Survey Research Methods (2023)
Vol. 17, No. 3, pp. 301-322
doi:10.18148/srm/2023.v17i3.7991
European Survey Research Association

© 2023Author(s)
CC BY 4.0

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

Tobias Rettig[1] and Bella Struminskaya[2]
[1]University of Mannheim
[2]Utrecht University

If respondents recognize repeated survey questions and remember their previous responses, this can result in measurement error. Most studies that have investigated respondents' recall of their prior answers have done so in the context of repeated measurements within cross-sectional surveys. 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 cross-sectional surveys. 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 to 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 months or longer.

Keywords: extreme responses; measurement error; memory effects; online panel; repeated measurement

1 Introduction

Repeated measurements of the same survey questions from the same respondents have several applications in survey research. Longitudinal surveys commonly repeat questions to measure change over time on the individual level (see, e.g. Lynn, 2009). Repeated measurements are also used in pretest-posttest experimental designs (Campbell & Stanley, 1966; Dimitrov & Rumrill, 2003) to evaluate measurement quality of survey instruments (e.g., in test-retest or quasi-simplex designs to estimate reliability (Alwin, 2007, 2010, 2011), and in multitrait-multimethod (MTMM) designs to estimate reliability and validity (Campbell & Fiske, 1959; Saris & Gallhofer, 2014; Saris et al., 2004). Those applications rely on the assumption that different measurements are independent, that is, respondents undergo the cognitive response process—comprehending the question, retrieving the relevant information from memory, integrating this information and forming a judgement, and selecting the appropriate response (Tourangeau et al., 2000)—each time and that responses to the repetitions are not influenced by the previous response (Rettig & Blom, 2021). However, respondents may edit their response before reporting it, for example, due to social desirability or to appear consistent. Respondents can also retrieve an already formed judgement and forego integrating retrieved information into a new judgement (Strack & Martin, 1987; Tourangeau, 2018). Moreover, respondents can 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).

In those cases, the assumption of independent measurement would be violated. Such a violation is problematic for data quality and can introduce bias into the results of any analyses that rely on measurement independence. The statistical models used to analyze data that rely on the assumption of measurement independence can thus be invalid. Respondents who are influenced by their earlier answers can provide responses with an inflated level of consistency, leading to an underestimation of changes or treatment effects, or an overestimation of reliability or measurement validity (Alwin, 2011; Rettig & Blom, 2021).

Contact information: Tobias Rettig, University of Mannheim, Mannheim Center for European Social Research (MZES), B6 30-32, 68131 Mannheim, Germany (E-mail: tobias.rettig@uni-mannheim.de).

In this study, we investigate how independent repeated survey measurements actually are by assessing how well respondents of a longitudinal survey remember their responses from a previous panel wave after four months. We aim at disentangling memory effects from other reasons why respondents repeat their previous answer to the same question. For example, respondents' true answer may simply not have changed in the meantime (i.e., stable attitudes) and/or respondents might select the same answer on a scale due to chance (van Meurs & Saris, 1990). We asked the respondents whether they remembered their responses and how certain they felt about the response they had given in the previous wave. Subsequently, we check whether respondents' answers match the answers they gave in the wave prior.

It would also be problematic if memory effects were not the same across constructs, which would mean that the independent observations assumption is dependent on the distribution of a variable. We compare respondents' memory of their previous responses across questions on beliefs, attitudes, and behaviors. In addition, we investigate whether memory effects are more pronounced for the endpoints of a response scale (i.e., extreme responses, which may be a sign of strong opinions).

Memory effects are, however, not always disadvantageous for data quality. Sometimes researchers want respondents to remember their answers, for example, for factual questions. In dependent interviewing, researchers present previous responses to respondents and ask whether the situation has changed. The goal of dependent interviewing is to improve measurement validity by reducing spurious reports of change or stability, and to reduce item nonresponse and response burden (Jäckle, 2008, 2009; Jäckle & Eckman, 2020). Presenting previous responses can help jog respondents' memory and make clear what type of information researchers are looking for; it simplifies respondents' cognitive task and provides a baseline from which respondents may adjust their response (Eggs & Jäckle, 2015; Hoogendoorn, 2004; Jäckle, 2008, 2009; Mathiowetz & McGonagle, 2000).

However, respondents might accept their previous response as a preexisting judgement and agree with it (Hoogendoorn, 2004; Mathiowetz & McGonagle, 2000), which leads to underreporting of changes (Eggs & Jäckle, 2015; Lugtig & Lensvelt-Mulders, 2014). If respondents retrieve their previous response from memory, they may treat it as an existing judgement and simply repeat it or edit their later response to be more consistent with their previous response (Rettig & Blom, 2021). Also, in dependent interviewing *all* respondents are reminded of their previous responses, while it can vary how well respondents remember their previous response. When not presented with their previous answer, respondents might use accurate, inaccurate, or no information about their previous response in processing the repeated question, which results in differences in accuracy that dependent interviewing seeks to alleviate.

In the present study, we focus on situations in which remembering is explicitly not wanted and may have undesirable effects on data quality. Most studies on memory effects investigate whether respondents remember their previous responses within one survey, which is relevant for research designs with measurement repetitions after a short time (e.g., pretest-posttest experiments or MTMM models). In this study, we expand upon the existing literature by investigating respondents' ability to remember their responses from a previous panel wave in a longitudinal setting.

2 Background

Several studies have investigated memory effects in repeated survey measurements (Höhne, 2021; Rettig & Blom, 2021; Rettig et al., 2023; Revilla & Höhne, 2021; Revilla et al., 2023; Schwarz et al., 2020; van Meurs & Saris, 1990). Most of these studies administered repeated measurements within the same survey. van Meurs and Saris (1990) found that about 70% of respondents correctly repeated their answers about 9 minutes after the initial questions within one survey, and about 40% repeated their answer correctly after two weeks. Rettig et al. (2023) found that respondents correctly repeated their previous responses in 61% of cases after about 20 minutes, while Revilla and Höhne (2021) and Schwarz et al. (2020) found that 60% and 88% of respondents correctly repeated a previous response within one interview, respectively.

Generally, people tend to forget information as time passes (Bradburn et al., 1987; Cannell & Fowler, 1965; Tourangeau et al., 2000), so one can expect that respondents will be less likely to remember previous responses after weeks or months than after a few minutes into a survey. Indeed, fewer respondents correctly repeated their response after two weeks than after 9 minutes (van Meurs & Saris, 1990). However, because many respondents still remembered their answers, two weeks may not be long enough to prevent memory effects. In contrast, McKelvie (1992) found that 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 can persist after 12–16 weeks (Salinsky et al., 2001).

