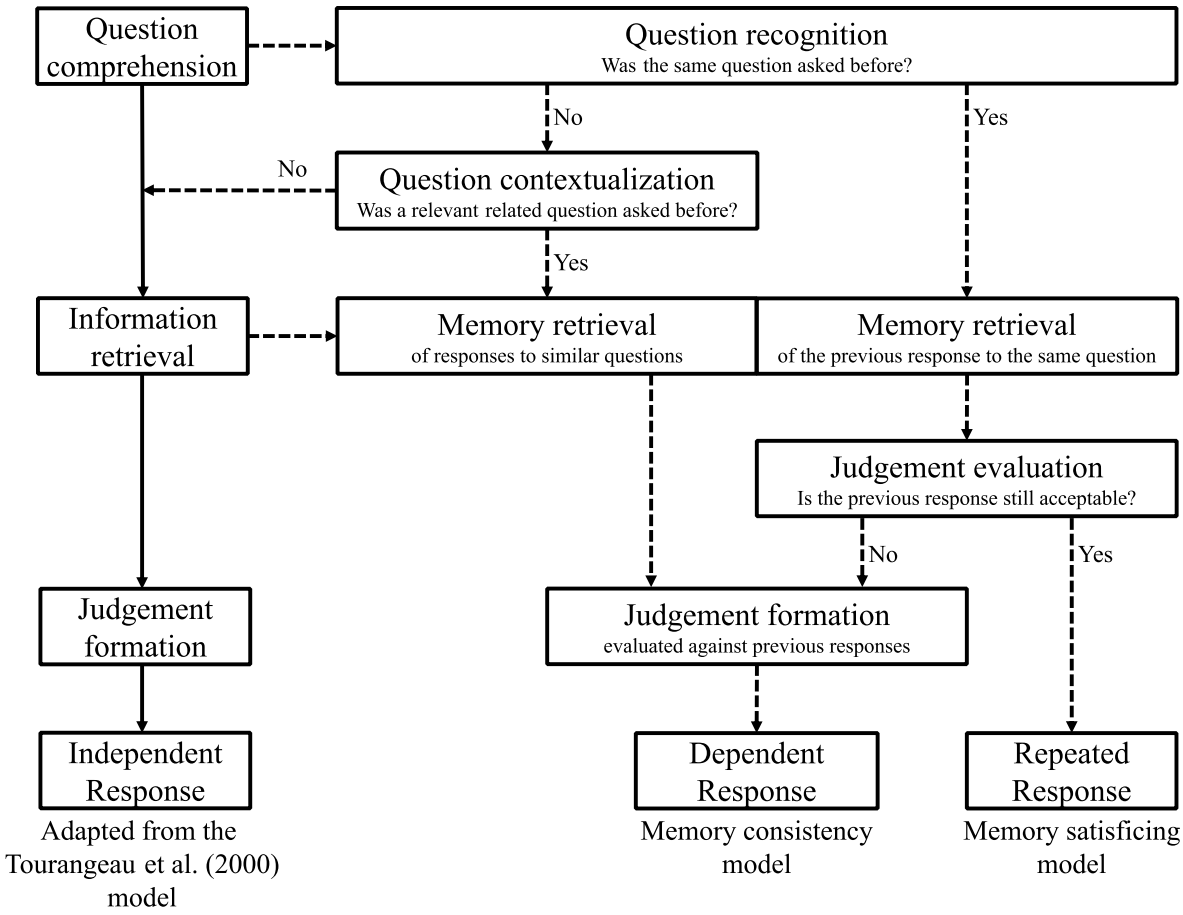

Figure 3.2. Models of specific memory effects in the cognitive response process.

3.2.2 Memory effects in a sequence of similar or related items

In a sequence of items concerning a common topic, the response to an item towards the end of the sequence is often different from what it would have been at the beginning of the sequence (see also the literature on question order effects, presented below). Question order effects may occur for many reasons, such as items providing further context to each other, setting a certain mood for respondents, and different information being made accessible. Question order effects are, however, not only caused by the act of merely asking certain questions in itself. Later responses are also influenced by the responses that respondents have given to these earlier questions. That is, respondents are influenced in their formation of judgements to later items by their memory of the judgements on the same topic they have made to earlier items (i.e. “I have rated party X 7/10 and I like party Y a bit less”). Later responses in the sequence

of questions are thus formed in relation to the earlier responses. This may happen in an attempt to present a consistent overall response on the topic, which may occur especially when such a consistent set of beliefs was not already consciously formed beforehand. Thus, the memory consistency model presented in Figure 3.2 also applies to sets of items that are related in topic.

In contrast, in some survey settings memory of previous questions and responses to these may lead respondents to exclude information from further consideration. This may be the case if respondents feel that they have already provided the same information in response to a different question. Such an exclusion may become a problem when it leads respondents to exclude information which still remains relevant to the new question at hand.

A hypothetical example: A political party suggests introducing more environmental protection and higher taxes to finance this protection. In a survey, respondents are asked to rate their agreement with the party's stance on environmental protection. A respondent agrees with more environmental protection but disagrees with higher taxes. This respondent may report moderate agreement with the party's stance on environmental protection, as they agree with the environmental protection itself, but disagree with the proposed means to fund it. In a similar survey, respondents are asked to rate their agreement with the party's stance on taxes and the party's stance on environmental protection separately. A similar respondent, who also agrees with more environmental protection, but disagrees with higher taxes, may report high agreement with the party's stance on environmental protection because they feel their disagreement with higher taxes was dealt with in a separate question. The two surveys may thus lead to different conclusions about whether respondents agree with the political party's stance on environmental protection.

Respondents in this case exclude information they feel has become irrelevant to their later judgements in an attempt to reduce redundancy in their responses. This too is a realization of

the memory consistency model, as it serves to create a clearer and more consistent overall view of the topic (see also the literature on question order effects presented below).

3.2.2.1 Memory effects in item batteries for latent constructs

Item batteries for latent constructs present a special case of related items, as the items are not only connected by a common topic, but even seek joint measurement of an underlying trait. Such item batteries are susceptible to very similar memory effects as other related questions: Respondents evaluate their later responses against their memory of earlier responses to related items and, thus, give more consistent responses. This is especially likely to happen if respondents realize what exactly researchers seek to measure and actively try to influence the result to match their perceived self-image. Responses to the items in the battery will then be more consistent with both each other and the desired overall result. Again, this phenomenon is described in the memory consistency model presented above.

Furthermore, if multiple items are very similar in nature and wording, as items in batteries for latent constructs tend to be, respondents may not understand a certain item as sufficiently different from a previous one. That is, respondents may misidentify a question as one that has been asked before or understand it to ask about exactly the same thing as a previous question. In this case, respondents may treat an item they misidentified as a repetition of a previous question the same way they would an actual repetition. This may lead to the same memory effects as actual repetitions of the same question described above. To a somewhat lesser degree, such misidentification is also a concern for any sequence of similar questions on a common topic.

3.2.3 The potential impact of memory effects

As described above, the primary effect of a respondent's memory of previous responses is an increase in the consistency of later responses. Such "memory-enhanced consistency" (Alwin,

2011) may have a profound impact on the results of substantive research. If respondents give more consistent or identical responses, actual change may go underreported. This will lead to an underestimation of treatment effects and of estimates of change in longitudinal survey data. Conversely, reliability will be inflated, leading to an overestimation of measurement quality (Alwin, 2011). In a series of related questions or item batteries for latent constructs, questions that are answered more consistently due to memory may also show inflated correlations. In contrast, the exclusion of certain information because of memory effects may lead to respondents' views seeming more differentiated than they actually are. This means that memory effects may potentially, depending on the situation, either inflate or mask certain effects and thus lead to incorrect inference from the data.

These errors become even more problematic when we take into account that the occurrence of memory effects may not be randomly distributed across respondents. As information (i.e. a previous response) must be present in respondents' memory in order to cause memory effects, they may be more common for respondents with better individual memory capabilities and higher cognitive abilities. These abilities may in turn be linked to respondent characteristics such as education or age (see, for example, Knäuper et al., 2016). The prevalence of memory effects may thus be systematically different across different groups of respondents.

3.3 The state of the literature on memory effects in surveys

To evaluate the state of the literature on memory effects, we first focus on the cognitive response process in theory, followed by empirical research on respondent memory, dependent interviewing as an alternative to question repetitions in longitudinal surveys, and finally question order effects as memory effects in related questions.

3.3.1 Literature on the cognitive response process

As presented above, the model of the cognitive response process as proposed by Tourangeau et al. (2000) includes four basic steps: The comprehension of the question itself, retrieval of relevant information, formation of a judgement, and finally, selection of a response. Strack & Martin (1987) propose a similar model with a secondary avenue that allows the circumvention of some of these steps: The authors postulate that respondents will, after interpreting the question, check whether a judgement on the matter is already present. A prior judgement may in this case be an opinion that respondents already formed and held independently of any survey, or indeed a judgement that was formed when respondents were asked about the same topic before. If such a pre-formed judgement is present, respondents will recall and use it to formulate their response instead of accessing relevant information in order to generate a new judgement (N. Schwarz & Strack, 1991; Strack & Martin, 1987).

While the presence of a prior judgement applies especially to repetitions of the same question, where respondents have formed a judgement on exactly the same question before, the information retrieval stage may also be affected by similar questions. As Tourangeau & Rasinski (1988) note, respondents will retrieve the information that is most accessible to them, which is not necessarily the most important information (see also Tourangeau et al., 1989). In addition, prior questions on a similar topic increase the accessibility of information used to answer them, as the same information was recently accessed before (N. Schwarz & Strack, 1991; Tourangeau & Rasinski, 1988). A recently generated memory of a judgement for a related question may be part of this more accessible information (Tourangeau et al., 1989). Finally, N. Schwarz & Strack (1991) also argue that respondents will stop retrieving more information once enough information to form a judgement has been accessed, rather than retrieve any related piece of information they possibly can (see also Tourangeau et al., 1989). When considering them in combination, these points indicate that when respondents retrieve information to form a judgement, they are likely to access the same information they

used to answer previous questions as well as their previous judgements, to the exclusion of other relevant information. Therefore, previous questions may have a strong influence on later responses.

While the information used to answer previous questions is likely to be retrieved during the response process for later questions, respondents still expect later questions to ask for different information than what they already provided (Strack & Martin, 1987). Therefore, while respondents may put their later responses in relation to earlier ones, they will generally avoid redundant responses and thus ignore information they have already provided (N. Schwarz & Strack, 1991; Strack & Martin, 1987).

3.3.2 Literature on respondent memory

Very little literature to date explicitly deals with the impact of respondents' memory of their previous responses. Most notably, van Meurs & Saris (1990) systematically investigate respondents' ability to recall their responses to a number of questions within one interview and after two weeks. They find that, when prompted to recall their previous response, about 70% of respondents can correctly repeat it after a period of about 9 minutes, and about 40% can do so after two weeks. Subsequent studies have come to similar conclusions: For a time period of 20 minutes between the question and the prompt to recall the response, H. Schwarz et al. (2020) report that 60% of respondents give the correct response; Rettig et al. (2019) replicate this finding with 61%. In addition, van Meurs & Saris (1990) report that respondents who take longer to complete the interview and, thus, for whom the time between giving a response and being prompted to recall it is longer, are less able to correctly recall their response. This effect, however, is not replicated by H. Schwarz et al. (2020).

In similar research, Alwin (2011) reports a slight decline over time in respondents' recall ability for nouns that were read out previously. Out of a list of 10 nouns an average respondent was able to recall 6 immediately afterwards and 5 after a period of 10 to 15

minutes. Overall, the time required for respondents to forget about their previous response is still unclear. What can be said is that a majority of respondents can recall their responses within the course of one interview. Furthermore, a period of two weeks is insufficient to reliably ensure that previous responses are forgotten.

Other than the time that has passed since a response was first given, van Meurs and Saris (1990) also suggest that memory is affected by the survey contents placed in-between a question and its repetition. The authors report that respondents are less likely to recall their response at a later point if questions on similar topics are asked in the meantime. Based on this finding, H. Schwarz et al. (2020) devised a memory interference task to reduce respondents' ability to recall their responses to questions placed before it. However, the authors find that this memory interference task did not lead to a reduced ability to recall their previous responses in their study.

Some studies also suggest that whether a previous response can be recalled is dependent on the content of the recalled information itself (Alwin, 2011; van Meurs & Saris, 1990). Rettig et al. (2019) find that recall ability differs by question type. The authors report that respondents can most easily recall their responses to questions on their behavior, followed by questions on attitudes. Responses to questions that deal with respondents' beliefs are the most difficult to recall.

Furthermore, some research suggests that respondents can more easily recall responses that are based on stronger opinions. Both van Meurs & Saris (1990) and Rettig et al. (2019) find that respondents who give an extreme response, i.e. on an endpoint of the response scale, are more likely to correctly recall it. In addition to it being an expression of a strong, salient opinion, an endpoint response may be easier to remember than one placed somewhere along the scale. Similarly, Rettig et al. (2019) report that respondents who correctly recall their response also report higher confidence about recalling their response. Jaspers et al. (2009)

find the same in their investigation of retrospective accounts of prior attitudes. High certainty about correctly recalling one's response may, thus, also be a sign of a strong, central, and salient opinion.

Finally, Rettig et al. (2019) find no difference between respondents with high panel experience and freshly recruited respondents in their ability to recall previous responses. However, while fresh and experienced respondents seem to be equally likely to recall their responses, their response behavior may still be differentially affected. The extent to which the ability to recall a previous response translates into memory effects has, at this point, not been researched yet. It may be different for more experienced respondents than those freshly recruited into a panel.

So far, no literature on memory effects in survey research explicitly deals with the role of respondents' individual cognitive ability or quality of memory. As mentioned above, cognitive ability, however, may play a role in causing memory effects, because information from memory can only influence responses if respondents are able to recall it. Further research on the role of cognitive ability in the context of memory effects is therefore needed.

3.3.3 Literature on dependent interviewing

Another source of relevant research literature is that on dependent interviewing in longitudinal surveys. Dependent interviewing is a technique used to update the information respondents provided in longitudinal settings without expressly asking the same questions again. Respondents are instead presented with their previous responses and are asked to indicate whether their situation has changed since the last interview. This technique is usually applied for factual information such as a respondent's employment status. The aim of dependent interviewing is a reduction of the response burden and of overreporting of change that results from measurement error rather than actual change (Hogendoorn, 2004; Jäckle, 2008; Jäckle & Eckman, 2020). Some research, however, suggests that presenting

respondents with their previous answers may also lead to underreporting of actual change (Eggs & Jäckle, 2015; Lugtig & Lensvelt-Mulders, 2014). The likely mechanism here is that, when presented with a question and their own prior judgement, some respondents choose to agree with their previous response rather than reconsider it.

This mechanism of reducing the cognitive effort of answering questions by reiterating a previous response rather than going through the cognitive response process anew is a form of satisficing (see Krosnick, 1991). Through the same mechanism, respondents may use their memory of previous responses to satisfice, even when they are not explicitly reminded of it (i.e. the memory satisficing model presented above).

3.3.4 Literature on question order effects

We also draw from an extensive body of literature on question order effects, which includes findings that can be applied (or even attributed) to memory effects in a sequence of related questions (see, for example, Rasinksi et al., 2012 for a comprehensive overview of question order effects). Schuman & Presser (1981) distinguish between unconditional question order effects, which are caused by the process of asking survey questions, and conditional question order effects, where subsequent responses are influenced by the responses to earlier questions. Conditional question order effects and memory effects are congruent concepts, because later responses are not affected by the response itself but by the memory of a response.

The order of questions is known to influence responses in several ways. Most prominent are consistency effects (also called carryover effects), where respondents generate more consistent responses toward the end of a series of related items than at the beginning (Knowles, 1988; Rasinski et al., 2012; Schuman & Presser, 1981; Schumann et al., 1981; Tourangeau & Rasinski, 1988). Later responses are thus influenced by those already given to form a more consistent overall picture. We describe this effect in the memory consistency model in Figure 3.2. Consistency effects also occur in item batteries for latent constructs,

where later items have been shown to produce higher reliability and to be better predictors of the overall score than earlier items (Knowles, 1988). Furthermore, later items are associated with lower within-subject variance and higher between-subject variance than earlier items (Knowles, 1988). This has two implications. First, respondents become more internally consistent as they give more responses to a series of strongly related questions. Second, across respondents, response profiles become more differentiated as more questions are asked.

The second type of question order effects are contrast effects (sometimes also called backfire effects). These contrast effects lead respondents to exclude information they already provided in response to an earlier question from their consideration of later questions (Rasiksi et al., 2012; Schuman & Presser, 1981; Schuman et al., 1981; Tourangeau & Rasinski, 1988; van de Walle & van Ryzin, 2011). Researchers have primarily attributed this phenomenon to an effort of reducing redundancy in a person's responses (Rasinski et al., 2012; Strack & Martin, 1987), which is another way of reaching internal consistency. This type of question order effect has been observed in settings where a specific question is followed by a more general question (Rasinski et al., 2012; Schuman & Presser, 1981; Strack & Martin, 1987; Tourangeau & Rasinski, 1988). A typical example of this is asking respondents first specifically about their marital satisfaction and subsequently about their general life satisfaction. In such a situation, respondents may perceive the life satisfaction question to mean satisfaction with areas in life other than their marriage (Rasinski et al., 2012). Contrast effects are consistent with the notion that respondents expect later questions to ask for new information (see also the literature on the cognitive response process above).

A third type of question order effects are assimilation effects, where respondents understand and treat a general question after a series of specific questions as a summary of the previous questions rather than as an independent judgement in itself (Rasinski et al., 2012). A typical example is, again, general life satisfaction being asked after a long series of specific questions

on satisfaction (Rasinski et al., 2012; Smith, 1982; Tourangeau & Rasinski, 1988). The response to the general question then becomes a summary of the responses to the previous questions. This means responses to the final general question are formed (nearly) exclusively on the basis of responses given to the earlier questions.

As demonstrated, the conditional question order effects described in the literature can be re-interpreted as memory effects according to the memory consistency model in Figure 3.2. Consistency effects, contrast effects and assimilation effects are all manifestations of respondents using their memory of previous responses to form their judgements to subsequent questions. They do this in three ways: as a basis to evaluate later responses against (consistency effects), to avoid redundancy of the information already provided (contrast effects), or as the basis for summarizing their previous responses (assimilation effects).

With respect to the persistence of question order effects, we again find similarities to memory effects. For example, Schuman et al. (1983) find that question order effects between two related questions still persist when 17 unrelated questions are placed between the related questions. The finding that question order effects can arise even without the affected questions being asked directly after each other strengthens the notion that the memory of previous responses is involved.

Furthermore, not all questions are equally susceptible to question order effects (Rasinski et al., 2012). According to McFarland (1981), vague and diffuse questions are more likely to produce these effects than clear and specifically worded questions. That is, respondents are more inclined to take cues from other questions, if the present question itself does not provide sufficient cues (see also Strack & Martin, 1987). Aside from the questions themselves, whether or not respondents already have a clearly structured belief system with strongly held opinions also plays a role. Responses are more affected by the question order if no such belief system is present (Tourangeau & Rasinski, 1988; van de Walle & van Ryzin, 2011).

Respondents are thus more inclined to refer to their other responses for cues, if their later response is not clearly determined by a strong opinion already.

3.4 Conclusion

When considering measurement errors in longitudinal surveys, the effect a respondent's memory plays in their response to repeated measurements is crucial. Our chapter conceptualizes such memory effects and proposes an extension of the cognitive response model introduced by Tourangeau et. al (2000). Our memory consistency and memory satisficing models displayed in Figure 3.2 describe how a respondent may use their memory of a response to a previous question. Either they use the remembered information to give a response that is consistent with their memory, or they use this information to cut the cognitive response process short and, thus, satisfice.

We demonstrate that various response effects described in the survey methodological literature can be understood in this framework, even if the effects have not traditionally been perceived as memory effects. In the longitudinal survey literature, we find memory effects in the context of dependent interviewing. In dependent interviewing, instances in which respondents use the supplied information on their previous answers to carefully consider whether their situation has changed since the previous interview can be viewed under our memory consistency model. However, instances in which respondents use the information provided to shortcut their response process should be viewed under our memory satisficing model.

In the general survey literature, which is naturally also relevant to longitudinal survey measurements, we further find memory effects in the context of question order effects. This literature distinguishes three types of effects: consistency effects, contrast effects, and assimilation effects. All three are manifestations of our memory consistency model.

Our model thus enables survey researchers to understand response effects that arise in repeated measurements due to respondents' memory. Despite their demonstrated impact on measurement, especially in the context of longitudinal surveys, memory effects have received surprisingly little attention from methodologists. As a consequence, we still see large gaps in the literature.