For longer timeframes, the influence of previous responses diminishes. 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

	Self-reported memory of previous response:	
	Yes	No
Correct repetition of previous response	Memory effects	
	Stable underlying information, correct guessing	
Incorrect repetition of previous response	misremembering	no memory

Figure 1

Different explanations for correct and incorrect repetitions of previous responses

10 years prior. Alwin (2011) suggests that an interval of two years between measurement repetitions would be sufficient to rule out memory effects.

However, most longitudinal surveys readminister questions more frequently than every few years. For example, the Survey of Income and Program Participation (SIPP) in the USA or the UK's Understanding Society annually readminister many questions. Panels such as the Dutch LISS panel, German GESIS Panel, or the German Internet Panel (GIP) contain annual modules which repeat the same questions. Some questions in those panels are repeated more frequently. The Current Population Survey (CPS) collects the same information from households monthly for four consecutive months and the GIP's 2020 Mannheim Corona Study (MCS) administered questions to respondents weekly for 16 weeks. A common response format for frequently readministered questions are Likert scales. In this study, we investigate how well respondents remember their responses to different types of questions (beliefs, attitudes, and behaviors) that were asked with 11-point scales in a previous panel wave after four months.

3 Hypotheses

Some respondents may correctly repeat their previous response without remembering it, either because they have not changed their mind or by chance (van Meurs & Saris, 1990). van Meurs and Saris (1990) suggested to use the proportion of respondents who correctly repeat their response—despite (self-reportedly) not remembering it—as an approximation for correct repetitions due to chance or stable opinion (i.e., not due to memory). Respondents who claim to remember their response may be misremembering it and thus not be able to correctly repeat it. The difference in correctly repeated responses between these two groups can be used as an approximation for correct repetitions due to memory effects (see Figure 1).

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). However, if respondents who claim to remember their response are indeed better at correctly repeating it, this would indicate memory effects. We hypothesize:

H1 After four months, respondents who self-report remembering their response are more likely to correctly recall it.

Different types of information are forgotten at different rates (Bradburn et al., 1987; Tourangeau et al., 2000), so respondents may 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, p. 82). Based on their beliefs, respondents form attitudes, which describe having positive or negative feelings towards an object or issue (e.g., "In general, how do you feel about nationwide legalization of abortion in the United States?"; Dillman, 1978, p. 81). Behaviors are formed on the basis of attitudes and describe respondents' actions (e.g., "Are you currently taking birth control pills?"; Dillman, 1978, p. 84; see also Fishbein and Ajzen, 1975).

These different types of questions ask respondents for different types of information, while they undergo the stages of the response process (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 have to recall their own actions, which is a more overt, directly accessible type of factual information.

Due to these differences in the retrieval as well as differences in the stability over time and accessibility of different types of information, how well respondents remember their response later might vary (Rettig et al., 2023). In the short-term when the underlying information is unlikely to change, more accessible information should be more easily reproduceable by respondents. Indeed, Rettig et al. (2023) 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, which may influence the ability to correctly repeat previous responses. We thus hypothesize:

H2 The likelihood of respondents remembering their previ-

ous responses differs across questions on beliefs, attitudes, and behaviors.

Similarly, a response based on a stronger opinion might be easier to remember. Some authors suggest that more salient (i.e., accessible) information is more likely to be retrieved during the cognitive response process and more likely to be used in forming the response (Schuman & Presser, 1981; Tourangeau & Rasinski, 1988). A stronger and more salient opinion may also be easier to repeat later. Several studies have found that respondents are more likely to remember their response if they originally chose an extreme response option (Rettig et al., 2023; van Meurs & Saris, 1990)[1]. We thus expect:

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

As memory effects are 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 could make the resulting measurement error more problematic, because it could systematically vary across groups of respondents. For example, Rettig et al. (2023) find that younger and higher educated respondents were more likely to remember their responses. Höhne (2021) finds the same effect of age, but not 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 or education, although this might be explained by the homogeneity of their sample of university students. Despite mixed findings, these studies indicate that better memory of previous responses may be linked to age and education. We, therefore, expect:

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.

Respondents may be less likely to remember a response if they generated it by superficially undergoing the cognitive response process (i.e., satisficed; see Krosnick, 1991). As several authors suggest, satisficing behavior is a way for respondents to deal with the cognitive demand of answering a survey and therefore increases with rising response burden and fatigue (Galesic & Bosnjak, 2009; Krosnick, 1991). By extension, we hypothesize:

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.

However, while satisficing may be caused by higher response burden, it is also a strategy respondents use to alleviate burden (Krosnick, 1991; Yan et al., 2020). Higher response burden thus might lead respondents to satisfice, which in turn reduces their response burden, which then may reduce their need to satisfice. We therefore also include response time in our analyses, as answering a survey very quickly (i.e., speeding) has commonly been associated with satisficing (see, e.g. Zhang & Conrad, 2014).

Finally, whether respondents remember their previous responses may be linked to other factors associated with more superficial response behavior. For instance, respondents who have participated in panels for a longer time have been shown to answer questions less carefully than less experienced respondents (Couper, 2000; Schonlau & Toepoel, 2015; Toepoel et al., 2008). In addition, answering a questionnaire may be more memorable to somebody who has not done it many times before. Research on human memory suggests that 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). While within one survey Rettig et al. (2023) did not find that the newly recruited respondents were more likely to remember their responses than experienced respondents, it is unclear whether an effect may exist for a longer time between repetitions. Respondents who have answered many panel waves may be less able to remember their responses from a specific wave than respondents who have participated in fewer waves. We hypothesize:

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

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 (Blom et al., 2019, 2020b), and the information 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 general population recruited from persons living in private households in Germany aged 16 to 75 years at the time of recruitment (Blom et al., 2022; Blom et al., 2017). Respondents of the GIP are surveyed online bimonthly, with a total of 6 waves per year. Each wave takes 20–25 minutes to complete. Respondents receive conditional incentives of 4€ for each wave in

[1] 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.

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.