In the context of longitudinal survey measurement, the most pressing open research questions relate to (a) dependent interviewing and (b) the time intervals after which responses to previous questions are forgotten. With respect to (a), our memory effects model highlights that more research on the underlying cognitive processes during dependent interviewing is needed. Furthermore, future research should identify the parameters that lead to either a memory consistency effect or a memory satisficing effect. Such insights would have practical implications, because the former is the declared goal of the dependent interviewing method, while the latter leads to underreporting of change and should be avoided.

Our literature review on true memory effects shows that only few studies have investigated the time period needed to avoid memory effects in repeated measurements. We know that during the course of an interview a majority of respondents can recall their responses to previous questions. Furthermore, we have some suspicion that even a period of two weeks may be insufficient to ensure that previous responses are forgotten. However, the empirical evidence is very sparse. Too sparse, in fact, to inform longitudinal questionnaire design in panel surveys that are repeated at short intervals. Recently, we have seen a surge in panel surveys that are conducted very frequently to densely monitor social change during the onset of the COVID-19 pandemic. Blom et al. (2020), for example, implemented a high-frequency rotating panel design, for which panelists are interviewed every week and the survey content remains largely unchanged. In such a survey setting, memory effects may well affect the longitudinal measurement if respondents remember their responses from the previous week.

More research on the decrease of memory effects is, thus, needed to inform both longitudinal survey design and estimation based on longitudinal data.

References

Alwin, D. F. (2011). Evaluating the Reliability and Validity of Survey Interview Data Using the MTMM Approach. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), *Question Evaluation Methods* (pp. 265–295). John Wiley and Sons.

Blom, A. G., Cornesse, C., Friedel, S., Krieger, U., Fikel, M., Rettig, T., Wenz, A., Juhl, S., Lehrer, R., Möhring, K., Naumann, E., & Reifenscheid, M. (2020). High-Frequency and High-Quality Survey Data Collection: The Mannheim Corona Study. *Survey Research Methods*, *14*(2), 171–178. https://doi.org/10.18148/srm/2020.v14i2.7735

Dillman, D. A. (1978). *Mail and Telephone Surveys: The Total Design Method*. Wiley and Sons.

Dimitrov, D. M., & Rumrill, P. D. (2003). Pretest-Posttest Designs and Measurement of change. *Work*, *20*(2), 159–165.

Eggs, J., & Jäckle, A. (2015). Dependent interviewing and sub-optimal responding. *Survey Research Methods*, *9*(1), 15–29. https://doi.org/10.18148/srm/2015.v9i1.5860

Eisenhower, D., Mathiowetz, N. A., & Morganstein, D. (1991). Recall Error: Sources and Bias Reduction Techniques. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Error in Surveys* (pp. 127–144). John Wiley & Sons. https://doi.org/10.1002/9781118150382

Hoogendoorn, A. (2004). A Questionnaire Design for Dependent Interviewing that Addresses the Problem of Cognitive Satisficing. *Journal of Official Statistics*, *20*(2), 219–232.

Jäckle, A. (2008). Dependent Interviewing: Effects on Respondent Burden and Efficiency of Data Collection. *Journal of Official Statistics*, *24*(3), 411–430.

Jäckle, A., & Eckman, S. (2020). Is That Still the Same? Has that Changed? On the Accuracy of Measuring Change with Dependent Interviewing. *Journal of Survey Statistics and Methodology*, *8*(4), 706–725. https://doi.org/10.1093/jssam/smz021

Jaspers, E., Lubbers, M., & De Graaf, N. D. (2009). Measuring Once Twice: An Evaluation of Recalling Attitudes in Survey Research. *European Sociological Review*, *25*(3), 287–301. https://doi.org/10.1093/esr/jcn048

Knäuper, B., Carrière, K., Chamandy, M., Xu, Z., Schwarz, N., & Rosen, N. O. (2016). How Aging Affects Self-Reports. *European Journal of Ageing*, *13*(2), 185–193. https://doi.org/10.1007/s10433-016-0369-0

Knowles, E. S. (1988). Item Context Effects on Personality Scales: Measuring Changes the Measure. *Journal of Personality and Social Psychology*, *55*(2), 312–320. https://doi.org/10.1037/0022-3514.55.2.312

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Lugtig, P., & Lensvelt-Mulders, G. J. L. M. (2014). Evaluating the Effect of Dependent Interviewing on the Quality of Measures of Change. *Field Methods*, *26*(2), 172–190. https://doi.org/10.1177/1525822X13491860

Lynn, P. (2009). Methods for Longitudinal Surveys. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (pp. 1–19). John Wiley & Sons. https://doi.org/10.1002/9780470743874.ch1

McFarland, S. G. (1981). Effects of Question Order on Survey Responses. *Public Opinion Quarterly*, *45*(2), 208–215. https://doi.org/10.1086/268651

Rasinski, K. A., Lee, L., & Krishnamurty, P. (2012). Question Order Effects. In Cooper, H., Camic. P. M., Long, D. L., Panter, A. T., Rindskopf, D., & Sher, K. J. (Eds.) *APA Handbooks in Psychology. APA Handbook of Research Methods in Psychology, Vol. 1. Foundations, Planning, Measures, and Psychometrics* (pp. 229–248). American Psychological Association.

Rettig, T., Höhne, J. K., & Blom, A. G. (2019). *Investigating Respondents' Ability to Recall Previous Responses to Different Types of Questions in a Probability-Based Online Panel*. Presentation at the European Survey Research Association (ESRA) 2019 Conference, Zagreb, Croatia.

Saris, W. E., & Gallhofer, I. N. (2014). *Design, Evaluation, and Analysis of Questionnaires for Survey Research* (2nd ed.). John Wiley and Sons.

Schuman, H., Kalton, G., & Ludwig, J. (1983). Context and Contiguity in Survey Questionnaires. *Public Opinion Quarterly*, *47*(1), 112–115. https://doi.org/10.1086/268771

Schuman, H., & Presser, S. (1981). *Questions and Answers in Attitude Surveys. Experiments on Question Form, Wording, and Context*. Academic Press.

Schuman, H., Presser, S., & Ludwig, J. (1981). Context Effects on Survey Responses to Questions About Abortion. *Public Opinion Quarterly*, *45*(2), 216–223. https://doi.org/10.1086/268652

Schwarz, H., Revilla, M., & Weber, W. (2020). Memory Effects in Repeated Survey Questions. Reviving the Empirical Investigation of the Independent Measurements Assumption. *Survey Research Methods*, *14*(3), 325–344. https://doi.org/10.18148/srm/2020.v14i3.7579

Schwarz, N. & Strack, F. (1991). Context Effects in Attitude Surveys: Applying Cognitive Theory to Social Research. *European Review of Social Psychology*, *2*(1), 31–50. https://doi.org/10.1080/14792779143000015

Smith, T. W. (1982). *Conditional Order Effects*. GSS Technical Report No. 33. National Opinion Research Center.

Strack, F., & Martin, L. L. (1987). Thinking, Judging, and Communicating: A Process Account of Context Effects in Attitude Surveys. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social Information Processing and Survey Methodology. Recent Research in Psychology.* (pp. 123–148). Springer.

Tourangeau, R., and K. A. Rasinski, 1988. "Congitive Processes Underlying Context Effects in Attitude Measurement". *Psychological Bulletin* 103(3): pp. 299–314.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin, 103*(3), 299–314. https://doi.org/10.1037/0033-2909.103.3.299

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

VandenBos, G. R. (Ed.). (2015). *APA Dictionary of Psychology* (2nd ed.). American Psychological Association.

van Meurs, A., & Saris, W. E. (1990). Memory Effects in MTMM Studies. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait Multimethod Studies* (pp. 134–146). North Holland.

van de Walle, S. & van Ryzin, G. G., (2011). The Order of Questions in a Survey on Citizen Satisfaction With Public Services: Lessons From a Split-Ballot Experiment. *Public Administration, 89*(4), 1436–1450. https://doi.org/10.1111/j.1467-9299.2011.01922.x

4. Memory Effects: A Comparison Across Question Types[8]

Abstract

A crucial assumption of survey measurements is that respondents carefully perceive, reflect upon, and provide an answer to a given question and that this process is independent of respondents' memory of their responses to previous questions. A violation of this assumption may considerably affect parameter estimations. To shed light on such memory effects, we investigate the ability of respondents to remember their answers to three types of survey questions (beliefs, attitudes, and behaviors) within one wave of a probability-based online panel survey. We find that respondents' ability to correctly reproduce their answers after 20 minutes is overall high and differs across questions on beliefs, attitudes, and behaviors. Furthermore, respondents who gave extreme answers are more likely to correctly reproduce their response than respondents who gave non-extreme answers.

Keywords

memory effects, online panel, extreme responses, measurement error, repeated measurement

Acknowledgements

This article uses data from the wave 38 of the German Internet Panel (GIP; DOI: 10.4232/1.13391; Blom et al. 2019). In addition, the following variables used in our analyses are available via the On-Site Data Access (ODA) facilities of the GIP: device, response time. A study description can be found in Blom et al. (2015). The GIP is funded by the German Research Foundation (DFG) as part of the Collaborative Research Center 884 (SFB 884; Project Number 139943784; Project Z1). The authors would like to thank Melanie Revilla for her helpful comments on an early version of this manuscript.

[8] This chapter is joint work with Annelies G. Blom and Jan Karem Höhne. The final version has been accepted for publication: Rettig, T., Blom, A.G., & Höhne, J.K. (forthcoming). Memory Effects: A Comparison Across Question Types. *Survey Research Methods*.

4.1 Introduction

In survey research, the questions respondents receive and the answers they provide will be stored in respondents' memory for some time. This natural process may pose a problem for research designs that rely on repeatedly asking respondents the same or very similar questions, if this information is still present in respondents' memory at the time a question is repeated to them. Respondents who recognize a question that has been asked before and remember their previous answer may use this information in their cognitive processing of the question repetition. This type of measurement error caused by respondents' memory of their previous answers is commonly referred to as "memory effects" (van Meurs & Saris, 1990).

Normally, respondents are expected to undergo the full cognitive response process independently for each question, including comprehending the question, retrieving relevant information, forming a judgement, and selecting the appropriate response (see Tourangeau et al., 2000, pp. 7–16). However, if their own previous answer is part of the information that respondents retrieve, they may take it as an existing judgement to either evaluate their later response against, or even take a cognitive shortcut and simply repeat it (see Rettig & Blom, 2021 for a conceptualization of memory effects in relation to Tourangeau et al.'s response process model).

Memory effects can pose a problem across a variety of survey designs that incorporate repeated measurements. In particular, longitudinal surveys, surveys that evaluate experimental treatment effects, and surveys that examine measurement quality commonly rely on some form of repetition of the same questions (Rettig & Blom, 2021).

In longitudinal surveys, the same respondents typically receive the same questions in regular intervals to measure change over time. The possibility of measuring change over time on the respondent level may even be considered the main reason for the importance and popularity of longitudinal study designs in behavioral and social research (Lynn, 2009). Recent trends

toward more frequent data collection, sometimes on a weekly or even daily basis, mean that respondents are often given less time to forget previous answers in longitudinal settings than they would, for example, have in more traditional annual surveys (Blom et al., 2020).

Repeated measurements after a relatively short time, usually within the same survey, are also commonly used in experimental research. For example, in pretest-posttest designs, which are especially popular in psychology and health research. Two identical measurements are taken to evaluate the effect of a treatment – one before and one after the treatment has been administered (Campbell & Stanley, 1966, pp. 7–25; Dimitrov & Rumrill, 2003).

The use of repeated measurements is also key for the evaluation of measurement quality across data collection methods, for instance in a test-retest or quasi-simplex design to estimate reliability (Alwin, 2007, pp. 95–110, 2010, 2011; Saris & Gallhofer, 2014, pp. 178–183) and in a multitrait-multimethod (MTMM) design to estimate reliability and validity (Alwin, 2007, pp. 67–93; Campbell & Fiske, 1959; Saris et al., 2004; Saris & Gallhofer, 2014, pp. 197–202). With the exception of quasi-simplex models, these designs are, again, usually reliant on measurements that are taken within the same survey with only a relatively short time between the repetitions.

In summary, there are various reasons for implementing repeated measurements in survey research. For any such application, measurement theories assume that the repeated measurements are independent from one another, in the sense that an earlier answer does not influence the answer given to a later question (Alwin, 2011; Saris & Gallhofer, 2014, pp. 181–182). This includes the assumption that for repeated questions respondents undergo the cognitive response process (see Tourangeau et al., 2000, pp. 7–16) anew; either with no memory of their previous response present or at least without using this information in forming their later response.

However, respondents who remember their previous answer may use it to evaluate their later answers and thus respond more consistently overall or simply repeat their previous response without rethinking it (independent of any actual change that may have occurred in the meantime). This may in turn result in a number of biases. In longitudinal surveys seeking to measure individual changes in beliefs, attitudes, and behaviors over time, these changes may be underestimated due to the more consistent responses. Similarly, in studies with a pretest-posttest design researchers may underestimate treatment effects. When evaluating the measurement quality of data collection methods, reliability and validity might be overestimated due to the artificially higher consistency of responses (Alwin, 2011). These biases may potentially have a profound impact on the conclusions drawn from studies with repeated measurements (Alwin, 2010, 2011; Saris et al., 2010; van Meurs & Saris, 1990).

Respondents choosing to repeat their previous answer (rather than rethink it) has also been shown to be a concern in dependent interviewing. Using this technique, instead of repeatedly asking the same questions, respondents are presented with their previous response and asked to indicate whether it has changed since the last time in order to reduce response burden and overreporting of change that results from measurement error rather than actual change (Hoogendoorn, 2004; Jäckle, 2008; Jäckle & Eckman, 2020). Some research, however, suggests that this may cause that respondents underreport actual change. Eggs and Jäckle (2015) as well as Lugtig and Lensvelt-Mulders (2014) demonstrated that respondents tend to underreport changes in dependent interviews, indicating that these respondents have taken a cognitive shortcut and, for instance, simply chosen to “agree” with their previous response, rather than rethink it. This cognitive shortcut to reduce response burden can be considered a form of satisficing (see Krosnick, 1991). In other words, some respondents who were presented with their previous response have been shown to simply reuse it instead of undergoing the cognitive response process to check whether their old response is still correct. These findings increase concerns that, in line with the “memory satisficing” model proposed

by Rettig and Blom (2021), respondents who remember their previous answer (without having it presented to them) may use it to satisfice in the same way.

4.2 Background and hypotheses

A small group of researchers has investigated memory effects in great detail with purposively designed experiments. Most notably, van Meurs and Saris (1990) used data from the Dutch NIPO telepanel, a predecessor of modern online panels, to investigate to what extent respondents were able to correctly repeat their previous answers to six political items depending on whether respondents indicated that they remembered their answers. This distinction between alleged and actual recall is useful because, as van Meurs and Saris (1990) argue, a correct repetition of the previous answer could also be a sign of an unchanged opinion or correct guessing due to chance. The proportion of respondents who correctly repeat their answer despite alleging that they do not remember it can serve as a baseline for these alternative explanations of correct repetitions. In turn, alleged recall by itself would not be a sufficient measure of respondents' recall ability either, as it can be expected that some respondents who claim to remember their answer will give an incorrect recollection.

The results by van Meurs and Saris (1990) showed that 70% of respondents correctly reproduced their previous answer (i.e., selected the same point on a 10-point scale) after a period of about 9 minutes. The proportion was lower for respondents who took more time answering the survey. This finding is in line with the long-established general concept that human memory tends to decline over time (Bradburn et al., 1987; Cannell & Fowler, 1965, pp. 11, 25; Tourangeau et al., 2000, pp. 82–88)[9]. In a follow-up study three decades later, Schwarz et al. (2020) conducted a lab experiment with a sample of college students and found

[9] It should be noted, however, that research on memory and forgetting generally examines longer time periods, usually in the order of weeks or years.

that 60% of respondents correctly reproduced their previous answer (to a single item with an 11-point response scale) after a period of about 20 minutes. Revilla and Höhne (2021) even report that 88% of respondents in a probability-based online panel correctly repeated their previous rating of their own political interest (on a fully labeled 5-point scale) within one survey. Yet, this finding may be an artifact of the shorter scale used (5 points versus 10 or 11 points). In contrast to van Meurs and Saris (1990), this more recent research did not find that a longer time interval between the repeated questions reduced respondents' recall ability. However, a study by Alwin (2011) confirmed a slight decline in recall ability over time. The author asked respondents to repeat a list of nouns after it was read out to them. On average, respondents were able to repeat 6 out of 10 nouns immediately afterwards, and 5 out of 10 nouns after 10 to 15 minutes.

Overall, whereas a 20-minute time interval between repeated questions has sometimes been suggested to be the minimum time required to avoid memory effects (Saris et al., 2010), the existing studies show that respondents have a relatively good recollection of their previous answers within the same survey. Thus, based on the previous research on this issue, we expect that respondents will be able to correctly reproduce answers that they have given within the same survey in a majority of cases.

H1: Respondents can correctly repeat their previous answers at the end of the survey in a majority of cases.

In addition, memory effects may be linked to the content of the information that is being recalled. Research suggests that different types of information are forgotten over time at different rates (see, e.g., Bradburn et al., 1987; Tourangeau et al., 2000, pp. 83–86).Moreover, van Meurs and Saris (1990) found that the proportion of respondents who correctly reproduced their previous answer varied across questions. They also found that respondents

were less likely to recall their previous answer when they had been presented with questions on similar topics in the meantime. Furthermore, Alwin (2011) noted that the number of respondents who remembered individual nouns varied greatly. Some nouns were remembered by nearly three times as many respondents as other nouns.

The question type has not been researched specifically in the context of memory effects. However, memory effects described above for questions carrying different information may well be driven by question type effects on respondents' recall ability. Following definitions by Dillman (1978, pp. 80–84), we distinguish three types of questions: beliefs, attitudes, and behaviors. Belief questions measure what people think is true or false, thereby eliciting their perceptions of past, present, or future reality. Attitude questions describe what people like or dislike, requiring them to indicate whether they have positive or negative feelings about an attitudinal object. Finally, behavior questions capture peoples' actions in the past, present, or future (see also Fishbein & Ajzen, 1975, pp. 11–13).

As Fishbein and Ajzen (1975, pp. 13–16) argue, these concepts do not exist independently of each other; they are interlinked. In the authors' conceptualization, beliefs are the most fundamental of these three concepts and are the basis on which attitudes are formed. Attitudes, in turn, influence the formation of behaviors. This implies that beliefs, attitudes, and behaviors lie at different "depths" and differ in terms of both their accessibility and stability. In addition, these different types of information also require respondents to undergo different cognitive retrieval processes (see Tourangeau et al., 2000, Chapters 3, 5 & 6). To answer belief questions, respondents may either retrieve an existing belief about an object (if present) or retrieve relevant information about the object to make a judgement on their factuality. Similarly, attitude questions require respondents to either retrieve an existing evaluation of the object or to retrieve facts and beliefs about the object and form an attitude judgement based on these (Strack & Martin, 1987; Tourangeau et al., 2000, pp. 165–178).

Behavior questions, however, require respondents to retrieve factual information about their own actions. These different paths for reaching the original answer may have an effect on how easily it can be accessed a second time (i.e., recalled) at a later point. Therefore, our second hypothesis is that respondents' recall ability differs across questions of these three types.

H2: Respondents' ability to recall previous answers differs across three types of questions: beliefs, attitudes, and behaviors.

Research has found another correlate of memory effects: the extremeness of respondents' beliefs, attitudes, and behaviors, which is observed through the response itself. Van Meurs and Saris (1990) found that respondents who provided an extreme answer (i.e., selected an endpoint of the response scale) were more likely to correctly reproduce their answer. The authors offered two explanations for this finding: First, extreme opinions are likely more salient and central to respondents (i.e., a sign of strong feelings towards the topic of interest). Salient topics may well be more accessible for respondents and thus more easily retrieved (Schuman & Presser, 1981, pp. 44–49; Tourangeau et al., 1989, 2000, pp. 167–172; Tourangeau & Rasinski, 1988). Second, respondents might find it easier to recall an answer that is (visually) represented by the endpoint of a scale. This leads us to our third hypothesis:

H3: Respondents who provide an extreme answer are more likely to correctly reproduce this answer than respondents who provide moderate answers.

Jaspers et al. (2009) found that the retrospective accounts of respondents who were more certain that they had accurately reproduced their previous answer were indeed more accurate. Thus, respondents seem to know quite well whether they remember their answers. Since human beings like to show consistent behavior (van Kampen, 2019), the certain knowledge of

a previous answer is likely to be used in answering the repeated question. We therefore extend the approach by van Meurs and Saris (1990) and, in addition to observing alleged recall and correct recall, also ask respondents how certain they felt about remembering their previous answer. In our hypothesis, we follow Jaspers et al. (2009)'s findings:

H4: Respondents who express higher certainty about remembering their previous answer are more likely to correctly reproduce it.