4.1 Experimental design

At the beginning of the GIP wave 38's questionnaire (November 2018), respondents received two questions on the topic of environmental awareness (the "test questions"; see Figure 2). Respondents were randomly assigned to receive these two questions either in the form of belief, attitude, or behavior questions. Each respondent received the two questions of the same type in a randomized order. All 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 (see Appendix Table A1 for the wordings of the 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 (claimed recall: yes/no). Depending on their answer, respondents were either asked to repeat their previous response (if "yes") or to give their best estimate (if "no"). 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 studies suggest that expressing high certainty about remembering a previous response is a good predictor of remembering it (e.g. Jaspers et al., 2009; Rettig et al., 2023), respondents were asked 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.

In another experiment, the same test questions in wave 38 were used with the same set of follow-up questions at the end of wave 38, approximately 20 minutes after respondents initially answered the test questions. Only about half of all respondents in wave 38 were randomly selected to receive the follow-up questions then. The results from that experiment are described in Rettig et al. (2023). We expand on their design in this study by repeating the same set of follow-up questions (claimed recall, correct recall, and recall certainty for each of the two test 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) was 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.

4.2 Sample and variables

In total, 4294 respondents participated in GIP's wave 38 (November 2018), which included the initial test questions. Of these, 3928 respondents (91%) also participated in wave 40 (March 2019), which included the follow-up questions. Because every respondent received a set of two test questions and respective follow-ups, we have two observations per respondent. However, we excluded incomplete observations due to (1) item nonresponse on test questions or follow-up questions (17 observations), (2) breakoff before or during the experiment (75 observations), and (3) missingness on other variables of interest (155 observations). This yielded an analytic sample of 7609 observations of 3809 respondents.

Of these 3809 respondents, 1248 (33%) received the belief questions in wave 38, 1271 (33%) received the attitude questions, and 1290 (34%) received the behavior questions. Across all question types, 1877 respondents (49%) previously received the follow-up questions in wave 38, the rest received them for the first time in wave 40. Extreme responses account for 17% of all answers. In terms of sociodemographics, 48% of respondents were female, 15% had low formal education, 29% medium-low, 22% medium-high, and 34% a high level of education, respectively. To address our hypothesis about age (H4.1), we split the sample into six groups (under 29 years, 29–38 years, 39–48 years, 49–58 years, 59–68 years, and over 68 years) based on respondents' year of birth (see Appendix Table A2 for details).

Regarding panel experience, 43% of all respondents were part of the newly recruited 2018 sample. The majority of respondents (75%) completed both waves 38 and 40 on a computer or tablet, 18% used a smartphone for both, and 8% switched between device types (e.g., from a computer to a smartphone). Most respondents (97%) also participated in the wave in-between (wave 39, January 2019). Finally, the actual time between the test questions and follow-up questions can be anywhere from 91 to 150 days, depending on when during the field times of waves 38 and 40 respondents participated (although most respondents fall around the 4-

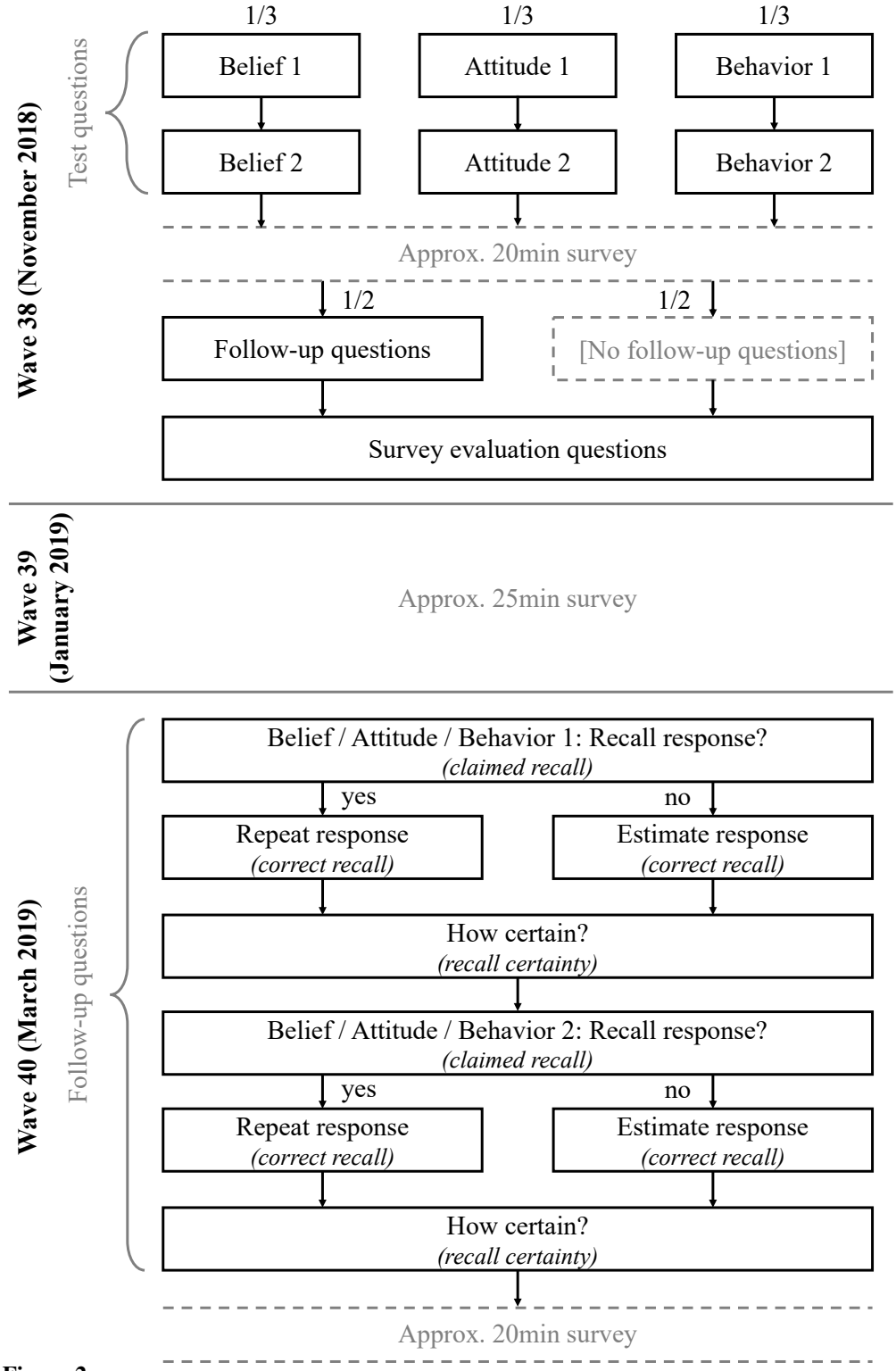

Figure 2

Illustration of the experimental design. Follow-up questions (recall or not; repeat/estimate response; certainty) 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 A1 for English translations of all questions and response scales.

month mark with a mode of 119 days and a median of 120 days). The time and date of respondents' participation as well as the time they spent to answer the test questions (response time) and the whole wave 38 questionnaire (interview length) were collected in full seconds.

To investigate the role of response burden and survey enjoyment, we created the response burden index from responses to questions about whether respondents found the survey to be "long", "difficult", and "too personal" and the survey enjoyment index from respondents' ratings of the survey as "interesting", "varied", and "relevant."[2] These survey evaluation questions are presented to respondents at the end of every GIP wave, and we used data from wave 38 to create the indices. 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).

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 the proportions of observations in which respondents correctly repeated their previous response (i.e., selected the exact same scale point) depending on whether recall was claimed (H1) both overall and by question type (H2). We then use weighted kappa statistics to study 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. 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 is the index for original responses and j for 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 1000 repetitions (Reichenheim, 2004).

To further investigate differences across question types as well as other correlates of remembering a previous response, we 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), gender and age (H4.1), education (H4.2), self-reported response burden (H5.1), survey enjoyment (H5.2), as well as the logarithmized time respondents spent answering the test question (response time) and the whole wave 38 questionnaire (interview length) as predictors. All models also include respondents' level of panel experience (newly recruited versus experienced respondents; H6). In addition, we add claimed recall as a predictor in the models on recall certainty and correct recall; and recall certainty 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) differ 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.

Our models also include several control variables. 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 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 on memory effects have so far not found differences across devices (Rettig et al., 2023; Revilla & Höhne, 2021), we control for the device type (computer or smartphone) or switches between them. Furthermore, we add the information whether respondents received the follow-ups before (in wave 38) 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 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.

[2]Exploratory factor analysis (not shown) confirmed a two-factor solution with the items "long", "difficult", and "too personal" loading highly on one factor and "interesting", "varied", and "relevant" on the other. For a validation of survey enjoyment, survey burden, and survey value as distinct concepts see also de de Leeuw et al. (2019).

5 Results

Overall, respondents claimed that they remembered their previous response in 31% of cases (Table 1). Of these, the correct response (i.e., the exact same scale point) was recalled in 34% of cases. Respondents also repeated their response correctly in 27% of cases where they had stated that they did not remember it, which is significantly lower (i.e., −7 percentage points; $t(4597) = -6.026$, $p = 0.000$)[3]. Supporting our first hypothesis (H1), this indicates that for some respondents, remembering previous responses seems to persist even after 4 months with another survey wave in-between. However, cases in which respondents claimed that they remembered their previous response and were subsequently able to correctly repeat it account for just 11% of all observations. In contrast, respondents were unable to correctly repeat their previous response in most cases (71%) and repeated their correct response despite claiming not to remember it (i.e., due to chance or an unchanged underlying information) in the remaining 19%.

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 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. Generally, the kappa statistic ranges around 0.4, indicating moderate agreement.

When further investigating the absolute difference between respondents' original response to the test question in wave 38 and their recollection in wave 40 (Figure 3), we 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 ($t(4597) = -2.096$, $p = 0.036$).

Table 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 38% for attitude questions to 22% for behavior questions. The overall proportion of correct recalls is also highest for attitude questions (32%), and significantly lower for both belief questions (27%) and behavior questions (29%; see Appendix Table A3 for t-statistics).

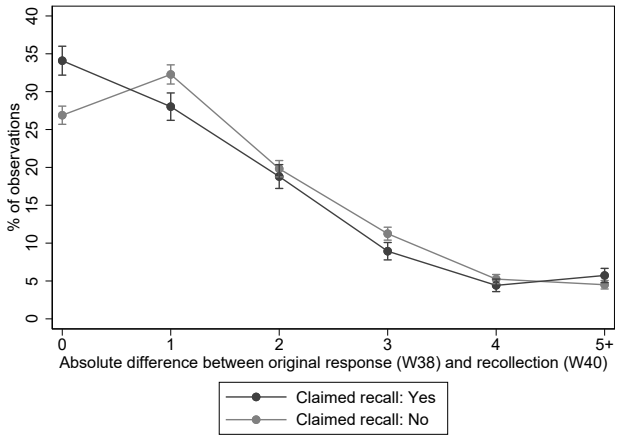

Figure 3

Differences between original response and recollection by claimed recall. 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.

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 A3). However, if no recall was claimed, the proportion of correct repetitions is significantly lower for belief questions than for attitudes ($t(1900) = 4.048$, $p = 0.000$) and behaviors ($t(2072) = 3.497$, $p = 0.001$). This finding 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. Combined, these results indicate that remembering previous responses differs across question types (H2).

The model of correct recall in Table 2 again confirms that a correct repetition of the previous response was significantly more likely when respondents claimed that they remembered their response (AME = 0.025, $p = 0.044$). This is in line with our expectation (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 but hint at persistence of memory effects across waves. Considering the interaction between claimed recall and question type, 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 ques-

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

Table 1

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

	Overall	Beliefs	Attitudes	Behaviors
Claimed recall: yes	2373 (31%)	835 (33%)	959 (38%)	579 (22%)
Of these:				
Correct recall	809 (34%)	282 (34%)	344 (36%)	183 (32%)
Weighted Kappa	0.444	0.445	0.436	0.411
CI[a]	[0.417, 0.470]	[0.403, 0.490]	[0.395, 0.481]	[0.364, 0.470]
Mean certainty[b]	6.7	6.6	6.8	6.4
Claimed recall: no	5236 (69%)	1658 (67%)	1577 (62%)	2001 (78%)
Of these:				
Correct recall	1408 (27%)	380 (23%)	465 (29%)	563 (28%)
Weighted Kappa	0.372	0.331	0.389	0.367
CI[a]	[0.356, 0.390]	[0.305, 0.364]	[0.362, 0.417]	[0.342, 0.397]
Mean certainty[b]	5.2	5.1	5.6	5.0
Overall				
Correct recall	2217 (29%)	662 (27%)	809 (32%)	746 (29%)
Weighted Kappa	0.412	0.386	0.426	0.388
CI[a]	[0.399, 0.426]	[0.363, 0.412]	[0.401, 0.449]	[0.361, 0.410]
Mean certainty[b]	5.7	5.6	6.1	5.3
Observations	7609	2493	2536	2580

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

tions than for belief questions (OR = 0.746, $p = 0.035$ and OR = 0.715, $p = 0.020$ respectively).