The literature on longitudinal panel surveys documents effects of panel experience (i.e., how long respondents have participated in a longitudinal survey) on response behavior. Respondents with higher panel experience tend to answer questions less carefully than respondents with lower panel experience (Couper, 2000; Schonlau & Toepoel, 2015; Toepoel et al., 2008). Our study was conducted in a bimonthly longitudinal survey, for which some panelists had been recruited in 2012 and 2014, while others had been recruited in September 2018, only two months prior to the implementation of our memory effects experiment. This data structure allows us to investigate whether experienced panelists differ in their recall ability from newly recruited panelists. In line with the literature, we expect panelists to become less careful respondents over time. In addition, some research on memory in general has suggested that rare and distinctive events are easier to recall than events that are more typical and similar to other events stored in respondents' memory (Bradburn et al., 1987; Cannell & Fowler, 1965, pp. 12, 26; Tourangeau et al., 2000, p. 91). Experienced respondents may therefore find it harder to remember previous answers than freshly recruited respondents, because the survey is a less memorable event for them. Both would lead to weaker memory effects among experienced panelists.

H5: Inexperienced panelists are more likely to correctly reproduce previous answers than experienced panelists.

4.3 Methods

4.3.1 Study design

We investigate alleged recall (i.e., whether respondents claim that they can remember their answers), correct recall (i.e., whether respondents can correctly reproduce their answers), and recall certainty (i.e., how certain respondents are about correctly reproducing their answers). To test our five hypotheses our experiment applied a between-subject design in which respondents were randomly assigned to one of three question types. Respondents received two questions of their assigned question type (the "test questions") at the beginning of the survey (see Table 4.1). The first experimental group received two belief questions (beliefs condition), the second received two attitude questions (attitudes condition), and the third received two behavior questions (behaviors condition).

Table 4.1. Wording and response scales of the test questions.

Question type	Question stem	Response scale
Beliefs	How likely do you think it is that you can help save the environment by buying environmentally friendly products?	0 not at all likely – 10 extremely likely
Beliefs	How likely do you think it is that you can help prevent climate change by reducing your power consumption?	0 not at all likely – 10 extremely likely
Attitudes	How acceptable would you find it to pay higher prices for environmentally friendly products?	0 not at all acceptable – 10 completely acceptable
Attitudes	How acceptable would you find it to reduce your power consumption to help prevent climate change?	0 not at all acceptable – 10 completely acceptable
Behaviors	How often do you pay attention to the environmental friendliness of the products you buy?	0 never – 10 always
Behaviors	How often do you pay attention to your power consumption in everyday life to prevent climate change?	0 never – 10 always

Note. Questions fielded in German, own translation.

In order to measure the pure effect of question type, we kept the topic of the test questions as similar as possible across question types. For this purpose, we developed pairs of comparable belief, attitude, and behavior questions on the topic of environmental and climate awareness. More specifically, one test question of each type was concerned with environmentally friendly products and the other one with saving energy (see Table 4.1). These questions were based on three questions regarding respondents' beliefs and behaviors on climate change and energy use from the 8th round of the European Social Survey (ESS; European Social Survey, 2016; see Appendix Table A4.1 for the original questions). The ESS questions were adapted to fit the three question types used in our experiment with comparable response scales for each question type. Each test question was presented on a separate survey page with unipolar, item-specific eleven-point response scales in vertical alignment with verbal labels on the endpoints and numeric labels (0–10) on all scale points. The labels of the endpoints were adapted to fit the respective question type. We chose this specific style of response scale as it is commonly used in survey research. For example, the ESS regularly employs endpoint-labelled 0–10 scales for items on several topics, such as social trust, immigration, and left-right placement (see, e.g., European Social Survey, 2016).

The test questions were followed by in-between survey questions that took respondents about 20 minutes to complete. As discussed above, 20 minutes had previously been suggested as a sufficient time interval for question repetitions within one survey (see Saris et al., 2010). In addition, 20 to 25 minutes is the typical overall length for a wave of the online panel in which this experiment was implemented (see below). Using a typical questionnaire length thus provides a realistic assessment of the feasibility of repeating questions within one panel wave and also serves to avoid other issues, such as respondents becoming suspicious or breaking off the survey due to an unusually long wave.

At the end of the survey, respondents received three follow-up questions for each test question in order to determine whether they were able to correctly reproduce their answers to the test questions. First, the test question was again shown to the respondents and they were asked to indicate whether they remembered their answer to it (alleged recall: yes/no). Subsequently, respondents were asked to reproduce their previous answer. By comparing this answer with their answer to the initial test question, we determined whether respondents correctly recalled their answer (i.e., picked the same scale point both times; correct recall: yes/no). Finally, respondents were asked to indicate how confident they were about recalling their previous answer (recall certainty). These follow-up questions were asked for each of the two test questions. Figure 4.1 displays the experimental design (see Appendix Table A4.2 for the wording of the follow-up questions).

The topics of the in-between questions were diverse and covered respondents' perception of political parties and European Union politics. Some of the in-between question pages did not provide respondents with the option to go back to previous questions. Respondents were thus prevented from looking up or changing their previous answers to the test questions after reaching the follow-up questions.

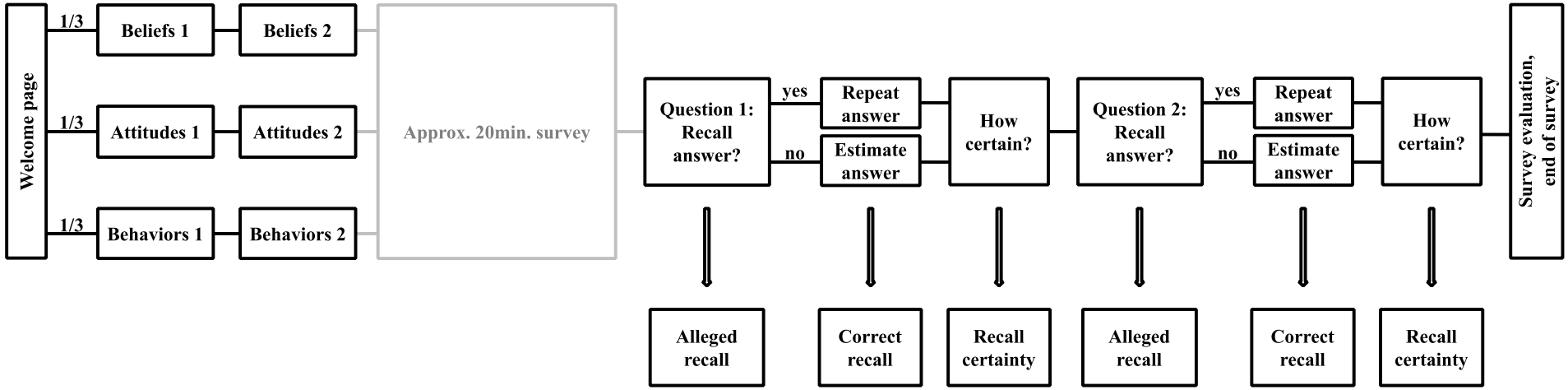

Figure 4.1. Experimental design.

Note. Within each group, the order of the two test questions (i.e. beliefs 1 and 2, attitudes 1 and 2, or behaviors 1 and 2) was randomized across respondents. The order of the follow-up questions reflected the order of the test questions (i.e. if attitudes 2 was shown first, the follow-up questions to attitudes 2 also came before the follow-up questions to attitudes 1)

4.3.2 Data

We implemented this experiment in the November 2018 wave of the German Internet Panel (GIP; Blom et al., 2019). The GIP is a probability-based online panel of the general population of Germany (see Blom et al., 2015). GIP respondents are surveyed bimonthly with each online wave taking about 20 to 25 minutes to complete. The GIP covers a diverse set of topics including national and international politics, policy preferences, and social issues. Our test questions on environmental awareness, therefore, blended well into the GIP context: However, these questions had never been asked in the GIP before, thus avoiding any possible influence of earlier repetitions of the same questions on our experiment. GIP panelists were recruited in 2012, 2014, and 2018 with a random probability sample of people living in private households in Germany and were 16 to 75 years old at the time of recruitment. During the 2012 and 2014 recruitments, respondents without internet access were equipped with devices to facilitate their participation (see Blom et al., 2017). For the 2018 sample, the November 2018 wave was their first regular survey wave. For panelists recruited in 2012 and 2014 it was the 38th and 24th wave, respectively. The sample design thus allows comparisons between new (inexperienced) respondents and those who had been with the GIP for several years.

In total, 4,294 GIP members participated in this wave. 2,119 (49.3%) of these were randomly selected to take part in our experiment. The median age of the respondents was 51 years and 48.4% were female. Overall, 15.3% had a no or a basic school degree (“Hauptschule”), 31.5% a vocational school degree (“Realschule” or equivalent), and 53.2% a high school degree that allows entering higher education (“Fachhochschulreife”, “Abitur”, or equivalent). In terms of the devices used to complete the survey, 23.0% of respondents used a smartphone, the remaining 77.0% used computers or tablets. Finally, 55.8% were experienced panelists, i.e., recruited in 2012 or 2014.

We conducted χ^2-tests to evaluate the effectiveness of the random assignment to the experimental groups (belief, attitude, and behavior conditions). The groups did not significantly differ with respect to the respondents' age (p=.605), gender (p=.256), education (p=.670), device (p=.365), and recruitment sample (p=.981; see Appendix Table A4.3 for the χ^2-statistics). Thus, we confirmed a uniform distribution of the sample across the experimental groups regarding these basic respondent characteristics.

4.3.3 Analytical strategy

Respondents received two test questions and were asked to recall their answers to both of them separately. Therefore, we gained two observations per respondent each consisting of the answer to the test question, the alleged recall (yes/no), the restated answer with which we derived the correct recall (yes/no), and recall certainty (0–10). Since these observations are clustered within respondents, they are not fully independent. To account for the clustered nature of our data in the statistical models, we computed cluster-robust standard errors and included a dummy variable indicating whether a given observation is from the first or second test question presented to a respondent.

We excluded a small proportion of cases due to missing data (1.0% broke off the survey before answering all of the follow-up questions; 0.2% were missing the test questions or follow-ups; 2.6% were missing the socio-demographic controls). Furthermore, the GIP allows respondents to interrupt the survey and return to it at a later point. Since such interruptions may affect respondents' recall, we excluded the affected cases from the analyses (8.2% closed the survey, 0.5% took a long break without closing the survey). Our analyses are thus based on 3,711 observations of 1,858 respondents.

To investigate respondents' recall ability and test our hypotheses, we first have a descriptive look at the number and proportion of alleged and correct recalls overall (*H1*) and separately by question type (*H2*). This also includes the mean recall certainty. We then separately look at

these indicators for cases with extreme answers (*H3*). To further investigate our hypotheses on the role of question type, extreme answers, recall certainty, and panel experience while controlling for socio-demographics, we compute multiple regression models. For the dependent variables alleged recall and correct recall, we compute multiple logistic regression models, and for recall certainty, we compute a linear regression model[10]. In all models, we include the effects of question type (*H2*), extreme responses (*H3*), and panel experience (*H5*). In addition, we add alleged recall as a predictor in the models on recall certainty and correct recall. We further add an interaction of alleged recall and question type to see whether the differences in recall certainty and correct recalls between respondents who said they remembered their answers and those who said they did not also differ across question types. In addition, we add an interaction of panel experience and question type to investigate whether cognitive differences between experienced and inexperienced panelists may be different for different types of questions. Furthermore, recall certainty is added as a predictor in the model on correct recall, to test whether higher self-reported certainty predicts correctly recalling a previous answer (*H4*).

All models control for respondents' socio-demographic characteristics age (in three groups of roughly equal size; <44 years, 44–58 years, >58 years), education (three groups), and gender (two groups). Furthermore, some research suggests that response behavior frequently differs between smartphone respondents and those using computers to answer the survey (Couper & Peterson, 2017; Krebs & Höhne, 2020; Lugtig & Toepoel, 2016; Struminskaya et al., 2015; Tourangeau et al., 2017). Therefore, we add the device respondents used (smartphones versus computers or tablets) as a control variable. In addition, we control for the question order of the test questions, the time respondents spent answering the test questions (server-side

[10] All analyses were computed in Stata 16 using the „logistic" command for logistic regression models or „regress" command for linear regression models respectively. We used the „cluster" option to compute cluster-robust standard errors in order to account for the clustered nature of our data with two observations per respondent.

response time in seconds), and the time between test questions and follow-up questions (server-side in-between time in seconds).

4.4 Results

In a first step, we investigate the proportion of respondents who said that they remembered their previous answer (alleged recall), the proportion of respondents that correctly reproduced their previous answer (correct recall), and the mean recall certainty of respondents (Table 4.2).

Table 4.2. Key indicators on alleged recall, correct recall, and recall certainty by question type.

	Overall		Beliefs		Attitudes		Behaviors	
Observations	3,711		1,229		1,230		1,252	
Alleged recall: yes	3,124	84.2%	1,024	83.3%	1,058	86.0%	1,042	83.2%
Of these:								
Correct recall	1,997	63.9%	558	54.5%	720	68.1%	719	69.0%
Mean certainty[1]		7.6		7.1		7.9		7.7
Alleged recall: no	587	15.8%	205	16.7%	172	14.0%	210	16.8%
Of these:								
Correct recall	258	44.0%	91	44.4%	71	41.3%	96	45.7%
Mean certainty[1]		5.5		5.3		5.6		5.7
Overall correct recall	2,255	60.8%	649	52.8%	791	64.3%	815	65.1%
Overall mean certainty[1]		7.3		6.8		7.6		7.4

Note. Two observations per respondent. Respondents were randomly allocated to one of the question types. [1]On an 11-point scale from 0 “not at all certain” to 10 “absolutely certain”.

Overall, respondents reported to remember their answer in 84.2% of the cases. The differences across question types are small with 83.3% for belief questions, 86.0% for

attitudes, and 83.2% for behaviors. A χ^2-test of differences in alleged recall across the three question types was not significant (p=.098; see Appendix Table A4.4 for χ^2-statistics).

Overall, 60.8% of all observations showed a correct reproduction of the previous answer. Out of the 84.2% where recall was alleged, 63.9% of the recalls were correct. In combination, this means that respondents alleged that they remembered their answer and subsequently gave a correct recollection in 53.8% of all cases. These results support our expectation that after a 20-minute time interval respondents are able to correctly recall their answers in a majority of cases (*H1*). Whereas answers to belief questions were correctly recalled in 52.8% of the cases, the proportions of correct recall for attitude questions and behavior questions are higher with 64.3% and 65.1%, respectively. A χ^2-test showed significant differences in correct recall across question types (p=.000; see Appendix Table A4.4 for χ^2-statistics). These differences seem to be primarily driven by cases in which respondents stated that they remembered their previous answer. In this group we find that 54.5% of the recalls were correct for belief questions, 68.1% for attitude questions, and 69.0% for behavior questions. These results are in line with our hypothesis that the responses to different types of questions are remembered at different rates (*H2*). Looking at cases where respondents stated that they did not remember their previous answer, the differences across question types are smaller. In this group, respondents correctly reproduced their answer in 44.0% of all cases, with 44.4% for belief questions, 41.3% for attitude questions, and 45.7% for behavior questions.

Respondents were relatively confident about remembering their answer. On an 11-point scale from 0 “not at all certain” to 10 “absolutely certain” respondents on average rated their certainty with 7.3. Comparing question types, the mean recall certainty was lowest for belief questions (6.8) and highest for attitude questions (7.6), followed by behavior questions (7.4). A one-way ANOVA showed significant differences in the mean recall certainty across question types (p=.000; see Appendix Table A4.4).

A similar pattern can be observed when only considering cases in which respondents stated that they remembered their answer. In this group, the overall mean recall certainty is 7.6, with 7.1 for belief questions, 7.9 for attitude questions, and 7.7 for behavior questions. Looking at cases where respondents stated that they did not remember their previous answer, the mean recall certainty is considerably lower with an overall mean of 5.5. The differences across question types are small with a mean recall certainty of 5.3 for belief questions, 5.6 for attitude questions, and 5.7 for behavior questions. Overall, respondents seem to be both more likely to correctly reproduce their answer and express higher certainty about remembering it if they alleged that they remembered their answer.

In order to investigate differences across extreme and non-extreme answers, we separately consider alleged recall, correct recall, and recall certainty for cases in which respondents provided extreme answers. 16.6% of all answers were extreme. Table 4.3 reports the descriptive results.

Table 4.3. Key indicators on alleged recall, correct recall, and recall certainty by question type, extreme answers only.

	Overall		Beliefs		Attitudes		Behaviors	
Observations	614	16.6%	188	15.3%	315	25.6%	111	8.9%
Alleged recall: yes	591	96.3%	176	93.6%	305	96.8%	110	99.1%
Alleged recall: no	23	3.7%	12	6.4%	10	3.2%	1	.9%
Overall correct recall	523	85.2%	143	76.1%	281	89.2%	99	89.2%
Overall mean certainty[1]		9.2		8.8		9.3		9.4

Note. [1]On an 11-point scale from 0 "not at all certain" to 10 "absolutely certain".

The proportion of alleged recall is very high for cases with extreme (96.3%). This is significantly higher than for cases with non-extreme answers (p=.000; see Appendix Table A4.5 for the χ^2-statistics). The differences across question types are significant (p=.041; see

Appendix Table A4.6) with 93.6% for belief questions, 96.8% for attitude questions, and 99.1% for behavior questions, respectively.

The correct answer was recalled in 85.2% of the cases; 76.1% for belief questions and 89.2% for both attitude and behavior questions. The differences across question types are again statistically significant (p=.000; see Appendix Table A4.6). The overall proportion of correct recalls is also significantly higher for cases with extreme answers than for cases with non-extreme answers (p=.000; see Appendix Table A4.5).

Furthermore, respondents with extreme answers show very high recall certainty with an overall mean of 9.2. Comparing question types, the mean recall certainty is 8.8 for belief questions, 9.3 for attitude questions, and 9.4 for behavior questions. The differences across question types are again statistically significant (p=.001; see Appendix Table A4.6). The overall mean recall certainty is significantly higher for cases with extreme answers than for cases with non-extreme answers (p=.000; see Appendix Table A4.5). In short, respondents are more likely to allege recall, more likely to provide a correct recall, and express higher recall certainty if they provided an extreme answer. This is in line with our expectation (*H3*).

Next, we computed three multiple regression models in order to model predictors of alleged recall, correct recall, and recall certainty while controlling for socio-demographics and other variables. Table 4.4 displays the results for all three models.

The first Model in Table 4.4 shows predictors of alleged recall. We find no difference in the likelihood of alleged recall across question types. However, we find that alleged recall is more likely if an extreme response was provided, respondents are inexperienced, and respondents are older than 44. Finally, alleged recall was significantly higher for the first set of questions than for the second set.

The second Model in Table 4.4 shows predictors of recall certainty. In contrast to the bivariate analysis, recall certainty does not significantly differ across question types. However, we find that respondents report higher certainty when they provided an extreme answer and when they alleged recall. Panel experience is not significantly related to recall certainty. Finally, respondents report significantly lower recall certainty in their first set of follow-up questions than in the second set.

Table 4.4. Regression models of alleged recall, recall certainty, and correct recall.

	Alleged recall			Recall certainty			Correct recall		
	OR	SE	p	b	SE	p	OR	SE	p
Question type (ref.: behaviors)									
Beliefs	.805	.154	.258	- .435	.306	.155	.857	.209	.526
Attitudes	.848	.167	.403	- .124	.301	.681	.598	.151	.042
Extreme response	5.622	1.278	.000	2.065	.095	.000	2.861	.397	.000
Experienced panelist	.624	.121	.015	.156	.164	.342	.775	.105	.061
Experienced * question type									
Beliefs	1.446	.395	.178	- .171	.240	.476	1.203	.233	.339
Attitudes	1.546	.437	.123	- .189	.218	.386	1.571	.310	.022
Age (ref.: <44 years)									
44–58 years	1.774	.265	.000	.045	.119	.702	.782	.081	.018
>58 years	1.884	.303	.000	- .252	.134	.060	.751	.085	.012
School degree (ref.: basic/none)									
Vocational	1.300	.240	.156	.090	.165	.587	1.357	.172	.016
High school	.964	.165	.829	.201	.155	.196	1.347	.162	.013
Female	.809	.093	.066	- .018	.095	.849	1.322	.107	.001
Smartphone respondent	1.124	.164	.425	- .103	.122	.396	.883	.091	.227
First question	1.555	.103	.000	- .129	.046	.005	.864	.059	.031
Response time	.998	.001	.105	.000	.001	.627	1.001	.001	.467
In-between time	1.000	.000	.228	.000	.000	.086	1.000	.000	.709
Alleged recall				1.821	.190	.000	1.693	.282	.002
Alleged recall * question type									
Beliefs				- .191	.290	.510	.578	.140	.024
Attitudes				.023	.300	.939	1.086	.279	.748
Recall certainty							1.266	.023	.000
Constant	3.707	.913	.000	5.770	.279	.000	.213	.054	.000
Pseudo-R^2McKelvey & Zavoina	.161						.208		
$R^2_{adj.}$				.197					
Observations	3,711			3,711			3,711		

Note. Odds ratios (OR) for logistic regressions (alleged recall and correct recall), b-coefficients for linear regression (recall certainty). Cluster-robust standard errors (SE) account for clustering of observations within respondents.