Across question types, compared to belief questions, recall to behavior questions was less likely claimed (AME = −0.109, $p = 0.000$). The reported certainty about remembering responses was higher for attitude questions ($b = 0.430$, $p = 0.001$). While claimed recall also correlated with higher certainty ($b = 1.349$, $p = 0.000$), the difference in certainty between respondents who did and those who did not claim recall was smaller for attitude questions ($b = -0.356$, $p = 0.026$). Recollections for attitude questions were also more likely correct (AME = 0.031, $p = 0.023$). In line with our expectations and the descriptive overview above, respondents remember their responses to questions on beliefs, attitudes, and behaviors from a previous panel wave at different rates (H2).

In line with some previous findings on extreme responses, we find that when respondents provided an extreme response (i.e., an endpoint of the response scale), they were more likely to claim that they remembered it (AME = 0.162, $p = 0.000$), reported higher certainty about remembering it ($b = 0.933$, $p = 0.000$), and were also more likely to give a correct recollection of what their response had been (AME = 0.046, $p = 0.003$). However, considering the interaction between extreme response and question type, the positive effect of extreme responses on correct repetitions vanishes for behavior questions (OR = 0.669, $p = 0.040$; see also Appendix Table A4). While these results are in line with our expectation that 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 ($b = -0.237$, $p = 0.001$) but were more likely to be correct in their recall than male respondents (OR = 1.131, $p = 0.028$). Respondents of the higher age groups (39 years and up) were more likely to claim that they remembered their response than respondents under 29 years (OR = 1.544, $p = 0.001$, OR = 1.829, $p = 0.000$, OR = 2.230, $p = 0.000$, and OR = 2.640, $p = 0.000$ respectively). With the exception of respondents over 68 years, the older age groups also reported higher certainty ($b = 0.302$, $p = 0.018$, $b = 0.431$, $p = 0.000$, and $b = 0.347$, $p = 0.006$ respectively). However, the likelihood of repeating the previous response correctly did not differ across age groups. We thus find no support 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 = 1.346, $p = 0.003$ and

OR = 1.233, $p = 0.038$ 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).

Regarding satisficing indicators, 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 = 0.688, $p = 0.028$), but we find no significant effect on the reported recall certainty or correct recall. Similarly, respondents who reported higher survey enjoyment were more likely to claim recall (OR = 1.401; $p = 0.036$) and reported higher certainty about remembering their response ($b = 0.883$, $p = 0.000$) but were not significantly more likely to correctly repeat their response. In addition, respondents who took longer to answer the test question were less likely to claim recall (OR = 0.800, $p = 0.000$) and reported lower certainty about recalling it ($b = -0.118$, $p = 0.011$), indicating that rather than speeding, a quick response to the question itself may be a sign of a salient, easily accessible judgement. Respondents who took longer to answer the whole questionnaire reported higher certainty ($b = 0.092$, $p = 0.007$). However, neither response time nor interview length is significantly correlated with correct recall. Respondents may recognize previous questions more easily when they found them to be more enjoyable and could answer them more easily but are not necessarily able to remember their response more easily. We thus do not find support for the hypotheses 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 = 1.380, $p = 0.003$) and reported higher certainty about remembering them ($b = 0.464$, $p = 0.000$) 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 = 1.460, $p = 0.001$), 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 = 1.316, $p = 0.000$) and in cases where they had been incorrect (OR = 1.223, $p = 0.011$). However, respondents who had previously received the follow-up questions also tended to report lower certainty about remembering their response ($b = -0.266$, $p = 0.000$ and $b = -0.401$, $p = 0.000$ respectively). If respondents correctly recalled their response in the previous wave, they were more likely correct again in the later wave (OR = 1.247, $p = 0.000$). 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. Finally, respondents were more likely to claim that they remembered their response in the first set of follow-up questions they received (OR = 1.403, $p = 0.000$) but also reported lower certainty about remembering it ($b = -0.136$, $p = 0.000$), while the likelihood of correct recalls did not significantly differ. 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). These models did not yield substantially different results.

6 Discussion and conclusion

In this study, we examined the ability of respondents from a probability-based online panel to remember their responses from a previous panel wave after 4 months. Table 3 provides a summary of our hypotheses and results.

Overall, respondents chose their correct previous response in about 29% of cases. This finding is in line with a downward trend in correct recalls over time compared to the 60–88% other studies have reported for questions readministered within one survey (Rettig et al., 2023; Revilla & Höhne, 2021; Schwarz et al., 2020; van Meurs & Saris, 1990), and the roughly 40% van van Meurs and 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 of 7 percentage points. Such 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 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

Table 2

Logistic and linear regression models of claimed recall, recall certainty, and correct recall