Finally, the third Model in Table 4.4 presents predictors of correct recall. In contrast to alleged recall, correct recall significantly differs across question types. Responses to attitude questions are recalled significantly less likely correctly than responses to behavior questions[11]. The likelihood of correct recall is significantly higher for extreme responses. This again supports our hypothesis that respondents can remember extreme answers more easily (*H3*). Investigating the effects panel experience and its interaction with question type, we see that experienced respondents are more likely to correctly reproduce their answers to attitude questions (but they do not differ from inexperienced respondents overall). Thus, while we find some connection between question type and panel experience, these results do not match our expectation that inexperienced respondents have a higher recall ability than experienced respondents. Thus, the evidence does not support our hypothesis (*H5*). Furthermore, correct answers are more likely recalled if recall is alleged. In line with our descriptive results, this difference is less pronounced for belief questions. Overall, these results support our hypothesis that recall ability differs by question type (*H2*). We also find a positive association between recall certainty and correct recall, which is in line with our hypothesis that respondents are more certain about remembering their answer when they correctly recall it (*H4*). Finally, correct recalls are more likely if respondents are under 44 years old, have a higher than basic school degree, are female, and for their second set of follow-up questions.

4.5 Discussion and conclusion

The aim of this experimental study was to investigate respondents' ability to recall previous answers in a probability-based online panel. More specifically, we looked at alleged recall (i.e., whether respondents say that they remember their previously given answer), correct

[11] Selecting belief questions as the reference in a separate model (not shown) revealed that they do not significantly differ from attitude questions.

recall (i.e., whether respondents pick the same scale point as previously selected), and recall certainty (i.e., how certain respondents are about remembering their previously given answer). For this purpose, we randomly assigned respondents to one out of three experimental groups that varied the question type (beliefs, attitudes, and behaviors).

Overall, we found that respondents claimed to remember their previous answer in 84.2% of all cases and were able to correctly repeat it in 63.9% of these. Moreover, respondents who indicated that they did not remember their previous answer correctly repeated it in 44.0% of the cases. As argued by van Meurs and Saris (1990), the main reason for this phenomenon is that many respondents are not likely to change their mind within a short period of time. Thus, there is a good chance that some respondents give the same answer again without this being due to a memory effect. Some respondents may also pick the correct answer by chance. However, respondents who said that they remembered their previous answer were considerably more likely to provide a correct recall than respondents who said that they did not remember their answer (63.9% vs. 44.0%). This 19.9 percentage point difference in correct recalls is smaller than the 34 percentage points found by van Meurs and Saris (1990), but similar to the 17 percentage points found by Schwarz et al. (2020). The authors of both studies used this difference as an approximation for the proportion of respondents for whom a memory effect might occur (i.e., respondents who repeated their answer correctly due to memory, rather than due to a stable opinion or correct guessing).

We found that the proportion of correctly recalled answers is high, especially when considering the relatively long response scales with 11 points that we used in this study and the strict definition of correct recall as picking the exact same scale point (see Höhne, 2021 for a discussion of less strict definitions of correct recall). Since so many respondents were able to correctly recall their answers, we conclude that a time interval of 20 minutes is insufficient to reliably prevent memory effects. In addition, differing rates of remembering

previous answers are linked to different question types, the answer extremeness, recall certainty, panel experience, age, school education, and gender. Comparisons of repeated survey measurements across question types or groups of respondents are thus likely to be biased due to memory effects. In light of these findings, we recommend that researchers use question repetitions within the same survey with caution.

Researchers have only recently begun integrating memory effects into the wider literature on measurement error in surveys and the cognitive response process (see, for instance, Rettig & Blom, 2021). We aimed to contribute to this integration process with our literature review. However, given the scarcity of research in this field, investigating memory effects offers further opportunities for future research. Most research on memory effects has thus far focused on whether respondents remember their previous answer. However, more research is needed to determine whether and how later answers will actually be influenced by this in an undesired way (i.e., respondents giving a different answer than they would have if they did not remember their previous answer). Consequently, a systematic investigation of how answers to a repeated survey measurement differ across respondents who can and those who cannot remember their answer to a previous iteration of the same question would be an interesting and worthwhile avenue for future research.

In addition, an influence of memory on later answers may not necessarily be undesirable in all cases. A simple repetition of the previous answer or inflated consistency across answers would be a source of measurement error. However, respondents may also use their memory of the previous answer to carefully consider whether and how their opinions have changed since the last time the question was asked (Rettig & Blom, 2021). Similar to dependent interviewing, where the previous answer is purposefully presented to respondents to minimize measurement error, this may even lead to a less biased response than a completely independent second response process. A way to distinguish these effects as well as an

investigation into which effect occurs more commonly in survey practice would thus be very valuable to survey researchers.

Furthermore, there may be other influences on respondents' ability to remember previous answers that were not investigated in this study, such as the question topic and how strongly respondents feel about it or the survey mode. The visual presentation of the response scale in a self-administered online survey may, for example, serve as a recall cue that makes it easier for respondents to recall their previous answers. In addition, memorizing which scale point one selected is easier when there are, for example, only 5 instead of the 11 scale points we used in this study (Höhne, 2021). Picking the correct scale point by chance is also more likely on a shorter response scale. Thus, our results might not be generalized to scales of different lengths.

Finally, this study adds to an emerging body of literature that suggests respondents are frequently able to remember and correctly repeat the exact answers they gave to earlier questions within one survey (Höhne, 2021; Revilla & Höhne, 2021; Schwarz et al., 2020; van Meurs & Saris, 1990). However, much less research has dealt with memory effects over longer time periods, which are more common for repeated survey measurements to observe change over time in longitudinal panel surveys. As we can generally expect memory to decline over time, respondents may be much less likely to remember their answers after several weeks or months. Further research on memory effects in longitudinal settings is therefore needed to guide researchers in establishing reasonable time intervals for repeated survey measurements.

References

Alwin, D. F. (2007). *Margins of Error. A Study of Reliability in Survey Measurement*. John Wiley & Sons.

Alwin, D. F. (2010). How Good is Survey Measurement? Assessing the Reliability and Validity of Survey Measures. In P. V. Marsden & J. Wright (Eds.), *Handbook of Survey Research* (2nd ed., pp. 405–434). Emerald Group Publishing.

Alwin, D. F. (2011). Evaluating the Reliability and Validity of Survey Interview Data Using the MTMM Approach. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), *Question Evaluation Methods* (pp. 265–295). John Wiley and Sons.

Blom, A. G., Cornesse, C., Friedel, S., Krieger, U., Fikel, M., Rettig, T., Wenz, A., Juhl, S., Lehrer, R., Möhring, K., Naumann, E., & Reifenscheid, M. (2020). High-Frequency and High-Quality Survey Data Collection: The Mannheim Corona Study. *Survey Research Methods*, *14*(2), 171–178. https://doi.org/10.18148/srm/2020.v14i2.7735

Blom, A. G., Fikel, M., Friedel, S., Höhne, J. K., Krieger, U., Rettig, T., Wenz, A., & SFB 884 ´Political Economy of Reforms´ Universität Mannheim. (2019). *German Internet Panel, Wave 38 (November 2018)*. GESIS Data Archive, Cologne. ZA6958 Data file Version 1.0.0. https://doi.org/10.4232/1.13391

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting Up an Online Panel Representative of the General Population: The German Internet Panel. *Field Methods*, *27*(4), 391–408. https://doi.org/10.1177/1525822X15574494

Blom, A. G., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U., & Bossert, D. (2017). Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence From the German Internet Panel. *Social Science Computer Review*, *35*(4), 498–520. https://doi.org/10.1177/0894439316651584

Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. *Science*, *236*(4798), 157–161. https://doi.org/10.1126/science.3563494

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multi-Trait-Multimethod Matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin Company.

Cannell, C. F., & Fowler, F. J., Jr. (1965). Comparison of Hospitalization Reporting in Three Survey procedures. *National Center for Health Statistics. Vital Health Stat*, *2*(8).

Couper, M. P. (2000). Web Surveys. *Public Opinion Quarterly*, *64*(4), 464–494. https://doi.org/10.1086/318641

Couper, M. P., & Peterson, G. J. (2017). Why Do Web Surveys Take Longer on Smartphones? *Social Science Computer Review*, *35*(3), 357–377. https://doi.org/10.1177/0894439316629932

Dillman, D. A. (1978). *Mail and Telephone Surveys: The Total Design Method*. Wiley and Sons.

Dimitrov, D. M., & Rumrill, P. D. (2003). Pretest-Posttest Designs and Measurement of change. *Work*, *20*(2), 159–165.

Eggs, J., & Jäckle, A. (2015). Dependent interviewing and sub-optimal responding. *Survey Research Methods*, *9*(1), 15–29. https://doi.org/10.18148/srm/2015.v9i1.5860

European Social Survey. (2016). *ESS Round 8 Source Questionnaire*. London: ESS ERIC Headquarters c/o City University London. https://www.europeansocialsurvey.org/docs/round8/fieldwork/source/ESS8_source_questionnaires.pdf

Fishbein, M., & Ajzen, I. (1975). *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research.* Addison-Wesley.

Höhne, J. K. (2021). New Insights on Respondents' Recall Ability and Memory Effects When Repeatedly Measuring Political Efficacy. *Quality and Quantity*. https://doi.org/10.1007/s11135-021-01219-2

Hoogendoorn, A. (2004). A Questionnaire Design for Dependent Interviewing that Addresses the Problem of Cognitive Satisficing. *Journal of Official Statistics*, *20*(2), 219–232.

Jäckle, A. (2008). Dependent Interviewing: Effects on Respondent Burden and Efficiency of Data Collection. *Journal of Official Statistics*, *24*(3), 411–430.

Jäckle, A., & Eckman, S. (2020). Is That Still the Same? Has that Changed? On the Accuracy of Measuring Change with Dependent Interviewing. *Journal of Survey Statistics and Methodology*, *8*(4), 706–725. https://doi.org/10.1093/jssam/smz021

Jaspers, E., Lubbers, M., & De Graaf, N. D. (2009). Measuring Once Twice: An Evaluation of Recalling Attitudes in Survey Research. *European Sociological Review*, *25*(3), 287–301. https://doi.org/10.1093/esr/jcn048

Krebs, D., & Höhne, J. K. (2020). Exploring Scale Direction Effects and Response Behavior Across Pc and Smartphone Surveys. *Journal of Survey Statistics and Methodology*, 1–19. https://doi.org/10.1093/jssam/smz058

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Lugtig, P., & Lensvelt-Mulders, G. J. L. M. (2014). Evaluating the Effect of Dependent Interviewing on the Quality of Measures of Change. *Field Methods*, *26*(2), 172–190. https://doi.org/10.1177/1525822X13491860

Lugtig, P., & Toepoel, V. (2016). The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey: Effects on Survey Measurement Error. *Social Science Computer Review*, *34*(1), 78–94. https://doi.org/10.1177/0894439315574248

Lynn, P. (2009). Methods for Longitudinal Surveys. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (pp. 1–19). John Wiley & Sons. https://doi.org/10.1002/9780470743874.ch1

Rettig, T., & Blom, A. G. (2021). Memory Effects as a Source of Bias in Repeated Survey Measurement. In A. Cernat & J. W. Sakshaug (Eds.), *Measurement Error in Longitudinal Data* (pp. 3–18). Oxford University Press.

Revilla, M., & Höhne, J. K. (2021). Repeatedly Measuring Political Interest: Can we Reduce Respondent' Recall Ability and Memory Effects in Surveys Using Memory Interference Tasks? *International Journal of Public Opinion Research*, *33*(3), 678–689. https://doi.org/10.1093/ijpor/edaa035

Saris, W. E., & Gallhofer, I. N. (2014). *Design, Evaluation, and Analysis of Questionnaires for Survey Research* (2nd ed.). John Wiley and Sons.

Saris, W. E., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, *4*(1), 45–59. https://doi.org/10.18148/srm/2010.v4i1.2682

Saris, W. E., Satorra, A., & Coenders, G. (2004). A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design. *Sociological Methodology*, *34*(1), 311–347. https://doi.org/10.1111/j.0081-1750.2004.00155.x

Schonlau, M., & Toepoel, V. (2015). Straightlining in Web Survey Panels over Time. *Survey Research Methods*, *9*(2), 125–137. https://doi.org/10.18148/srm/2015.v9i2.6128

Schuman, H., & Presser, S. (1981). *Questions and Answers in Attitude Surveys. Experiments on Question Form, Wording, and Context*. Academic Press.

Schwarz, H., Revilla, M., & Weber, W. (2020). Memory Effects in Repeated Survey Questions. Reviving the Empirical Investigation of the Independent Measurements Assumption. *Survey Research Methods*, *14*(3), 325–344. https://doi.org/10.18148/srm/2020.v14i3.7579

Strack, F., & Martin, L. L. (1987). Thinking, Judging, and Communicating: A Process Account of Context Effects in Attitude Surveys. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social Information Processing and Survey Methodology. Recent Research in Psychology.* (pp. 123–148). Springer.

Struminskaya, B., Weyandt, K., & Bosnjak, M. (2015). The Effects of Questionnaire Completion Using Mobile Devices on Data Quality. Evidence from a Probability-based General Population Panel. *Methods, Data, Analyses*, *9*(2), 261–292. https://doi.org/10.4232/1.12245.

Toepoel, V., Das, M., & Van Soest, A. (2008). Effects of Design in Web Surveys: Comparing Trained and Fresh Respondents. *Public Opinion Quarterly*, *72*(5), 985–1007. https://doi.org/10.1093/poq/nfn060

Tourangeau, R., Maitland, A., Rivero, G., Sun, H., Williams, D., & Yan, T. (2017). Web Surveys by Smartphone and Tablets. Effects on Survey Responses. *Public Opinion Quarterly*, *81*(4), 896–929. https://doi.org/10.1093/poq/nfx035

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin*, *103*(3), 299–314. https://doi.org/10.1037/0033-2909.103.3.299

Tourangeau, R., Rasinski, K. A., Bradburn, N., & D'Andrade, R. (1989). Belief Accessibility and Context Effects in Attitude Measurement. *Journal of Experimental Social Psychology*, *25*(5), 401–421. https://doi.org/10.1016/0022-1031(89)90030-9

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

van Kampen, H. S. (2019). The principle of consistency and the cause and function of behaviour. *Behavioural Processes*, *159*, 42–54. https://doi.org/10.1016/j.beproc.2018.12.013

van Meurs, A., & Saris, W. E. (1990). Memory Effects in MTMM Studies. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait Multimethod Studies* (pp. 134–146). North Holland.

Appendix

Table A4.1. Original questions from ESS Round 8.

Question stem	Response scale
If you were to buy a large electrical appliance for your home, how likely is it that you would buy one of the most energy efficient ones?	00 Not at all likely – 10 Extremely likely
[…] In your daily life, how often do you do things to reduce your energy use?	01 Never 02 Hardly ever 03 Sometimes 04 Often 05 Very often 06 Always
How likely do you think it is that limiting your own energy use would help reduce climate change?	00 Not at all likely – 10 Extremely likely

Note. See European Social Survey, 2016, pp. 30, 37.

Table A4.2. Wording and response scales of the follow-up questions.

Question type	Question stem	Response scale
Alleged recall (if first follow-up)	Earlier we asked you the following question: *[TEST QUESTION TEXT]* Can you recall your exact answer to it?	yes / no
Alleged recall (if second follow-up)	We also asked you the following question: *[TEST QUESTION TEXT]* Can you recall your exact answer to it?	yes / no
Correct recall (if alleged recall: yes)	Please indicate what your answer was.	same scale as test question
Correct recall (if alleged recall: no)	Even if you do not exactly recall: Please estimate, what your answer was.	same scale as test question
Recall certainty	How certain are you about your answer?	0 not at all certain – 10 absolutely certain

Note. Questions fielded in German, own translation.

Table A4.3. Chi-squared tests of differences across experimental groups.

	χ^2	df	p
Age	23.50	26	.605
Gender	2.73	2	.256
School degree	9.38	12	.670
Device	2.01	2	.365
Recruitment sample	.42	4	.981

Table A4.4. Chi-squared tests and one-way ANOVA for differences across question types.

Chi-squared tests for differences on…	χ^2	df	p
Alleged recall	4.65	2	.098
Correct recall	48.97	2	.000
ANOVA for differences on…	**F**	**df**	**p**
Mean recall certainty	30.23	2	.000

Table A4.5. Chi-squared tests and one-way ANOVA for differences between extreme and non-extreme answers.

Chi-squared tests for differences on…	χ^2	df	p
Alleged recall	80.52	1	.000
Correct recall	183.94	1	.000
ANOVA for differences on…	F	df	P
Mean recall certainty	487.61	1	.000

Table A4.6. Chi-squared tests and one-way ANOVA for differences across question types (extreme answers only).

Chi-squared tests for differences on…	χ^2	df	p
Alleged recall	6.40	2	.041
Correct recall	17.83	2	.000
ANOVA for differences on…	F	df	p
Mean recall certainty	6.64	2	.001

5. Memory Effects in Online Panel Surveys: Investigating Respondents' Ability to Recall Responses from a Previous Panel Wave[12]

Abstract

If respondents recognize repeated survey questions and remember their previous responses, this can result in measurement error. Most studies to date that have investigated respondents' recall of their prior answers have done so in the context of repeated measurements within one cross-sectional survey. The present study extends this research to a longitudinal panel context by investigating whether respondents remember their previous responses to different types of survey questions (beliefs, attitudes, and behaviors) from a previous wave in a probability-based online panel in Germany. We find evidence that some respondents remember their responses from a previous panel wave even after four months, but at a considerably lower rate than previous studies found within a single cross-sectional survey. Respondents who could not remember their response were most commonly off by only a single scale point. Respondents remembered their responses to different types of questions at different rates and were more likely remember an extreme response. Female respondents were more likely to remember their responses, but we find no link to age, education, perceived response burden, survey enjoyment or online panel experience. As respondents could not remember their previous responses in most cases and we find little evidence for a systematic variation of memory effects across groups of respondents, we conclude that the potential for measurement error due to memory effects across panel waves is low after four or more than four months.

[12] This chapter is joint work with Bella Struminskaya. It is at the time of writing under peer review: Rettig, T. & Struminskaya, B. (under review). Memory Effects in Online Panel Surveys: Investigating Respondents' Ability to Recall Responses from a Previous Panel Wave.

Keywords

extreme responses, measurement error, memory effects, online panel, repeated measurement

Acknowledgements

This article uses data from the waves 38–40 of the German Internet Panel (GIP; DOIs: 10.4232/1.13391; 10.4232/1.13585; 10.4232/1.13463), (Blom et al., 2019, 2020a, 2020b). The following variable used in our analyses is available via the On-Site Data Access (ODA) facilities of the GIP for data protection reasons: device. A study description can be found in Blom et al. (2015). The GIP is part of the Collaborative Research Center 884 (SFB 884) funded by the German Research Foundation (DFG) – Project Number 139943784 – SFB 884. The authors gratefully acknowledge the valuable feedback from Annelies G. Blom and Jan Karem Höhne in designing and implementing this experiment.

5.1 Introduction

Repeated measurements of the same survey questions from the same respondents at different points in time have several applications in survey research. In longitudinal studies, survey questions are commonly repeated to the same respondents at set time intervals in order to measure change over time on the individual level (see, e.g., Lynn, 2009). In addition, repeated measurements are used in pretest-posttest experimental designs, in which the effectiveness of a treatment is evaluated by taking identical measurements before and after its administration (Campbell & Stanley, 1966; Dimitrov & Rumrill, 2003). Furthermore, repeated measurements are used to evaluate the measurement quality of survey instruments, such as in test-retest or quasi-simplex designs to estimate reliability (Alwin, 2007, 2010, 2011), or in multitrait-multimethod (MTMM) designs to estimate reliability and validity (Campbell & Fiske, 1959; Saris et al., 2004; Saris & Gallhofer, 2014).

All of these applications for repeated measurements rely to some extent on the assumption that the different measurements are independent in the sense that respondents undergo the full cognitive response process each time and that responses to the repetitions are not influenced by the previous response (Rettig & Blom, 2021). The response process includes four steps: comprehension of the question, retrieval of the relevant information from memory, formation of a judgement based on the retrieved information, and selection of a fitting response (Tourangeau et al., 2000). A violation of the assumption that this process is fully and independently undergone for repeated questions may occur when respondents recognize that they were asked the same question before, remember their previous response, and use this information in their processing of the repeated question (Struminskaya & Bosnjak, 2021).

First, instead of merely comprehending the question when it is asked for the subsequent times, respondents may also recognize this question and remember some of the cues from the previous time that they searched their memory, thus, retrieving not only the information they

need to form a new judgement, but also the information that they used for their previous response. Upon retrieving those cues, respondents might form a judgement that is not independent of the judgement they formed previously. Second, respondents might retrieve their previous response from memory and use this response as a basis for reflection against which they evaluate their current stance in forming a judgement to the repeated question. Third, respondents may also accept their previous response as a preexisting judgement and simply repeat it without further reflection (Rettig & Blom, 2021).

Thus, when the exact previous response cannot be remembered, recognizing the question and remembering the act of previously answering it may provide respondents with additional retrieval cues. These cues may make it easier for respondents to retrace their previous retrieval and judgement processes (and thus arrive at the same response) instead of forming a fully independent judgement.

Either option presumably leads to more consistent (or even identical) responses, which would in turn lead to an underestimation of changes over time, an underestimation of treatment effects, or an overestimation of reliability and, thus, also an overestimation of measurement quality (Alwin, 2011; Rettig & Blom, 2021). The resulting measurement error, commonly referred to as a memory effect, has recently gained renewed interest from researchers (Höhne, 2021; Rettig et al., 2019; Rettig & Blom, 2021; Revilla & Höhne, 2021; Schwarz et al., 2020). However, most studies that systematically investigate whether respondents remember their previous responses have done so within one survey wave. These studies offer valuable insight for study designs that incorporate measurement repetitions within a single survey wave, such as pretest-posttest experiments and MTMM designs. In this study, we add to this emerging body of research by investigating respondents' ability to remember their responses from a previous panel wave in a longitudinal setting.

5.2 Background

The issue of memory effects in repeated survey measurements has been investigated in relatively few studies. Notably, van Meurs and Saris (1990) investigated how well respondents remembered their previous responses both within a single survey (with a period of about 9 minutes between the initial questions and repetitions) and after a period of two weeks. The authors found that about 70% of respondents correctly repeated their answers within one survey, and about 40% did so after two weeks. Expanding on this study, some recent studies have investigated memory effects in repeated survey measurements within the context of one survey. Rettig et al. (2019) report that respondents correctly repeated their previous responses in 61% of cases after a period of about 20 minutes. This is in line with Schwarz et al. (2020), who found that 60% of respondents did so. Revilla and Höhne (2020) reported that in their experiment, an even larger proportion of respondents (88%) correctly repeated a previous response within one interview. These studies give a relatively consistent indication that a large portion of respondents remember their responses within one survey.

Beyond cross-sectional surveys, respondents' ability to recall previous responses has seldom been examined. Human memory in general declines over time (i.e., people tend to forget information as time passes; see, e.g., Bradburn et al., 1987; Cannell & Fowler, 1965; Tourangeau et al., 2000). Thus, we can expect that respondents will be less likely to remember previous responses after weeks or months than after a few minutes. In line with this expectation, respondents were less able to correctly repeat their response after two weeks than after 9 minutes (van Meurs & Saris, 1990). However, because a sizeable proportion of respondents still remembered their answers, a period of two weeks may not be long enough to reliably prevent memory effects. In contrast, McKelvie (1992) found that instructing respondents to use their previous response – encouraging respondents to repeat a pervious response or discouraging its use in forming the later response – had negligible effects on the test-retest reliability of repeated measurements after 17–25 days. However, McKelvie (1992)

also notes that practice effects from the previous measurement were present in the repetition, which Salinsky et al. (2001) also found to persist after 12–16 weeks.

For a considerably longer time frame, Jaspers et al. (2009) found that when asked for retrospective accounts of their attitudes after more than 10 years, respondents adjusted their recollection to their current attitudes rather than adjusting their current attitude to their recollection. For example, those who presently held a more favorable attitude towards homosexuality tended to also falsely report having held a more favorable view over 10 years prior: their attitudes were in fact not as favorable at that time as they reported them to be. This finding indicates that after such a long time, there no longer seems to be any influence of previous responses. Alwin (2011) suggests that an interval of two years between measurement repetitions would be sufficient to rule out the occurrence of memory effects.

Noteworthy, memory effects do not necessarily need to be negative. Respondents' answers can contain measurement error, and reminding respondents about their answers to the same questions asked closer to the event can simplify the respondents' task and increase data quality by reducing negative effects such as misclassification of events or omission of events in surveys that ask about life events. Panel surveys implement special techniques such as dependent interviewing to aid the respondents and avoid either spurious reports of change or spurious reports of stability (Jäckle, 2009). For certain research questions, reminding the respondents about their past situation and asking, "Is it still the case?" produces most accurate data on change and stability as Jäckle and Eckman (2020) demonstrated through the use of survey data and administrative records.

Whether remembering the answer is positive or negative for survey data quality, it remains an empirical question how well respondents remember their responses from previous panel waves for different types of questions. With this paper, we investigate panel survey respondents' reported and actual recall after four months.

5.3 Hypotheses

Some respondents may correctly repeat their previous response without remembering it, either because they have not changed their mind in the meantime – and thus their response itself has not changed – or by chance (van Meurs & Saris, 1990). Meurs and Saris (1990) separately investigated the responses of respondents who reported that they remembered their response and those who said that they did not. The proportion of respondents who correctly repeat their response – despite (self-reportedly) not remembering it – is used as an approximation for correct repetitions due to chance or stable opinion. In turn, respondents who claim to remember their response may be misremembering it in some cases and thus not be able to correctly repeat it. However, van Meurs and Saris (1990) suggest that the difference in correctly repeated responses between these two groups can be used as an approximation for correct repetitions due to memory effects. Following this approach, if no memory effects were present, we would expect to see no difference in correctly repeated responses between respondents who claim to remember their response and those who do not (i.e., all correct repetitions can be explained by stable opinion or answering the same by chance). In turn, if memory effects are present, we hypothesize that:

H1: Respondents are more likely to correctly repeat their responses from a previous panel wave when they claim that they remember it.

Different types of information are forgotten at different rates (Bradburn et al., 1987; Tourangeau et al., 2000). Respondents may thus remember their answers to different types of questions (i.e., beliefs, attitudes, and behaviors) to differing degrees. Conceptually, beliefs deal with respondents' perception of reality, describing what they think is true or false (e.g., "Do you think that making abortions legal everywhere in the United States will lead to an actual decrease in our country's population?"; Dillman, 1978: 82). Based on their beliefs,

respondents form attitudes, which describe whether people have positive or negative feelings towards an object or issue, describing what they like or dislike (e.g., “In general, how do you feel about nationwide legalization of abortion in the United States?”; Dillman, 1978: 81). Behaviors are then formed on the basis of attitudes and describe respondents’ actions (e.g., “Are you currently taking birth control pills?”; Dillman, 1978: 84; see also Fishbein & Ajzen, 1975).

These different types of questions ask respondents for different types of information, which require respondents to undergo different cognitive processes in retrieving the information and answering the question (see Tourangeau et al., 2000). Belief questions require respondents to retrieve facts and beliefs about the topic to form a judgement based on this information or retrieve an existing judgement. To answer an attitudinal question, respondents either retrieve their feelings towards the object (i.e., an existing attitude judgement) or their beliefs and factual information to form an attitude judgement (see also Strack & Martin, 1987; Tourangeau et al., 1989). To answer behavior questions, respondents need to recall their own actions, which is a more overt, directly accessible type of factual information.

These differences in the cognitive processes respondents undergo to reach a response as well as differences in the stability over time and accessibility of these different types of information may translate to differences in how well respondents can remember their response at a later point (Rettig et al., 2019). In the short-term when the underlying information is unlikely to change in the meantime, the more accessible information should be more easily reproduceable for respondents. In line with this expectation, Rettig et al. (2019) found that within one survey, the proportion of correctly repeated responses was higher for behavior and attitude questions than for belief questions. However, over a longer time interval the underlying information is more likely to change between the repeated questions, which

may additionally influence the ability to correctly repeat previous responses across question types. We thus hypothesize:

H2: The likelihood of respondents remembering their previous responses differs across questions on beliefs, attitudes, and behaviors.

Similarly, a response that was based on a stronger and more salient opinion has also been suggested to be easier to remember. Some authors suggest that more salient (and thus accessible) information is more likely to be retrieved during the cognitive response process and, therefore, more likely to be used in forming the response (Schuman & Presser, 1981; Tourangeau & Rasinski, 1988). A stronger and more salient opinion may thus also be easier to repeat at a later point. Several studies have found that respondents are more likely to remember their response if they originally chose an extreme response option (i.e., an endpoint of the response scale; Rettig et al., 2019; van Meurs & Saris, 1990)[13], which reflects a strong opinion. We thus expect:

H3: Respondents are more likely to remember extreme responses than non-extreme responses.

As the occurrence of memory effects is tied to a previous response being present in respondents' memory, a link between memory effects and individual memory capacity and cognitive ability may exist. Such a link would potentially make the resulting measurement error more problematic, because it could systematically vary across groups of respondents, such as different levels of education or across different age groups. Within one survey, Rettig et al. (2019) find that younger and higher educated respondents were more likely to remember

[13] However, Revilla and Höhne (2021) do not find this effect and Höhne's results (2021) even suggest that respondents were less likely to remember extreme responses.

their responses. Höhne (2021) finds the same effect of age, but no effect for education, while Revilla and Höhne (2021) find an effect for high education but not for age. Schwarz et al. (2020) find no effect for age and for holding a university degree, albeit in a sample of university students aged 19 to 29 years, which may as a whole be considered exclusively young and highly educated. While these studies present mixed findings, they give some indication that better memory of previous responses may be linked to younger age and higher education levels. We, therefore, expect the following:

H4.1: Younger respondents are more likely to remember their previous responses than older respondents.

H4.2: Respondents with higher education are more likely to remember their previous responses than respondents with lower educational levels.

Furthermore, it seems plausible that a more accessible and persistent memory of a response may be formed when respondents process a question more thoroughly. In other words, respondents may be less likely to remember a response if they generated it by only superficially undergoing the cognitive response process (i.e., if they satisficed; see Krosnick, 1991). Because respondents presumably use such cognitive shortcuts to alleviate the cognitive effort and burden of answering a survey (Krosnick, 1991; Yan et al., 2020), respondents who perceive answering a survey as more burdensome and less enjoyable may less easily remember their responses. We therefore hypothesize that such a link between respondents' ability to recall previous responses, their perceived response burden, and their enjoyment of the survey exists:

H5.1: Respondents who perceived the previous panel wave as more burdensome are less likely to remember their responses.

H5.2: Respondents who perceived the previous panel wave as more enjoyable are more likely to remember their responses.

Finally, whether respondents remember their previous responses may also potentially be linked to other factors typically associated with more superficial and less careful response behavior. For instance, in a longitudinal survey context, respondents who have participated for a longer time and therefore have a higher level of panel experience have been shown to interact with a questionnaire differently and answer questions less carefully than less experienced respondents (Couper, 2000; Schonlau & Toepoel, 2015; Toepoel et al., 2008). In addition to this less thorough processing of questions and their responses, the act of answering a questionnaire may be more memorable to somebody who has not done it many times before. Research on human memory suggests that individual events are harder to recall when they are similar to other events stored in a person's memory, whereas more unique and rare events are easier to remember (Bradburn et al., 1987; Tourangeau et al., 2000). In the one study to examine this so far, Rettig et al. (2019) did not find the expected effect of newly recruited respondents being more likely to remember their responses than experienced respondents within one survey. However, it is unclear whether an effect may exist for a longer time period between repetitions. Outside of the context of the same survey, respondents who have answered many panel waves may be less able to remember their responses from a specific wave than respondents who have only participated in a few waves. Thus, we hypothesize:

H6: Newly recruited respondents are more likely to remember their responses than experienced respondents.

5.4 Data and method

This study uses data from an experiment fielded in the November 2018 and March 2019 waves of the German Internet Panel (GIP; Blom et al., 2019, 2020b), as well as information about whether respondents participated in the wave in-between these two (January 2019; Blom et al., 2020a). The GIP is a probability-based online panel of the German population recruited from persons living in private households in Germany aged 16 to 75 years at the time of recruitment (Blom et al., 2015, 2017). Respondents of the GIP are surveyed online bimonthly, with a total of 6 waves per year. Each wave takes about 20 to 25 minutes to complete. Respondents receive conditional incentives of 4€ for each wave in which they participate and a bonus of 10€ for participating in all 6 waves in a year or 5€ for participating in 5 out of 6 waves. Incentives are paid out twice a year, and respondents can choose between a bank transfer, an Amazon voucher, or a donation to a charitable organization. The November 2018 wave was the 38th wave of the GIP overall, but the first regular wave of a newly recruited 2018 refresher sample, thus allowing for comparisons across freshly recruited respondents and experienced respondents who had been panelists for several years.

5.4.1 Experimental design

At the beginning of GIP wave 38 (November 2018), respondents received two questions on the topic of environmental awareness (the "test questions"; see Figure 5.1). These questions were placed at the beginning of the questionnaire. Respondents were randomly assigned to receive these two questions either in the form of belief, attitude, or behavior questions with each respondent receiving two questions of the same type in a randomized order. All of the test questions were presented to respondents on individual pages in the online questionnaire with 11-point response scales. The response scales were unipolar, numerically labelled 0 to 10 on all scale points, and had verbal labels on their endpoints. The endpoint labels were adapted

to fit the three respective question types (Appendix Table A5.1 provides English translations of all questions and response scales).

At the beginning of wave 40 (March 2019; 4 months after wave 38), all respondents who had participated in wave 38 received a set of follow-up questions: First, they were presented with the first question they had answered in wave 38 and asked to indicate whether or not they remembered their response to it (claimed recall: yes/no). Depending on whether or not respondents claimed that they remembered their response, respondents were then either asked to repeat their response or to give their best estimate. Comparing this repeated response to respondents' original response from wave 38 allows us to examine whether or not respondents repeated their previous response correctly (correct recall: yes/no). Finally, because some studies have suggested that expressing high certainty about remembering a previous response is a good predictor of remembering it (Jaspers et al., 2009; Rettig et al., 2019), respondents were asked to indicate how certain they felt about remembering their response (recall certainty). The same set of follow-up questions was then repeated for the second question respondents answered in wave 38.

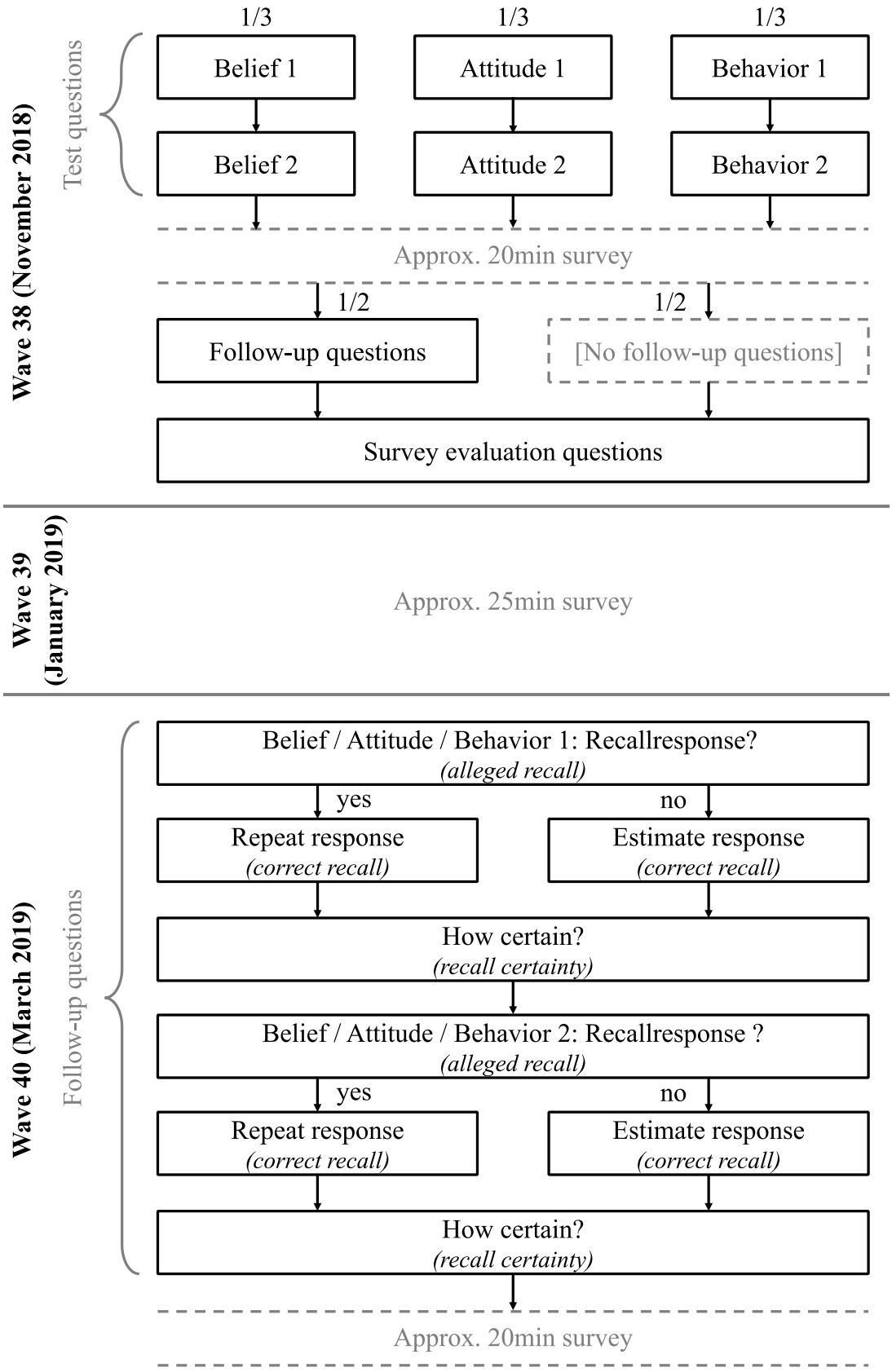

Figure 5.1. Illustration of the experimental design.

Note. Follow-up questions in wave 38 were the same as in wave 40. Test questions were shown in randomized order, follow-up questions matched the order of the test questions. See Appendix Table A5.1 for English translations of all questions and response scales.

About half of all respondents in this experiment had previously been randomly chosen to receive the same set of follow-up questions at the end of wave 38, approximately 20 minutes after they initially answered the test questions. The results from that experiment are described in Rettig et al. (2019). We expand on their design in this study by repeating the same set of follow-up questions after 4 months and to a larger pool of respondents – both those who previously received the follow-up questions in wave 38 and those who had not seen the follow-up questions before. The in-between wave (wave 39, January 2019) constituted a typical GIP wave (20–25 minutes, 4€ conditional incentive) with questions on respondents' position on the labor market and their perceptions of the welfare state, gender roles, tax evasion, economic inequality, and the European Union. It contained no questions on environmental awareness (the topic of our test questions) and no experiments that substantially varied the topic, the number of questions or the overall length of the questionnaire.

5.4.2 Sample and variables

In total, 4,294 respondents participated in GIP wave 38 (November 2018), which included the initial test questions. Of these, 3,928 respondents (91.5%) also participated in wave 40 (March 2019), which included the follow-up questions used in this study. Because every respondent received a set of two test questions and respective follow-ups, we have two observations per respondent. However, we excluded some of these observations because they were incomplete due to item nonresponse on test questions or follow-up questions (17 observations), breakoff before or during the experiment (75 observations), or missingness on other variables of interest (155 observations). This yielded an analytic sample of 7,609 observations of 3,809 respondents.

Of these 3,809 respondents, 1,248 (32.8%) received the belief questions in wave 38, 1,271 (33.4%) received the attitude questions, and 1,290 (33.9%) received the behavior questions.

Across all question types, 1,877 respondents (49.3%) previously received the follow-up questions in wave 38, the rest received them for the first time in wave 40. Extreme responses account for 16.8% of all answers. In terms of socio-demographics, 48.0% of respondents were female, 14.9% had low formal education (see Appendix Table A5.2 for details), 29.3% medium-low, 22.0% medium-high, and 33.8% a high level of education, respectively. To address our hypothesis about age (*H4.1*), we split the sample into three groups of roughly equal size (<44 years, 44–58 years, and >58 years).