	Claimed recall				Recall certainty		Correct recall			
	OR[a]	Std.Err.[b]	AME[c]	Std.Err.[b]	Coef.[d]	Std.Err.[b]	OR[a]	Std.Err.[b]	AME[c]	Std.Err.[b]
Question type (ref.: beliefs)										
Attitudes	1.034	0.111	0.023	0.016	0.430***	0.129	1.187	0.127	0.031*	0.013
Behaviors	0.566***	0.065	−0.109***	0.015	−0.106	0.130	1.368**	0.141	0.025	0.014
Extreme response	2.015***	0.256	0.162***	0.017	0.933***	0.159	1.347*	0.166	0.046**	0.016
Extreme response × question type (ref.: non-extreme, beliefs)										
Attitudes	1.098	0.183	-	-	0.174	0.203	1.179	0.192	-	-
Behaviors	1.050	0.211	-	-	−0.492	0.258	0.669*	0.131	-	-
Female	0.916	0.059	−0.017	0.013	−0.237***	0.070	1.131*	0.063	0.024*	0.011
Age (ref.: <29 years)										
29–38 years	1.248	0.159	0.037	0.021	0.148	0.128	1.093	0.117	0.017	0.021
39–48 years	1.544***	0.193	0.077***	0.022	0.302*	0.128	1.169	0.125	0.031	0.021
49–58 years	1.829***	0.222	0.111***	0.021	0.432***	0.120	1.123	0.114	0.023	0.020
59–68 years	2.230***	0.281	0.152***	0.023	0.347**	0.125	1.008	0.107	0.002	0.021
>68 years	2.640***	0.381	0.190***	0.028	0.256	0.153	1.055	0.134	0.010	0.025
Education (ref.: low)										
Medium-low	1.346**	0.136	0.058**	0.019	0.112	0.115	0.967	0.085	−0.007	0.017
Medium-high	1.127	0.124	0.023	0.021	0.052	0.122	1.121	0.106	0.023	0.019
High	1.233*	0.124	0.040*	0.019	0.171	0.114	1.043	0.090	0.008	0.017
Response burden (W38)	0.688*	0.117	−0.074*	0.034	−0.060	0.186	0.816	0.117	−0.040	0.029
Survey enjoyment (W38)	1.401*	0.225	0.067*	0.032	0.883***	0.175	0.821	0.108	−0.039	0.026
Log response time (test question)	0.800***	0.046	−0.044***	0.011	−0.118*	0.046	0.924	0.044	−0.016	0.009
Log interview length (W38)	1.043	0.032	0.008	0.006	0.092**	0.034	1.026	0.026	0.005	0.005
Newly recruited respondent	1.380**	0.149	0.072***	0.013	0.464***	0.125	0.861	0.088	−0.021	0.011
Newly recruited × question type (ref.: newly recruited resp., beliefs)										
Attitudes	1.130	0.168	-	-	−0.132	0.170	1.132	0.156	-	-
Behaviors	0.974	0.156	-	-	0.059	0.170	1.008	0.137	-	-

Continues on next page

Continued from last page

	Claimed recall				Recall certainty		Correct recall			
	OR[a]	Std.Err.[b]	AME[c]	Std.Err.[b]	Coef.[d]	Std.Err.[b]	OR[a]	Std.Err.[b]	AME[c]	Std.Err.[b]
Device (ref.: both waves computer)										
Both waves smartphone	1.146	0.105	0.027	0.019	−0.090	0.098	0.917	0.073	−0.017	0.016
Device switch	1.460**	0.171	0.078**	0.025	0.017	0.123	0.986	0.105	−0.003	0.021
Follow-ups W38 (ref.: no follow-ups)										
Correct recall in W38	1.316***	0.092	0.055***	0.014	−0.266***	0.076	1.247***	0.077	0.045***	0.013
Incorrect recall in W38	1.223*	0.097	0.040*	0.016	−0.401***	0.086	0.897	0.068	−0.021	0.014
Days between waves	1.002	0.004	0.000	0.001	−0.004	0.004	1.004	0.003	0.001	0.001
In-between wave (W39)	0.882	0.159	−0.025	0.036	0.054	0.211	1.036	0.172	0.007	0.033
First question	1.403***	0.061	0.067***	0.008	−0.136***	0.035	1.005	0.052	0.001	0.010
Claimed recall (ref.: no)	-	-	-	-	1.349***	0.118	1.402***	0.143	0.025*	0.012
Claimed recall × question type (ref.: claimed recall: no, beliefs)										
Attitudes	-	-	-	-	−0.356*	0.160	0.746*	0.104	-	-
Behaviors	-	-	-	-	0.007	0.174	0.715*	0.103	-	-
Recall certainty	-		-		-	-	1.126***	0.014	0.024***	0.002
Constant	0.159**	0.090	-	-	4.342***	0.629	0.095***	0.047	-	-
Pseudo-R^2 McKelvey & Zavoina		0.105					0.054			
R^2 Adj.					0.131					
Observations		7609			7609			7609		

Two observations per respondent. See also Appendix Figure B1 for a coefficient plot of the presented models.

[a] Odds Ratios from logistic regression models [b] Cluster-robust standard errors [c] Average marginal effects from logistic regression models

[d] Unstandardized linear regression coefficients,

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

Table 3

Summary of our hypotheses and results

Hypothesis		Supported?
H1:	After four months, respondents who self-report remembering their response are more likely to correctly recall 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)[a]
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

[a] for belief and attitude questions, not for behavior questions.

effects persist. Indeed, the 7 percentage point difference we found after 4 months is smaller than the roughly 20 percentage points reported by Rettig et al. (2023) for the same test questions after 20 minutes, 17 percentage points in Schwarz et al. (2020) and 34 percentage points in van Meurs and Saris (1990). This again points towards some memory of previous responses persisting even after 4 months, albeit much less than within the same survey. While Revilla and Höhne (2021) reported a similarly small difference within one survey, this may be due to the very high proportion of correct repetitions across both groups in their experiment.

Additionally, we found that in almost 90% of cases respondents were either unable to correctly repeat their previous response or stated that they could not remember it. Respondents who could not correctly recall their previous response were also most commonly off by just one scale point. Combined, 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 the resulting measurement error) are likely very small. Additionally, 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. 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.

Regarding question types, we found that responses to belief questions were correctly repeated least often overall but had the largest difference between respondents who said they could and those who could not remember their response. This finding is in contrast to the one of Rettig et al. (2023) that 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. (2023) argued, differences across question types may be driven by their differences in accessibility and stability. However, as van Meurs and Saris (1990) pointed out, most respondents will likely not change their opinions or behaviors during 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.

Noteworthy, our results are based on a set of questions which respondents were asked for the first time which had a goal of avoiding a confounding of different effects in our experiment. However, measurements in longitudinal studies are often repeated at regular intervals. As a link between the repetition of information and its retention in memory exists (see, e.g. Hintzman, 1976), memory effects in repeated survey measurements may also be 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 indicate that this repeated presentation and additional attention towards the test questions made respondents more likely to recognize them or

to make cues more accessible when retrieving relevant information during the response process. This links our research to studies on panel conditioning (i.e., learning effects that occur across panel waves). Panel conditioning can occur due to respondents learning the content of the questionnaire or learning the procedure. The negligible differences that we find in actual recall despite greater claimed recall (and it being greater for the less experienced respondents) serve as evidence of the mechanism of respondents learning the survey procedure and provide additional insight on why panel conditioning effects that have been found were small or nonexistent—respondents might not actually remember their responses (Struminskaya & Bosnjak, 2021). A further investigation into memory effects in the context of questions which have been frequently repeated to respondents would be a useful avenue for future research.