Regarding panel experience, 43.4% of all respondents were part of the newly recruited 2018 sample. The majority of respondents (74.6%) completed both waves 38 and 40 on a computer or tablet, 17.6% used a smartphone for both, and 7.8% switched between device types (e.g., from a computer to a smartphone; not accounting for switches between two computers etc.). Most respondents (97.2%) also participated in the wave in-between (wave 39, January 2019). Finally, while waves 38 and 40 were fielded 4 months apart, the actual time between the test questions and follow-up questions may be anywhere from 91 to 150 days, depending on when during the field times of both waves respondents chose to participate (although most respondents fall around the 4-month mark with a mode of 119 days and a median of 120 days).

To investigate the role of response burden and survey enjoyment, we computed indices of respondents' perceived response burden and survey enjoyment in wave 38 from their responses to a set of survey evaluation questions at the end of the wave. These questions are presented to respondents at the end of every GIP wave. From these six items we computed two indices: The response burden index was created from responses to questions about whether respondents found the survey to be "long", "difficult", and "too personal". The survey enjoyment index was created from respondents' ratings of the survey as "interesting", "varied", and "relevant". Each of these items was presented with a 4-point response scale with

endpoints labelled “not at all” and “very much”. We computed each index by summing respondents’ respective ratings and standardizing the indices to take values from 0 (the lowest possible burden or enjoyment, i.e., checking “not at all” on all items) to 1 (the highest possible burden or enjoyment, i.e., checking “very much” on all items).

5.4.3 Analytical strategy

To investigate how well respondents can remember their responses from a previous panel wave and test our hypotheses, we first give a descriptive overview of the claimed recall, correct recall, and recall certainty. We compare in which proportion of observations respondents correctly repeated their previous response (i.e., selected the exact same scale point) depending on whether recall was claimed (*H1*) both overall and separately by question type (*H2*). We then take a closer look at the differences between the original responses and recollections given 4 months later to gain further insight into how far off respondents were in cases where they repeated their previous response incorrectly using weighted kappa statistic. The (linear) weighted kappa measures the agreement between two ratings (in this case the original response in wave 38 and the repeated response in wave 40) while controlling for chance and penalizing larger differences between the two (see, e.g., Vanbelle & Albert, 2009). Disagreements are weighted using the formula $1-|i-j|/(k-1)$, in which i and j index the rows and columns of original and repeated responses (i.e., 1–11 for the original scale points numbered 0–10) and k is the maximum number of ratings (i.e., 11 for the eleven scale points). We computed 95% confidence intervals for the kappa using a bootstrapping procedure with 1,000 repetitions (Reichenheim, 2004).

To further investigate differences across question types as well as other correlates of remembering a previous response, we then compute three regression models: A logistic regression model of claimed recall, a linear regression model of recall certainty, and a logistic regression model of correct recall. In all models, we include the information whether an

extreme response was given (*H3*), socio-demographics: gender, age (*H4.1*) and education (*H4.2*), self-reported response burden (*H5.1*), survey enjoyment (*H5.2*), and panel experience (newly recruited versus experienced respondents; *H6*) as predictors. In addition, claimed recall is added as a predictor in the models on recall certainty and correct recall, and recall certainty is added as a predictor of correct recall. We also add an interaction effect of claimed recall with question type to see if differences in correct recalls between cases with and without claimed recall (i.e., the proportion of correctly repeated responses not explained by chance or stable opinion) differs across the question types, which would indicate a different proportion of respondents at risk for memory effects across question types. To investigate whether the effects of an extreme response and of panel experience differ across question types, we add interactions between those variables.

We also add several control variables to our models. Some research has suggested that response behavior may differ across respondents who use different devices to complete a survey, with some studies voicing concern that respondents who use smartphones to complete a survey may be more prone to some types of satisficing (Keusch & Yan, 2017; Krebs & Höhne, 2020; Lugtig & Toepoel, 2016; Struminskaya et al., 2015; Tourangeau et al., 2017). This difference in question processing may, in turn, be reflected in different rates of remembering a previous response. While studies investigating memory effects have so far generally not found differences across devices (Rettig et al., 2019; Revilla & Höhne, 2021), we therefore control for whether respondents answered the two survey waves on a computer, a smartphone, or switched between them. Furthermore, as some respondents in our experiment had already received the same follow-up questions within the same survey, we add the information whether respondents received the follow-ups before and whether they had been correct or incorrect in their previous recollection. As we expect respondents to forget their previous responses over time, we also add the number of days between their participation in both waves (91–150) and the information whether they also participated in the

panel wave in-between. Finally, to account for the clustered nature of our data with two observations per respondent, we add a dummy variable that indicates whether an observation is from the first or second set of follow-up questions a respondent received and compute cluster-robust standard errors in the regression models as well as cluster-adjusted *t*-tests for bivariate comparisons.

5.5 Results

Overall, respondents claimed that they remembered their previous response in 31.2% of all cases (Table 5.1). Of these, the correct response (i.e., the exact same scale point) was recalled in 34.1% of cases. Respondents also repeated their response correctly in 26.9% of cases where they had stated that they did not remember it, which is significantly lower (i.e., -7.2 percentage points; $t(4,597) = -6.023$, $p = .000$)[14]. Regarding our first hypothesis (*H1*), this indicates that for some respondents, remembering previous responses seems to persist even after a period of 4 months with another survey wave in-between, which is in line with our expectations. However, cases in which respondents claimed that they remembered their previous response and were subsequently able to correctly repeat it account for just 10.6% of all observations. In contrast, respondents were unable to correctly repeat their previous response in most cases (70.9%) and repeated their correct response despite claiming not to remember it (i.e., due to chance or an unchanged underlying information) in the remaining 18.5%.

[14] One-tailed, cluster-adjusted *t*-test to account for clustering in the data with two observations per respondent.

Table 5.1. Claimed recall, correct recall, weighted kappa, and mean recall certainty by question type.

	Overall		Beliefs		Attitudes		Behaviors	
Observations	7,609		2,493		2,536		2,580	
Claimed recall: yes	2,373	31.2%	835	33.5%	959	37.8%	579	22.4%
Of these:								
Correct recall	809	34.1%	282	33.8%	344	35.9%	183	31.6%
Weighted Kappa	0.444		0.445		0.436		0.411	
[CI]	[0.417, 0.470]		[0.403, 0.490]		[0.395, 0.481]		[0.364, 0.470]	
Mean certainty[1]		6.7		6.6		6.8		6.4
Claimed recall: no	5,236	68.8%	1,658	66.5%	1,577	62.2%	2,001	77.6%
Of these:								
Correct recall	1,408	26.9%	380	22.9%	465	29.5%	563	28.1%
Weighted Kappa	0.372		0.331		0.389		0.367	
[CI]	[0.356, 0.390]		[0.305, 0.364]		[0.362, 0.417]		[0.342, 0.397]	
Mean certainty		5.2		5.1		5.6		5.0
Overall								
Correct recall	2,217	29.1%	662	26.6%	809	31.9%	746	28.9%
Weighted Kappa	0.412		0.386		0.426		0.388	
[CI]	[0.399, 0.426]		[0.363, 0.412]		[0.401, 0.449]		[0.361, 0.410]	
Mean certainty		5.7		5.6		6.1		5.3

Note. Two observations per respondent. Respondents were randomly allocated to one of the question types. CI = 95% confidence interval. [1]On an 11-point scale from 0 "not at all certain" to 10 "absolutely certain".

To investigate how far off respondents' recollections were from their original responses, we computed the weighted kappa statistic as a measure of the agreement between their original response and the repetition (Table 5.1). The weighted kappa is consistently higher for cases in which recall was claimed than cases in which it was not. This indicates that respondents tended to give recollections closer to their original response when they claimed to remember it. We find this difference both overall and for each question type individually. Generally, the kappa statistic ranges around 0.4, indicating moderate agreement.

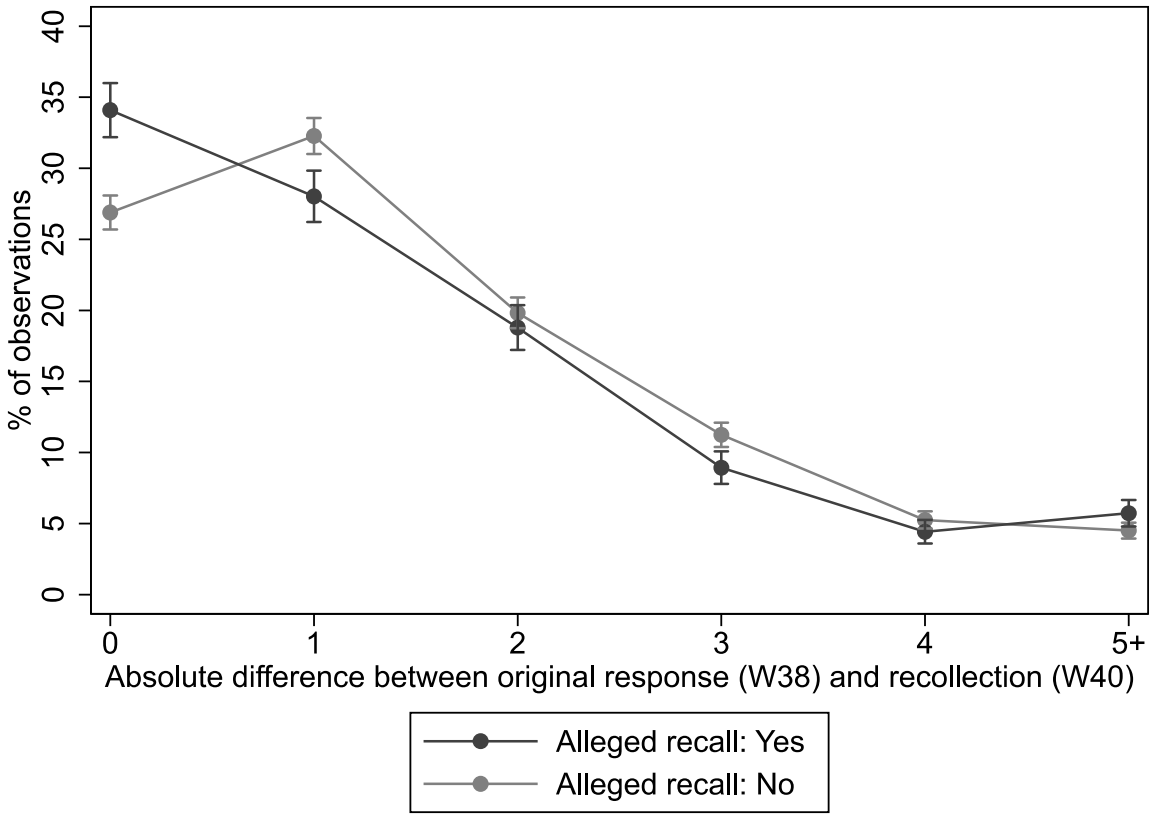

Figure 5.2. Differences between original response and recollection by claimed recall.

Note. Differences of 5 and more scale points collapsed into one category. The maximum possible difference depended on the original response and could not exceed 5 if respondents originally chose the midpoint category. Error bars represent the 95% confidence interval.

When further investigating the absolute difference between respondents' original response to the test question in wave 38 and their recollection in wave 40 (Figure 5.2), we can see that a correct recall (i.e., a difference of 0) is the most common outcome in cases where respondents claimed that they remembered their response. In cases where respondents reported not to remember their response but gave their best estimate, they were most commonly off by just a single scale point. Larger deviations from the original response are less likely. However, interestingly, respondents were also more likely to give a completely wrong recollection (i.e., off by 5 or more scale points) if they claimed to remember their response ($t(4,597) = -2.097$, $p = .018$).

Table 5.1 also provides an overview of claimed recall, correct recall, and mean recall certainty by question type. Notably, the proportion of cases in which respondents claimed they remembered their response differs considerably across question types and ranges from 37.8% for attitude questions to 22.4% for behavior questions. The overall proportion of correct recalls is also highest for attitude questions (31.9%), but lowest for belief questions (26.6%) with behavior questions in-between (28.9%). *T*-tests for differences between each pair of question types revealed these differences to be statistically significant (see Appendix Table A5.3 for *t*-statistics).

In cases where respondents claimed that they remembered their response, the proportions of correct recalls do not significantly differ across question types (see Appendix Table A5.3). However, if no recall was claimed, the proportion of correct repetitions (due to chance or stable underlying information) is significantly lower for belief questions than for attitudes ($t(1,900) = 4.048$, p = .000) and behaviors ($t(2,072) = 3.497$, p = .000). This in turn indicates that the difference in correct recalls between cases where recall was and was not claimed, and thus the proportion of correct recalls not explained by chance or stable opinion (i.e., due to memory), are different across question types. When taken in combination, these results indicate that remembering previous responses, does seem to differ across question types (*H2*).

The model of correct recall in Table 5.2 again confirms that a correct repetition of the previous response was significantly more likely when respondents claimed that they remembered their response (OR = 1.412, $p = .001$). This is in line with our expectations (*H1*) and indicates that after 4 months, not all correct repetitions of responses from a previous panel wave can be explained by unchanged underlying information or chance (i.e., evidence for the persistence of memory effects across panel waves).

Table 5.2. Logistic and linear regression models of claimed recall, recall certainty, and correct recall.

	Claimed recall		Recall certainty		Correct recall	
	OR	(SE)	b	(SE)	OR	(SE)
Question type (ref.: beliefs)						
Attitudes	1.065	(0.113)	0.450***	(0.130)	1.197	(0.128)
Behaviors	0.598***	(0.068)	-0.078	(0.130)	1.390**	(0.142)
Extreme response	2.085***	(0.266)	0.952***	(0.160)	1.363*	(0.168)
Extreme response * question type (ref.: non-extreme, beliefs)						
Attitudes	1.112	(0.186)	0.186	(0.204)	1.182	(0.192)
Behaviors	1.045	(0.211)	-0.495	(0.259)	0.666*	(0.130)
Female	0.910	(0.058)	-0.233***	(0.070)	1.130*	(0.063)
Age (ref.: <44 years)						
44–58 years	1.383***	(0.115)	0.313***	(0.089)	1.071	(0.076)
>58 years	1.796***	(0.157)	0.201*	(0.094)	0.946	(0.072)
Education (ref.: low)						
Medium-low	1.366**	(0.137)	0.123	(0.115)	0.972	(0.086)
Medium-high	1.115	(0.123)	0.052	(0.122)	1.122	(0.106)
High	1.258*	(0.126)	0.184	(0.114)	1.058	(0.091)
Response burden (W38)	0.653*	(0.110)	-0.052	(0.184)	0.808	(0.115)
Survey enjoyment (W38)	1.346	(0.215)	0.868***	(0.175)	0.809	(0.106)
Newly recruited respondent	1.347**	(0.145)	0.477***	(0.124)	0.855	(0.087)
Newly recruited * question type (ref.: newly recruited resp., beliefs)						
Attitudes	1.102	(0.164)	-0.153	(0.170)	1.125	(0.155)
Behaviors	0.948	(0.151)	0.029	(0.170)	1.002	(0.136)
Device (ref.: both waves computer)						
Both waves smartphone	1.117	(0.102)	-0.073	(0.098)	0.924	(0.073)
Device switch	1.413**	(0.165)	0.025	(0.123)	0.984	(0.104)
Follow-ups W38 (ref.: no follow-ups)						
Correct recall in W38	1.320***	(0.093)	-0.265***	(0.076)	1.250***	(0.077)
Incorrect recall in W38	1.241**	(0.098)	-0.384***	(0.086)	0.903	(0.068)
Days between waves	1.002	(0.004)	-0.004	(0.004)	1.004	(0.003)
In-between wave (W39)	0.878	(0.160)	0.038	(0.213)	1.034	(0.172)
First question	1.285***	(0.048)	-0.182***	(0.030)	0.975	(0.047)
Claimed recall (ref.: no)			1.365***	(0.118)	1.412***	(0.143)
Claimed recall * question type (ref.: claimed recall: no, beliefs)						
Attitudes			-0.360*	(0.160)	0.745*	(0.104)
Behaviors			-0.001	(0.174)	0.714*	(0.103)
Recall certainty					1.127***	(0.014)
Constant	0.159***	(0.083)	4.743***	(0.584)	0.099***	(0.045)
Pseudo-$R^2_{McKelvey \& Zavoina}$	0.096				0.053	
$R^2_{Adj.}$			0.130			
Observations	7,609		7,609		7,609	

Note. OR = Odds Ratios from logistic regression models, b = unstandardized linear regression coefficients, SE = cluster-robust standard errors. Two observations per respondent.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Across question types, we can see that compared to belief questions, recall to behavior questions was less likely claimed (OR = .598, p = .000), but recollections were more likely correct (OR = 1.390, p = .001). The reported certainty about remembering responses was higher for attitude questions (b = .450, p = .001). While both claimed recall and higher certainty correlate with a higher likelihood of correctly recalling a previous response (OR = 1.412, p = .001 and OR = 1.127, p = .000 respectively), the effect of claimed recall (i.e., the proportion of correct repetitions not explained by stable opinion or chance) is less pronounced for attitude and behavior questions than for belief questions (OR = .745, p = .034 and OR = .714, p = .020 respectively). In line with our expectations and the descriptive overview above, we thus conclude that respondents remember their responses to questions on beliefs, attitudes, and behaviors from a previous panel wave at different rates (*H2*).

In line with some of the previous findings on extreme responses, we find that when respondents provided an extreme response (i.e., selected an endpoint of the response scale as their original answer), they were more likely to claim that they remembered it (OR = 2.085, p = .000), reported higher certainty about remembering it (b = .962, p = .000), and were also more likely to give a correct recollection of what their response had been (OR = 1.363, p = .012). However, the positive effect of extreme responses on correct repetitions vanishes for behavior questions (OR = .666, p = .038; see also Appendix Table A5.4). While these results are in line with our expectation that overall, extreme responses are more likely to be remembered by respondents (*H3*), we only find this effect for questions on beliefs and attitudes but not behaviors.

When investigating socio-demographic correlates of remembering previous responses, we find few significant effects. Female respondents reported significantly lower certainty about remembering their response (b = -.233, p = .001) but were more likely to be correct in their recall than male respondents (OR = 1.130, p = .029). Respondents of the higher age groups

were more likely to claim that they remembered their response than respondents under 44 years (OR = 1.383, p = .000 and OR = 1.796, p = .000 for 44 to 58 years and over 58 years, respectively). The older age groups also reported higher certainty (b = .313, p = .000 and b = .201, p = .033 respectively), however, the likelihood of repeating the previous response correctly did not differ across age groups. We thus find no supporting evidence for our hypothesis that younger respondents would be more likely to remember their responses (*H4.1*). Similarly, respondents with medium-low or high formal education were more likely to claim that they remembered their response than respondents with low education (OR = 1.366, p = .002 and OR = 1.258, p = .022 respectively), but the reported certainty and likelihood to correctly repeat a previous response did not significantly differ across education levels. We thus find no evidence to support our hypothesis that respondents with higher education would be more likely to remember a previous response (*H4.2*).

Respondents who reported higher response burden in the previous panel wave were less likely to claim that they remembered their response (possibly in an effort to avoid further follow-up questions; OR = .653, p = .011), but we find no significant effect on the reported recall certainty or correct recall. Similarly, respondents who reported higher survey enjoyment reported higher certainty about remembering their response (b = .868, p = .000) but were not significantly more likely to claim recall or to correctly repeat their response. We thus do not find evidence to support the notion that remembering previous responses is linked to either response burden (*H5.1*) or survey enjoyment (*H5.2*).

Furthermore, newly recruited respondents were more likely to claim that they remembered their responses (OR = 1.347, p = .006) and reported higher certainty about remembering them (b = .477, p = .000) but did not differ from experienced respondents in their likelihood of correctly repeating their previous response. We also find no significant interaction effect between panel experience and question type regarding either claimed recall, recall certainty,

or correct recall. The evidence does therefore not support our expectation that newly recruited respondents would be more likely to remember their responses (*H6*).

In addition, respondents who switched devices between waves were more likely to claim that they remembered their responses than respondents who answered both waves on a computer (OR = 1.413, $p = .003$), but we find no significant relationship of device use with recall certainty or correct recall. Respondents who had previously received the follow-up questions at the end of wave 38 were more likely to claim that they could remember their response than those who received them for the first time in wave 40. This is true both in cases where respondents had previously correctly repeated their response (OR = 1.320, $p = .000$) and in cases where they had been incorrect (OR = 1.241, $p = .006$). However, respondents who had previously received the follow-up questions also tended to report lower certainty about remembering their response (b = -.265, $p = .001$ and b = -.384, $p = .000$ respectively). If respondents correctly recalled their response in the previous wave, they were more likely correct again in the later wave (OR = 1.250, $p = .000$). We find no significant difference in claimed recall, correct recall, and recall certainty across respondents who answered the test questions and follow-up questions with an in-between time interval from 91 days to 150 days, and across respondents who did and those who did not answer the in-between panel wave 39 in January 2019. Finally, respondents were more likely to claim that they remembered their response in the first set of follow-up questions they received (OR = 1.285, $p = .000$) but also reported lower certainty about remembering it (b = -.182, $p = .000$), while the likelihood of correct recalls did not significantly differ. Figure 5.3 displays a coefficient plot of the three regression models on claimed recall, recall certainty, and correct recall. As a sensitivity analysis, we additionally computed the model of correct recall without claimed recall and recall certainty as additional predictors and separately for cases in which recall was claimed and not claimed (see Appendix Table A5.5). These models did not yield any substantively different results.