Respondents who previously received the same follow-up questions also present a concern 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 distinguish whether some respondents may have recalled their responses to the previous follow-up questions instead of their actual previous response. This is especially true in cases when the correct response was recalled both times: Correctly recalling the initial response twice and correctly recalling the previous recollection identical to the initial response would lead to identical results.

Finally, we found neither an effect of the number of days between the administration of the test questions and follow-up questions (from 91 to 150 days), nor of the participation in the in-between panel wave on respondents' recall ability. However, the time interval between the participations or participation in the in-between wave were not randomly assigned. Additionally, we analyzed data from respondents who participated in both the wave with the test questions (wave 38) and the wave with the follow-up questions (wave 40). 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 fully translate to respondents who participate infrequently or to newly recruited respondents who dropped out of the panel soon after their recruitment. While we could expect an additional panel wave to disrupt respondents' memory of previous responses, our findings are in line with recent research where a longer time to answer the questionnaire and memory interference tasks did not 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 be required to fully investigate the effects of these factors on memory effects in a longitudinal context.

Overall, our results have important implications 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 these groups do not differ in a number of characteristics, including age, education level, panel experience or the device used for survey completion. After four or more months, memory effects will likely affect only a small number of respondents, not substantially, and these respondents 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. This is good news for studies that wish to avoid memory effects when repeatedly administering questions (e.g., for attitudes) but not as good news for studies in which memory may be desirable (e.g., factual questions and assessing change). Our study has provided a basis for future research that should further investigate the time intervals after which memory effects are no longer present and whether they differ for different questions and respondents. Future studies should focus on practical measures to avoid memory effects and how to stimulate them. An extension of our study can be a randomized experiment combining asking to recall previous answers and dependent interviewing for different question types, with a potential goal of developing targeted dependent interviewing strategies that vary by type of question and interval between repetitions.

References

Alwin, D. F. (2007). *Margins of error. A study of reliability in survey measurement*. Wiley.

Alwin, D. F. (2010). How good is survey measurement? Assessing the reliability and validity of survey measure. 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). Wiley.

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]. 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]. 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]. https://doi.org/10.4232/1.13463

Blom, A. G., Gathmann, C., & Krieger, U. (2022). Setting up an online panel representative of the general population. *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 multitrait-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.

Cannell, C. F., & Fowler, F. J. (1965). Comparison of hospitalization reporting in three survey procedures. *National Center for Health Statistics. Vital Health Stat*, *2*(8), 1–17.

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

de Leeuw, E., Hox, J., Silber, H., Struminskaya, B., & Vis, C. (2019). Development of an international survey attitude scale: Measurement equivalence, reliability, and predictive validity. *Measurement Instruments for the Social Sciences*, *1*(1). https://doi.org/10.1186/s42409-019-0012-x

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. Wiley.

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

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Addison-Wesley.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

Hintzman, D. L. (1976). Repetition and memory. In *Psychology of learning and motivation* (pp. 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 & Quantity*, *56*(4), 2549–2566. https://doi.org/10.1007/s11135-021-01219-2

Hoogendoorn, A. W. (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. (2009). Dependent interviewing: A framework and application to current research. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 93–111). Wiley. 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*, *9*(3), 477–495. 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. *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). Wiley. https://doi.org/10.1002/9780470743874.ch1

Mathiowetz, N. A., & McGonagle, K. A. (2000). An assessment of the current state of dependent interviewing in household surveys. *Journal of Official Statistics*, *16*(4), 401–418.

McKelvie, S. J. (1992). Does memory contaminate test-retest reliability? *The 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., Blom, A. G., & Höhne, J. K. (2023). Memory effects: A comparison across question types. *Survey Research Methods*, *17*(1), 37–50. https://doi.org/10.18148/SRM/2023.v17i1.7903

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

Revilla, M., Höhne, J. K., & Rettig, T. (2023). Differences in measurement quality depending on recall: Results for a question about trust in the parliament. *Quality & Quantity*, *57*(3), 2125–2146. https://doi.org/10.1007/s11135-022-01441-6

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.). Wiley.

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 contex*. 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* (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). Wiley.

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.

Toepoel, V., Das, M., & Soest, A. V. (2008). Effects of design in web surveys. *Public Opinion Quarterly*, *72*(5), 985–1007. https://doi.org/10.1093/poq/nfn060

Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education*, *26*(2), 169–181. https://doi.org/10.1108/qae-06-2017-0034

Tourangeau, R., Maitland, A., Rivero, G., Sun, H., Williams, D., & Yan, T. (2017). Web surveys by smartphone and tablets. *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.

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

van Meurs, A., & Saris, W. E. (1990). Memory efects in MTMM studies. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies* (pp. 160–167). 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). Wiley. https://doi.org/10.1002/9781119263685.ch8

Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*(2), 127–135. https://doi.org/10.18148/srm/2014.v8i2.5453

Appendix A
Tables

Table A1

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)

Questions fielded in German, own translation.

Table A2

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)	
over 68 years	1935–1939
	1940–1944
	1945–1949
59 to 68 years	1950–1954
	1955–1959
49 to 58 years	1960–1964
	1965–1969
39 to 48 years	1970–1974
	1975–1979
29 to 38 years	1980–1984
	1985–1989
under 29 years	1990–1994
	1995–1999
	2000 and later

Table A3

T-tests for differences in claimed recall and correct recall across question types

	% of observations			*t*-test (two-tailed, cluster-adjusted)		
	Beliefs	Attitudes	Behaviors	*t*	*df*	*p*
Claimed recall: yes	33.5	37.8	-	2.607	2517	0.009
	33.5	-	22.4	−7.186	2536	0.000
	-	37.8	22.4	9.944	2559	0.000
Correct recall						
...overall	26.6	31.9	-	3.848	2517	0.000
	26.6	-	28.9	1.781	2536	0.075
	-	31.9	28.9	2.193	2559	0.028
...if claimed recall: yes	33.8	35.9	-	0.845	1180	0.398
	33.8	-	31.6	−0.811	950	0.418
	-	35.9	31.6	1.605	1032	0.109
...if claimed recall: no	22.9	29.5	-	4.048	1900	0.000
	22.9	-	28.1	3.497	2072	0.001
	-	29.5	28.1	0.856	2052	0.392