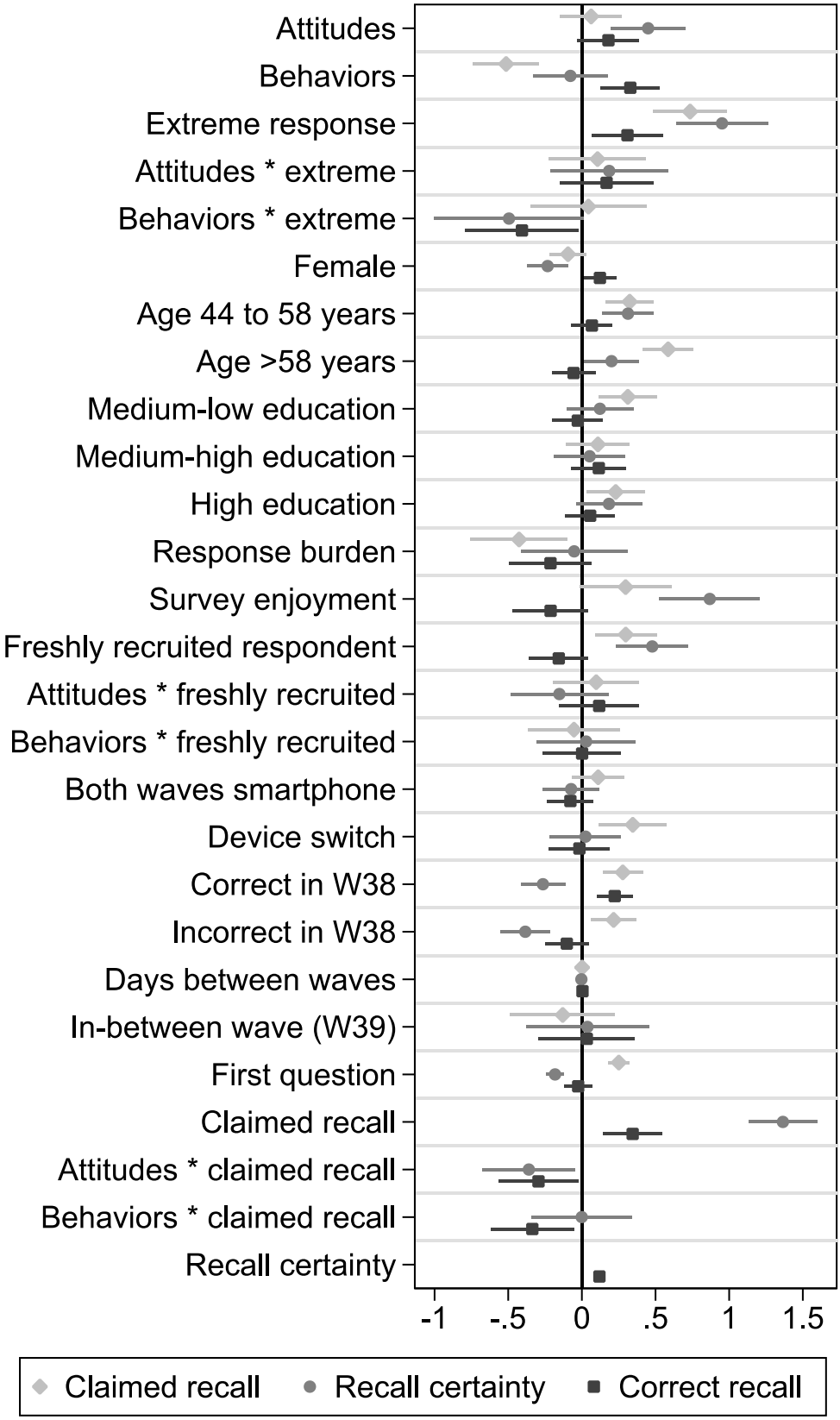

Figure 5.3. Coefficient plot for the regression models of claimed recall, recall certainty, and correct recall.

Note. Error bars represent the 95% confidence interval based on cluster-robust standard errors to account for clustering in our data with two observations per respondent.

5.6 Discussion and conclusion

In this study, we examined the ability of respondents from a probability-based online panel survey to remember their responses from a previous panel wave 4 months earlier. Table 5.3 provides a summary of our results regarding whether we found supporting evidence for each of our hypotheses.

Table 5.3. Summary of results regarding our hypotheses.

Hypothesis		Supported?
H1:	Respondents are more likely to correctly repeat their responses from a previous panel wave when they claim that they remember it.	Yes
H2:	The likelihood of respondents remembering their previous responses differs across questions on beliefs, attitudes, and behaviors.	Yes
H3:	Respondents are more likely to remember extreme responses than non-extreme responses.	(Yes)[1]
H4.1:	Younger respondents are more likely to remember their previous responses than older respondents.	No
H4.2:	Respondents with higher education are more likely to remember their previous responses than respondents with lower educational levels.	No
H5.1:	Respondents who perceived the previous panel wave as more burdensome are less likely to remember their responses.	No
H5.2:	Respondents who perceived the previous panel wave as more enjoyable are more likely to remember their responses.	No
H6:	Newly recruited respondents are more likely to remember their responses than experienced respondents.	No

Note. [1]for belief and attitude questions, not for behavior questions.

Overall, we find that respondents picked their correct previous response in about 29% of all cases. As may be expected, this finding is in line with a downward trend in correct recalls over time compared to the 61% Rettig et al. (2019) reported for the same test questions with follow-up questions after 20 minutes, the 60–88% some other studies have reported within one survey (Revilla & Höhne, 2021; Schwarz et al., 2020; van Meurs & Saris, 1990), and the roughly 40% van Meurs & Saris (1990) found after two weeks.

Furthermore, respondents who claimed that they remembered their response were significantly more likely to correctly recall it than respondents who said they could not remember it but gave their best estimate with a difference in correct repetitions of 7.2

percentage points. This difference in correct repetitions of the previous response has often been used as an estimate for the prevalence of memory effects (van Meurs & Saris, 1990). This practice follows the idea that correct repetitions by respondents who say they cannot remember their response can be explained by stable opinions (i.e., not recalling the response but repeating the same response because the underlying information has not changed) or correct guessing (van Meurs & Saris, 1990). Following this rationale, once respondents who claim they can remember their response are no better at correctly repeating it than those who say they cannot recall their response, all correct repetitions can be attributed to stable opinion or random chance and hence, no memory effects persist. In line with this idea, the 7.2 percentage point difference we found after 4 months is smaller than the roughly 20 percentage points reported by Rettig et al. (2019) for the same test questions after 20 minutes, 17 percentage points in Schwarz et al. (2020) and 34 percentage points in van Meurs & Saris (1990). This again points towards some memory of previous responses persisting even after 4 months, albeit to a much smaller degree than within the same survey. While Revilla and Höhne (2020) reported a similarly small difference within one survey, this may be due to the generally very high proportion of correct repetitions across both groups in their experiment.

Adding to this, we found that in almost 90% of all cases, respondents were either unable to correctly repeat their previous response or stated that they could not remember it (and presumably only repeated it correctly due to a stable opinion or correct guessing). Respondents who could not correctly recall their previous response were also most commonly off by just one scale point. In combination, these results imply that after four months, the group of respondents whose responses are affected by memory effects, their difference from non-affected respondents, and thus the resulting measurement error, are likely very small. In addition, while recall ability differed between genders, we did not find differences across age groups, education levels, panel experience or devices. After four months, memory effects are thus not likely to systematically vary across groups of respondents, which would be

problematic for comparisons across groups that were found to experience different levels of memory effects (Rettig & Blom, 2021). Repeated survey measurements with no or negligible memory effects may therefore be possible after a much shorter time than the two years Alwin (2011) suggested as a safe interval.

In our comparison of different question types, we found that after four months, responses to belief questions were correctly repeated the least often overall but had the largest difference between respondents who said they could and those who could not remember their response (i.e., the most correct responses due to memory). This is in contrast to the findings of Rettig et al. (2019), who found responses to belief questions had the smallest amount of correct repetitions that were not explained by stable underlying information or chance. As Rettig et al. (2019) argued, differences across question types may be driven by their differences in accessibility (salience) and stability. However, as van Meurs and Saris (1990) pointed out, most respondents will likely not change their opinions (or behaviors) over the course of just one survey. Therefore, accessibility and how strongly respondents felt about the topic may have played a larger role in their recall ability in the short term, while stability may be an additional factor primarily in the longer term.

It should be noted that our results are based on a set of test questions which respondents were asked for the first time. The use of completely new questions across all three question types served to avoid a confounding of different effects in our experiment (such as simultaneously making comparisons across experienced and inexperienced respondents a comparison across respondents who had answered the same question before and those who had not). However, measurements in longitudinal panel studies are often not just repeated once or twice, but many times and in regular intervals (e.g., every year). As a link between the repetition of information and its retention in memory has long been established in psychology (see, e.g., Hintzman, 1976), memory effects in repeated survey measurements may also be different

(and presumably more pronounced) for questions which respondents have answered many times before than for questions which were only repeated once. Our finding that previously receiving the follow-up questions in the same wave increased respondents' likelihood of reporting that they remembered their response (independently of whether they were correct) may be an indication that this repeated presentation and additional attention towards the test questions made respondents more likely to recognize them or to make cues more accessible at the process of retrieval of relevant information during the response process. An investigation into memory effects in the context of questions which have already been repeated to respondents frequently and regularly would therefore be an interesting avenue for future research.

Respondents who previously received the same follow-up questions also present a challenge in our analyses. While our finding that a response which was correctly recalled after 20 minutes was also more likely to be correctly recalled after four months seems intuitive, we cannot fully distinguish whether some respondents may have recalled their responses to the previous follow-up questions instead of or in addition to their actual previous response to the test question. This is especially true in cases where the correct response was recalled both times, as independently correctly recalling the initial response both times and correctly recalling the previous recollection (which was in turn correct, i.e., identical to the initial response) would lead to identical and virtually indistinguishable results.

Finally, we found no effect of the number of days between the administration of the test questions and follow-up questions (from 91 to 150 days), neither of the participation in the in-between panel wave on respondents' recall ability. However, neither the time interval between the participations nor participation in the in-between wave were randomly assigned. They were instead self-selected by respondents and not evenly distributed with only a small minority that did not participate in the in-between wave. In addition, we were only able to

analyze data from respondents who did participate in both the wave which included the test questions (wave 38, November 2018) and the wave which included the follow-up questions (wave 40, March 2019). Our data may therefore not be fully suited to investigate the effects of different between-participations time intervals and an additional panel wave on respondents remembering their previous responses. Our results may also not necessarily fully translate to respondents who participate infrequently or to newly recruited respondents who dropped out of the panel soon after their recruitment. While it would seem reasonable to expect an additional panel wave to disrupt respondents' memory of previous responses, our findings are in line with recent research in the context of a single survey, where a longer time to answer the questionnaire and memory interference tasks designed to reduce recall ability were also not found to reduce recall ability (Revilla & Höhne, 2021; Schwarz et al., 2020). A systematic variation of the time between panel waves and the inclusion of additional panel waves in-between would, however, be required to fully investigate the effects of these factors on memory effects in a longitudinal panel context.

Overall, however, our results bring good news for research practice of longitudinal studies with measurement repetitions: We find that after four months, the group of respondents who can remember their responses from a previous panel wave is small, their responses are not far off from those of respondents who cannot remember their previous response, and we find no evidence that these groups differ in a number of characteristics, including their age, education level, panel experience or the device used to complete the survey. We therefore conclude that after four months or more, memory effects will likely affect only a small number of respondents in a small way and that these are unlikely to be systematically different from respondents who are not experiencing memory effects. The resulting measurement error may therefore be negligible for panel studies in practice.

References

Alwin, D. F. (2007). *Margins of Error. A Study of Reliability in Survey Measurement*. John Wiley & Sons.

Alwin, D. F. (2010). How Good is Survey Measurement? Assessing the Reliability and Validity of Survey Measures. In P. V. Marsden & J. Wright (Eds.), *Handbook of Survey Research* (2nd ed., pp. 405–434). Emerald Group Publishing.

Alwin, D. F. (2011). Evaluating the Reliability and Validity of Survey Interview Data Using the MTMM Approach. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), *Question Evaluation Methods* (pp. 265–295). John Wiley and Sons.

Blom, A. G., Fikel, M., Friedel, S., Höhne, J. K., Krieger, U., Rettig, T., Wenz, A., & SFB 884 ´Political Economy of Reforms´ Universität Mannheim. (2019). *German Internet Panel, Wave 38 (November 2018)*. GESIS Data Archive, Cologne. ZA6958 Data file Version 1.0.0. https://doi.org/10.4232/1.13391

Blom, A. G., Fikel, M., Friedel, S., Höhne, J. K., Krieger, U., Rettig, T., Wenz, A., & SFB 884 ´Political Economy of Reforms´ Universität Mannheim. (2020a). *German Internet Panel, Wave 39 (January 2019)*. GESIS Data Archive, Cologne. ZA7588 Data file Version 2.0.0. https://doi.org/10.4232/1.13585

Blom, A. G., Fikel, M., Friedel, S., Höhne, J. K., Krieger, U., Rettig, T., Wenz, A., & SFB 884 ´Political Economy of Reforms´ Universität Mannheim. (2020b). *German Internet Panel, Wave 40 (March 2019)*. GESIS Data Archive, Cologne. ZA7589 Data file Version 1.0.0. https://doi.org/10.4232/1.13463

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting Up an Online Panel Representative of the General Population: The German Internet Panel. *Field Methods*, *27*(4), 391–408. https://doi.org/10.1177/1525822X15574494

Blom, A. G., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U., & Bossert, D. (2017). Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence From the German Internet Panel. *Social Science Computer Review*, *35*(4), 498–520. https://doi.org/10.1177/0894439316651584

Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. *Science*, *236*(4798), 157–161. https://doi.org/10.1126/science.3563494

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multi-Trait-Multimethod Matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin Company.

Cannell, C. F., & Fowler, F. J., Jr. (1965). Comparison of Hospitalization Reporting in Three Survey procedures. *National Center for Health Statistics. Vital Health Stat*, *2*(8).

Couper, M. P. (2000). Web Surveys. *Public Opinion Quarterly*, *64*(4), 464–494. https://doi.org/10.1086/318641

Dillman, D. A. (1978). *Mail and Telephone Surveys: The Total Design Method*. Wiley and Sons.

Dimitrov, D. M., & Rumrill, P. D. (2003). Pretest-Posttest Designs and Measurement of change. *Work*, *20*(2), 159–165.

Fishbein, M., & Ajzen, I. (1975). *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research.* Addison-Wesley.

Hintzman, D. L. (1976). Repetition and Memory. *Psychology of Learning and Motivation - Advances in Research and Theory*, *10*, 47–91. https://doi.org/10.1016/S0079-7421(08)60464-8

Höhne, J. K. (2021). New Insights on Respondents' Recall Ability and Memory Effects When Repeatedly Measuring Political Efficacy. *Quality and Quantity*. https://doi.org/10.1007/s11135-021-01219-2

Jäckle, A. (2009). Dependent Interviewing: A Framework and Application to Current Research. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (pp. 93–111). https://doi.org/10.1002/9780470743874.ch6

Jäckle, A., & Eckman, S. (2020). Is That Still the Same? Has that Changed? On the Accuracy of Measuring Change with Dependent Interviewing. *Journal of Survey Statistics and Methodology*, *8*(4), 706–725. https://doi.org/10.1093/jssam/smz021

Jaspers, E., Lubbers, M., & De Graaf, N. D. (2009). Measuring Once Twice: An Evaluation of Recalling Attitudes in Survey Research. *European Sociological Review*, *25*(3), 287–301. https://doi.org/10.1093/esr/jcn048

Keusch, F., & Yan, T. (2017). Web Versus Mobile Web: An Experimental Study of Device Effects and Self-Selection Effects. *Social Science Computer Review*, *35*(6), 751–769. https://doi.org/10.1177/0894439316675566

Krebs, D., & Höhne, J. K. (2020). Exploring Scale Direction Effects and Response Behavior Across Pc and Smartphone Surveys. *Journal of Survey Statistics and Methodology*, 1–19. https://doi.org/10.1093/jssam/smz058

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Lugtig, P., & Toepoel, V. (2016). The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey: Effects on Survey Measurement Error. *Social Science Computer Review*, *34*(1), 78–94. https://doi.org/10.1177/0894439315574248

Lynn, P. (2009). Methods for Longitudinal Surveys. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys* (pp. 1–19). John Wiley & Sons. https://doi.org/10.1002/9780470743874.ch1

McKelvie, S. J. (1992). Does memory contaminate test-retest reliability? *Journal of General Psychology*, *119*(1), 59–72. https://doi.org/10.1080/00221309.1992.9921158

Reichenheim, M. E. (2004). Confidence Intervals for the Kappa Statistic. *The Stata Journal*, *4*(4), 421–428. https://doi.org/10.1177/1536867x0400400404

Rettig, T., & Blom, A. G. (2021). Memory Effects as a Source of Bias in Repeated Survey Measurement. In A. Cernat & J. W. Sakshaug (Eds.), *Measurement Error in Longitudinal Data* (pp. 3–18). Oxford University Press.

Rettig, T., Höhne, J. K., & Blom, A. G. (2019). *Investigating Respondents' Ability to Recall Previous Responses to Different Types of Questions in a Probability-Based Online Panel*. Presentation at the European Survey Research Association (ESRA) 2019 Conference, Zagreb, Croatia.

Revilla, M., & Höhne, J. K. (2021). Repeatedly Measuring Political Interest: Can we Reduce Respondent' Recall Ability and Memory Effects in Surveys Using Memory Interference Tasks? *International Journal of Public Opinion Research*, *33*(3), 678–689. https://doi.org/10.1093/ijpor/edaa035

Salinsky, M. C., Storzbach, D., Dodrill, C. B., & Binder, L. M. (2001). Test-retest bias, reliability, and regression equations for neuropsychological measures repeated over a 12-16-week period. *Journal of the International Neuropsychological Society*, *7*(5), 597–605. https://doi.org/10.1017/S1355617701755075

Saris, W. E., & Gallhofer, I. N. (2014). *Design, Evaluation, and Analysis of Questionnaires for Survey Research* (2nd ed.). John Wiley and Sons.

Saris, W. E., Satorra, A., & Coenders, G. (2004). A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design. *Sociological Methodology*, *34*(1), 311–347. https://doi.org/10.1111/j.0081-1750.2004.00155.x

Schonlau, M., & Toepoel, V. (2015). Straightlining in Web Survey Panels over Time. *Survey Research Methods*, *9*(2), 125–137. https://doi.org/10.18148/srm/2015.v9i2.6128

Schuman, H., & Presser, S. (1981). *Questions and Answers in Attitude Surveys. Experiments on Question Form, Wording, and Context*. Academic Press.

Schwarz, H., Revilla, M., & Weber, W. (2020). Memory Effects in Repeated Survey Questions. Reviving the Empirical Investigation of the Independent Measurements Assumption. *Survey Research Methods*, *14*(3), 325–344. https://doi.org/10.18148/srm/2020.v14i3.7579

Strack, F., & Martin, L. L. (1987). Thinking, Judging, and Communicating: A Process Account of Context Effects in Attitude Surveys. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social Information Processing and Survey Methodology. Recent Research in Psychology.* (pp. 123–148). Springer.

Struminskaya, B., & Bosnjak, M. (2021). Panel Conditioning: Types, Causes, and Empirical Evidence of What We Know So Far. In P. Lynn (Ed.), *Advances in Longitudinal Survey Methodology* (pp. 272–301). John Wiley & Sons. https://doi.org/10.1002/9781119376965.ch12

Struminskaya, B., Weyandt, K., & Bosnjak, M. (2015). The Effects of Questionnaire Completion Using Mobile Devices on Data Quality. Evidence from a Probability-based General Population Panel. *Methods, Data, Analyses*, *9*(2), 261–292. https://doi.org/10.4232/1.12245.