Table A4

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

	% of observations		*t*-test (two-tailed, cluster-adjusted)		
Correct recall	Extreme	Non-extreme	*t*	*df*	*p*
Beliefs	36.3	24.6	−4.651	1422	0.000
Attitudes	44.0	28.0	−6.927	1494	0.000
Behaviors	29.7	28.8	−0.276	1451	0.783
Overall	38.8	27.2	−7.889	4371	0.000

Table A5

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	1.124	0.108	0.815	0.145	1.256*	0.144
Behaviors	1.194	0.113	0.986	0.184	1.282*	0.141
Extreme response	1.632***	0.210	3.386***	0.650	0.630*	0.131
Extreme response × question type (ref.: non-extreme, beliefs)						
Attitudes	1.131	0.189	0.859	0.216	1.720*	0.446
Behaviors	0.603*	0.123	0.497*	0.161	1.053	0.311
Female	1.097	0.061	1.069	0.106	1.109	0.075
Age (ref.: <29 years)						
29–38 years	1.127	0.118	1.294	0.277	1.054	0.129
39–48 years	1.238*	0.130	1.140	0.246	1.252	0.154
49–58 years	1.219*	0.122	1.239	0.249	1.131	0.135
59–68 years	1.098	0.115	1.110	0.230	1.034	0.130
>68 years	1.147	0.144	1.192	0.280	1.014	0.155
Education (ref.: low)						
Medium-low	0.998	0.088	0.832	0.129	1.044	0.110
Medium-high	1.140	0.108	1.051	0.176	1.154	0.131
High	1.083	0.094	0.866	0.137	1.131	0.117
Response burden (W38)	0.797	0.114	0.847	0.224	0.795	0.139
Survey enjoyment (W38)	0.932	0.121	0.859	0.206	0.930	0.145
Log response time (test question)	0.902*	0.043	0.923	0.082	0.919	0.051
Log interview length (W38)	1.040	0.026	1.059	0.046	1.020	0.032
Newly recruited respondent	0.938	0.094	0.733	0.124	0.980	0.123
Newly recruited × question type (ref.: newly recruited resp., beliefs)						
Attitudes	1.100	0.150	1.374	0.314	1.066	0.185
Behaviors	0.995	0.134	1.212	0.300	0.941	0.153
Device (ref.: both waves computer)						
Both waves smartphone	10.914	0.072	0.948	0.140	0.882	0.083
Device switch	1.008	0.106	1.061	0.196	0.960	0.131
Follow-ups W38 (ref.: no follow-ups)						
Correct recall in W38	1.222**	0.075	1.452***	0.157	1.075	0.082
Incorrect recall in W38	0.867	0.064	0.727*	0.100	0.928	0.081
Days between waves	1.004	0.003	1.011	0.006	1.001	0.004
In-between wave (W39)	1.019	0.167	1.109	0.325	1.018	0.205
First question	1.008	0.052	0.926	0.084	1.007	0.065
Constant	0.181***	0.089	0.091**	0.084	0.288*	0.167
Pseudo-R^2 McKelvey & Zavoina	0.027		0.101		0.016	
Observations	7609		2373		5236	

Two observations per respondent.

[a] Odds Ratios from logistic regression models [b] Cluster-robust standard errors

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

Appendix B
Figures

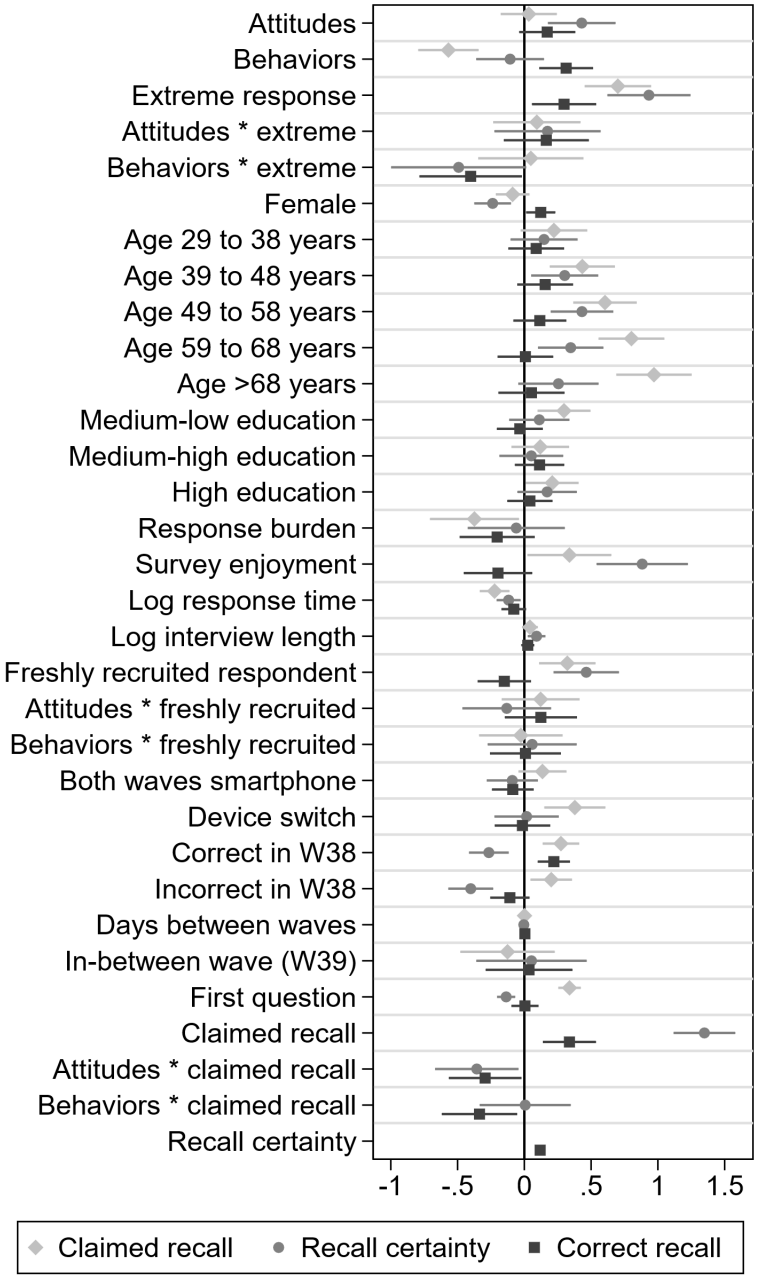

Figure B1

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