Toepoel, V., Das, M., & Van Soest, A. (2008). Effects of Design in Web Surveys: Comparing Trained and Fresh Respondents. *Public Opinion Quarterly*, *72*(5), 985–1007. https://doi.org/10.1093/poq/nfn060

Tourangeau, R., Maitland, A., Rivero, G., Sun, H., Williams, D., & Yan, T. (2017). Web Surveys by Smartphone and Tablets. Effects on Survey Responses. *Public Opinion Quarterly*, *81*(4), 896–929. https://doi.org/10.1093/poq/nfx035

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. *Psychological Bulletin*, *103*(3), 299–314. https://doi.org/10.1037/0033-2909.103.3.299

Tourangeau, R., Rasinski, K. A., Bradburn, N., & D'Andrade, R. (1989). Belief Accessibility and Context Effects in Attitude Measurement. *Journal of Experimental Social Psychology*, *25*(5), 401–421. https://doi.org/10.1016/0022-1031(89)90030-9

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

van Meurs, A., & Saris, W. E. (1990). Memory Effects in MTMM Studies. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait Multimethod Studies* (pp. 134–146). North Holland.

Vanbelle, S., & Albert, A. (2009). A Note on the Linearly Weighted Kappa Coefficient for Ordinal Scales. *Statistical Methodology*, *6*(2), 157–163. https://doi.org/10.1016/j.stamet.2008.06.001

Yan, T., Fricker, S., & Tsai, S. (2020). Response Burden: What Is It and What Predicts It? In P. C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G. B. Willis, & A. Wilmot (Eds.), *Advances in Questionnaire Design, Development, Evaluation and Testing* (pp. 193–212). John Wiley & Sons. https://doi.org/10.1002/9781119263685.ch8

Appendix

Table A5.1. English translations of the test questions and follow-up questions.

Test questions	Question text	Response scale
Belief 1	How likely do you think it is that you can help save the environment by buying environmentally friendly products?	0 not at all likely – 10 extremely likely
Belief 2	How likely do you think it is that you can help prevent climate change by reducing your power consumption?	0 not at all likely – 10 extremely likely
Attitude 1	How acceptable would you find it to pay higher prices for environmentally friendly products?	0 not at all acceptable – 10 completely acceptable
Attitude 2	How acceptable would you find it to reduce your power consumption to help prevent climate change?	0 not at all acceptable – 10 completely acceptable
Behavior 1	How often do you pay attention to the environmental friendliness of the products you buy?	0 never – 10 always
Behavior 2	How often do you pay attention to your power consumption in everyday life to prevent climate change?	0 never – 10 always
Follow-up questions	**Question text**	**Response scale**
Claimed recall (if first follow-up)	In November, we asked you the following question: *[test question text]* Can you recall your exact answer to it?	yes / no
Claimed recall (if second follow-up)	We also asked you the following question: *[test question text]* Can you recall your exact answer to it?	yes / no
Correct recall (if claimed recall: yes)	Please indicate what your answer was.	[*same scale as test question*]
Correct recall (if claimed recall: no)	Even if you do not exactly recall: Please estimate, what your answer was.	[*same scale as test question*]
Recall certainty	How certain are you about your answer?	0 not at all certain – 10 absolutely certain

Note. Questions fielded in German, own translation.

Table A5.2. Coding scheme for education and age.

	Recoded categories	Original categories
Education		(highest school & professional degree)
	Low	No degree yet (still student)
		Left school with no degree
		Volks- / Hauptschule (or equivalent)
	Medium-low	Mittlere Reife / Realschule (or equivalent)
	Medium-high	Fachhochschulreife
		Abitur (or equivalent)
	High	Bachelor's degree
		Diploma / Master's (vocational university)
		Diploma / Master's (university)
		Ph.D.
Age		(year of birth categories)
	59 years and over	1935–1939
		1940–1944
		1945–1949
		1950–1954
		1955–1959
	44 to 58 years	1960–1964
		1965–1969
		1970–1974
	under 44 years	1975–1979
		1980–1984
		1985–1989
		1990–1994
		1995–1999
		2000 and later

Table A5.3. *T*-tests for differences in claimed recall and correct recall across question types.

	% of observations			t-test (one-tailed, cluster-adjusted)		
	Beliefs	Attitudes	Behaviors	t	df	p
Claimed recall: yes	33.5%	37.8%		2.607	2,517	.005
	33.5%		22.4%	-7.186	2,536	.000
		37.8%	22.4%	9.944	2,559	.000
Correct recall						
…overall	26.6%	31.9%		3.848	2,517	.000
	26.6%		28.9%	1.781	2,536	.038
		31.9%	28.9%	2.193	2,559	.014
…if claimed recall: yes	33.8%	35.9%		.845	1,180	.199
	33.8%		31.6%	- .811	950	.209
		35.9%	31.6%	1.605	1,032	.054
…if claimed recall: no	22.9%	29.5%		4.048	1,900	.000
	22.9%		28.1%	3.497	2,072	.000
		29.5%	28.1%	.856	2,052	.196

Table A5.4. *T*-tests for differences in correct recall between extreme and non-extreme responses by question types.

	% of observations		t-test (one-tailed, cluster-adjusted)		
Correct recall	Extreme	Non-extreme	t	df	p
Beliefs	36.3%	24.6%	- 4.643	1,422	.000
Attitudes	44.0%	28.0%	- 6.918	1,494	.000
Behaviors	29.7%	28.8%	- .276	1,451	.391
Overall	38.8%	27.2%	- 7.876	4,371	.000

Table A5.5. Logistic regression models of correct recall without claimed recall and recall certainty as additional predictors and separately by claimed recall.

	Correct recall		Correct recall (claimed recall: yes)		Correct recall (claimed recall: no)	
	OR	(SE)	OR	(SE)	OR	(SE)
Question type (ref.: beliefs)						
Attitudes	1.138	(0.109)	0.822	(0.146)	1.272 *	(0.146)
Behaviors	1.218 *	(0.114)	1.000	(0.184)	1.299 *	(0.141)
Extreme response	1.662 ***	(0.214)	3.464 ***	(0.667)	0.638 *	(0.133)
Extreme response * question type (ref.: non-extreme, beliefs)						
Attitudes	1.137	(0.191)	0.855	(0.215)	1.726 *	(0.448)
Behaviors	0.599 *	(0.122)	0.492 *	(0.160)	1.057	(0.311)
Female	1.096	(0.061)	1.072	(0.106)	1.107	(0.075)
Age (ref.: <44 years)						
44–58 years	1.130	(0.080)	1.042	(0.135)	1.125	(0.095)
>58 years	1.003	(0.076)	0.974	(0.132)	0.960	(0.088)
Education (ref.: low)						
Medium-low	1.007	(0.089)	0.832	(0.129)	1.056	(0.111)
Medium-high	1.141	(0.108)	1.045	(0.174)	1.158	(0.131)
High	1.101	(0.095)	0.878	(0.138)	1.152	(0.119)
Response burden (W38)	0.788	(0.111)	0.837	(0.221)	0.782	(0.135)
Survey enjoyment (W38)	0.915	(0.119)	0.837	(0.201)	0.911	(0.141)
Newly recruited respondent	0.932	(0.093)	0.728	(0.123)	0.975	(0.122)
Newly recruited * question type (ref.: newly recruited resp., beliefs)						
Attitudes	1.089	(0.149)	1.381	(0.315)	1.052	(0.182)
Behaviors	0.983	(0.132)	1.201	(0.297)	0.936	(0.152)
Device (ref.: both waves computer)						
Both waves smartphone	0.923	(0.073)	0.962	(0.140)	0.895	(0.084)
Device switch	1.006	(0.106)	1.068	(0.196)	0.959	(0.130)
Follow-ups W38 (ref.: no follow-ups)						
Correct recall in W38	1.225 **	(0.076)	1.450 ***	(0.157)	1.076	(0.082)
Incorrect recall in W38	0.875	(0.064)	0.734 *	(0.101)	0.931	(0.080)
Days between waves	1.004	(0.003)	1.010	(0.006)	1.001	(0.004)
In-between wave (W39)	1.016	(0.167)	1.102	(0.324)	1.022	(0.206)
First question	0.969	(0.046)	0.899	(0.076)	0.975	(0.059)
Constant	0.200 ***	(0.091)	0.142 *	(0.119)	0.281 *	(0.152)
Pseudo-R^2 McKelvey & Zavoina	0.025		0.098		0.015	
Observations	7,609		2,373		5,236	

Note. OR = Odds Ratios from logistic regression models, SE = cluster-robust standard errors. Two observations per respondent. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6. Conclusion

Surveys remain a popular widespread tool for data collection across a variety of disciplines and whenever researchers use survey data, they need to be certain that respondents' answers and the conclusions drawn based on these are accurate. The present dissertation contributes to the current understanding of potential errors during respondents' cognitive response process in several ways. In this chapter, I discuss the key insights and implications from each of the four papers in turn, followed by a summary of the overall research contribution and an outline of the avenues for further research identified in this dissertation.

Paper I: Investigating Respondent Attention to Experimental Text Lengths

In this paper, I investigated the relationship of respondents' attention to treatment texts with the length of these texts, respondents' socio-demographic and participation characteristics, and the performance of response time as an attention indicator. While a small group of respondents skipped reading even the shortest text, for a much larger group whether they read a text depends on its length. Attention levels differ across socio-demographic groups, levels of panel experience, and may also be linked to the time of day and point during the fieldtime respondents choose to participate. As removing inattentive respondents is thus likely to result in a biased sample, optimizing surveys to facilitate attentive participation should be preferred over removing inattentive respondents. Most respondents are not likely to read an excessively long treatment text and their use should thus be avoided. Additional caution should be taken when researchers make comparisons across treatments of differing lengths or across groups of respondents that are associated with differing attention levels. A fast response time is not a suitable standalone indicator of inattention and may systematically misclassify respondents, potentially creating or masking biases in the sample. It should thus not be used as the sole method of identifying inattentive respondents.

However, the paper also outlines the need for additional research on respondent attention to inform researchers on how to deal with inattentive respondents. In particular, further research should explore whether the commonly voiced concerns about angering respondents with attention checks are justified, and thus whether researchers should or should not routinely include them in their surveys. Researchers should also explore possibilities of treating inattentive respondents with targeted interventions to encourage attentive participation instead of dropping these respondents from their studies. Finally, more research is needed to guide survey practitioners' decisions on how to treat inattentive respondents in a longitudinal panel context. In particular, additional insights into whether inattention during one interview predicts repeated inattentive responding in future survey waves and whether inattentive respondents can successfully be treated to participate attentively are required. Such insights could inform practitioners on whether such interventions are necessary, useful, cost-effective, and preferable to removing long-term panelists, which is particularly costly.

Paper II: Memory Effects as a Source of Bias in Repeated Survey Measurement

The second paper conceptualized memory effects and how these interfere with respondents' ideal cognitive response process, integrated memory effects into the wider survey-methodological literature on adjacent concepts such as question order effects, and provided a literature review of the prior research on memory effects. In the conceptual framework of memory effects in the cognitive response process, I propose two avenues in which respondents' later responses may be influenced by their memory of a previous response: In the memory consistency model, the previous response is retrieved as part of the information respondents retrieve to answer the repeated question and used as a basis for reflection. The later response is then adjusted to the previous response. In the memory satisficing model, the previous response is instead judged to still be acceptable and reiterated instead of forming a new judgement.

The literature review revealed a small body of previous research on memory effects that provides some evidence that respondents are mostly able to remember their previous responses within one survey. It further gives some indication that responses after as long as two weeks may not be completely free from memory effects. However, the paper also outlines the need for further research on respondents remembering previous responses to identify potential correlates that may increase or decrease the likelihood of respondents remembering their answers, as well as guidance for researchers in determining adequate time intervals for repeated survey measurements that are free from memory effects. Furthermore, additional research is needed to investigate how much respondents' later responses differ from those they would have given without the influence of memory effects in practice. More research is also needed to investigate whether respondents more commonly use their memory to further reflect upon their view and how it has changed since the previous interview, or as a means to reduce response burden by repeating their previous response without reevaluating it.

Paper III: Memory Effects: A Comparison Across Question Types

In the third paper, I investigated to what extent respondents remembered their previous responses within one survey depending on the type of question, whether the original response was extreme, respondents' level of panel experience, and across socio-demographic groups. While responses to questions on beliefs, attitudes, and behaviors are remembered at different rates, respondents overall remember their responses within one survey in most cases across all three question types. Memory effects are thus likely to occur in repeated survey measurement after a short time, such as in pretest-posttest or test-retest designs that incorporate repeated measurements within one survey. This may result in measurement error due to artificially consistent responses introducing bias into analyses and thus lead to inaccurate conclusions. In addition, different rates of remembering previous responses are linked to respondents' age,

gender, and education level. Measurement error due to memory effects may thus also systematically vary across groups of respondents.

These results indicate that measurement error due to respondents remembering their previous responses is a concern that should be considered in research designs that incorporate repeatedly asking respondents the same questions after a short in-between time. However, while the evidence suggests that respondents are likely to remember their responses, more research is needed to determine how much this will affect their responses to the later iterations of the same questions in practice. To accurately assess the potential magnitude of measurement error resulting from memory effects, researchers need to determine how likely the presence of a previous response in respondents' memory is to shift their later responses and by how much. Furthermore, additional research is needed to investigate how these results translate to a longer time frame and to different survey modes. It may, for example, be harder for respondents to remember their previous response in a telephone interview where the response scale is not visually presented to them, which may otherwise serve as a recall cue.

Paper IV: Memory Effects in Online Panel Surveys: Investigating Respondents' Ability to Recall Responses from a Previous Panel Wave

In the fourth paper, I expanded the investigation into memory effects presented in the third paper to a longer time frame between the original measurement and its repetition in a longitudinal panel context. Some respondents still remember their responses from a previous panel wave after four months and the likelihood of remembering a response differs across question types and between genders. However, most respondents cannot remember their responses after four months. These respondents are most commonly off by only a single scale point and thus their responses are not far off from those who can remember theirs. Remembering a response from a previous panel wave was also not found to be correlated with other socio-demographic and participation characteristics, including respondents' age,

education level, panel experience, self-reported response burden and survey enjoyment, the device used to complete the survey, or whether respondents participated in another panel wave in-between. Neither the prevalence of respondents remembering their responses nor its systematic variation across groups of respondents found in the short term in the third paper thus seem to persist in the longer term. Thus, after a period of four or more months, the group of respondents at risk of experiencing memory effects is likely to be small and not systematically different from unaffected respondents. The resulting measurement error in longitudinal study designs with measurement repetitions after at least four months may therefore be negligible, which is good news for panel research.

However, research on memory effects has so far been based on a single repetition of a previously asked question. Particularly for panel designs that repeatedly collect the same information from respondents more than twice, often in regular intervals, further research is therefore needed to investigate whether these positive results translate to multiple repetitions over longer periods. Respondents who cannot remember their answer to a question that they have only been asked once before may yet recognize a question that is repeated every year and form a response routine to such a question.

6.1 Overall research contribution and future prospects

The present dissertation contributes to the literature on inattention and memory effects as potential sources of interference with respondents' cognitive response process in several ways. In this dissertation, I synthesized the present literature on respondent attention and on memory effects and identified gaps in the existing research. I further expanded the conceptual understanding of these two concepts, added to the empirical evidence on respondent attention specifically to treatment texts, as well as on memory effects both in the short and longer term, and thereby contributed to closing some of the previously identified gaps in the literature.

The investigation of respondent attention to treatment texts not only showed that a substantial group of respondents will not read an excessively long text, but the link between attention and text length also indicates that many of them would have read a shorter text. Paying attention is thus not the sole responsibility of respondents. Instead, these results once again outline the need for researchers to optimize and improve their surveys to facilitate attentive participation. While this may seem like an obvious conclusion that survey researchers should already be aware of, Couper's observation that "[t]here has been an increasing disconnect between what we are asking for, and what people may think is reasonable to provide" (2013, p. 152) rings true here. To receive high quality survey data, it is important that researchers adjust the cognitive demands their surveys put on respondents to a reasonably low level.

Similarly, to avoid measurement error from respondents remembering their previous responses, researchers need to design their studies with repeated survey measurements around respondents' memory capacity. Despite previous suggestions that respondents may not remember their responses after as little as 20 minutes (see Saris et al., 2010), both the review of recent research on memory effects and the results of the third paper give a relatively consistent indication that respondents are likely remember their responses from the same survey. As recent attempts at disrupting respondents' memory of their prior responses within one survey have been unsuccessful (Revilla & Höhne, 2021; Schwarz et al., 2020), it seems that at this point, the only reliable way to prevent memory effects is to design repeated survey measurements in such a way that enough time passes between the repetitions, thus allowing respondents to forget about their answers.

In this regard, the fourth paper in this dissertation made an important step in investigating how long the time between repeated survey measurements should be to achieve repeat interviews without measurement error from memory effects. The literature previously gave some indication that two weeks may not be long enough (van Meurs & Saris, 1990) on the

one hand and that several years would likely suffice (Alwin, 2007; Jaspers et al., 2009) on the other. I contributed to filling this significant gap in the suggested time frames for reinterviews by demonstrating that respondents remembering their previous responses and the resulting measurement error due to memory effects may become negligible in practice after a few months, rather than years.

Finally, the present dissertation contributes to the field by pointing towards clear opportunities for further research. Regarding respondent attention, it would be interesting to measure attention repeatedly over several panel waves in a longitudinal context. This would allow a more in-depth investigation into how much inattention is circumstantial, dependent on the survey content, and to what degree it is a more persistent respondent characteristic or even a learned behavior to deal with the continued cognitive demand of participating in a panel survey (i.e., a form of panel conditioning; see, e.g., Warren & Halpern-Manners, 2012). In addition, it would be interesting to measure respondents' attitudes towards these attention measurements to answer whether concerns about angering respondents by implementing attention checks are justified, and whether respondents' attitudes towards attention checks change when they are presented repeatedly. Respondents may, for example, not be offended by attention checks initially but eventually grow tired of being monitored continually. Furthermore, it would be interesting to investigate whether inattentive respondents can be encouraged to participate more attentively with different targeted interventions. All of these points could in principle be investigated within a single longitudinal attention experiment.

Regarding memory effects, research has so far established how well respondents can remember their previous responses. However, in a next step it would be necessary to investigate to what extent holding the previous response in one's memory translates to an actual change in the later responses in practice. This includes investigating which proportion of respondents will actually be influenced, in which way, and by how much this alters their

responses. In addition, it would be interesting to compare memory effects across different survey modes, as the current research is nearly exclusively based on self-administered web surveys. However, a comprehensive investigation into memory effects outside of measurement repetitions in a cross-sectional study and across survey modes would require implementation in a longitudinal mixed-mode survey with random assignment of the respondents to the different survey modes, which is not a very common survey design. It would furthermore be interesting to investigate both whether respondents remember their responses and the influence this has on their later responses depending on how often and in which interval respondents have already been asked the same question in a longitudinal context. A question that is repeated regularly may be easier to recognize and recall the response to than one that has only been asked once or twice before, and respondents may lose motivation to repeatedly form a new judgement to the same questions after a while. Measurement error due to memory effects may thus become more pronounced over time. This could for example be investigated by implementing a long-term experiment in which panelists are randomly assigned to receive the same question in different regular intervals (e.g., one group receives the question once a year, one twice a year, one quarterly, and one only at the beginning and the end of the overall duration of the experiment).

Overall, this dissertation has addressed several important questions on the role of inattention and memory effects as sources of measurement error in survey research, and particularly in online panels. At the same time, these topics provide rich opportunities for further research. Given the continued importance of high-quality survey data for several disciplines, working towards filling the remaining gaps in the survey-methodological literature on measurement error is an important and worthwhile pursuit.

References

Alwin, D. F. (2007). *Margins of Error. A Study of Reliability in Survey Measurement*. John Wiley & Sons.

Couper, M. P. (2013). Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods*, *7*(3), 145–156.

Jaspers, E., Lubbers, M., & De Graaf, N. D. (2009). Measuring Once Twice: An Evaluation of Recalling Attitudes in Survey Research. *European Sociological Review*, *25*(3), 287–301. https://doi.org/10.1093/esr/jcn048

Revilla, M., & Höhne, J. K. (2021). Repeatedly Measuring Political Interest: Can we Reduce Respondent' Recall Ability and Memory Effects in Surveys Using Memory Interference Tasks? *International Journal of Public Opinion Research*, *33*(3), 678–689. https://doi.org/10.1093/ijpor/edaa035

Saris, W. E., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, *4*(1), 45–59. https://doi.org/10.18148/srm/2010.v4i1.2682

Schwarz, H., Revilla, M., & Weber, W. (2020). Memory Effects in Repeated Survey Questions. Reviving the Empirical Investigation of the Independent Measurements Assumption. *Survey Research Methods*, *14*(3), 325–344. https://doi.org/10.18148/srm/2020.v14i3.7579

van Meurs, A., & Saris, W. E. (1990). Memory Effects in MTMM Studies. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait Multimethod Studies* (pp. 134–146). North Holland.

Warren, J. R., & Halpern-Manners, A. (2012). Panel Conditioning in Longitudinal Social Science Surveys. *Sociological Methods and Research*, *41*(4), 491–534. https://doi.org/10.1177/0049124112460374