Survey Research Methods (2023)
Vol. 17, No. 1, pp. 37-50
doi:10.18148/srm/2023.v17i1.7903
European Survey Research Association

Memory Effects: A Comparison Across Question Types

Tobias Rettig[1,2], Annelies G. Blom[3,4], and Jan Karem Höhne[5,6]
[1]University of Mannheim, Mannheim Center for European Social Research (MZES)
[2]University of Mannheim, School of Social Sciences
[3]University of Bremen, SOCIUM Forschungszentrum Ungleichheit und Sozialpolitik
[4]University of Bergen, Department of Government
[5]University of Duisburg-Essen, Institute for Political Science
[6]Universitat Pompeu Fabra, RECSM

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

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,

Corresponding author: 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).

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; Alwin, 2010; Alwin, 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.

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).[1] In a follow-up study three decades later, Schwarz et al. (2020) conducted a lab experiment with a sample of college students and found that 60% of respondents correctly reproduced their previous answer (to a single

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

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; Tourangeau et al., 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.

3 Methods

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 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).

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 1). These questions were based on three questions regarding respondents' beliefs and behaviors on climate change and energy use from the 8$^{\text{th}}$ round of the European Social Survey (European Social Survey, 2016, see Appendix Table A1 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 1 displays the experimental design

Table 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

Questions fielded in German, own translation.

(see Appendix Table A2 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.

4 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 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=0.605), gender (p=0.256), education (p=0.670), device (p=0.365), and recruitment sample (p=0.981; see Appendix Table A3 for the χ^2-statistics). Thus, we confirmed a uniform distribution of the sample across the experimental groups regarding

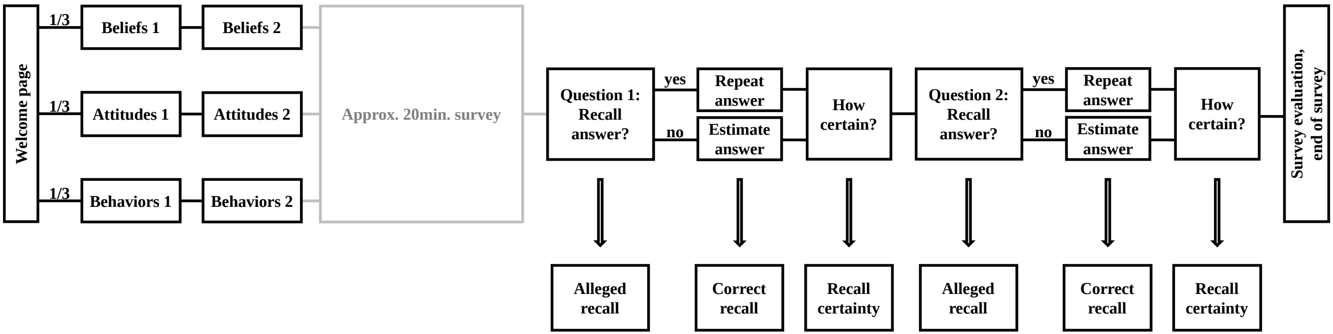

Figure 1

Experimental design. 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)

these basic respondent characteristics.

4.1 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.[2] 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 response time in seconds), and the time between test questions and follow-up questions (server-side in-between time in seconds).

[2] 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.

5 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 2).

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=0.098; see Appendix Table A4 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=0.000; see Appendix Table A4 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=0.000; see Appendix Table A4).

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 3 reports the descriptive results.

The proportion of alleged recall is very high for cases with extreme answers (96.3%). This is significantly higher than for cases with non-extreme answers (p = 0.000; see Appendix Table A5 for the χ^2-statistics). The differences across question types are significant (p = 0.041; see Appendix Table A6) 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 = 0.000; see Appendix Table A6). The overall proportion of correct recalls is also significantly higher for cases with extreme answers than for cases with non-extreme answers (p = 0.000; see Appendix Table A5).

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=0.001; see Appendix Table A6). The overall mean recall certainty is significantly higher for cases with extreme answers than for cases with non-extreme answers (p=0.000; see Appendix Table A5). 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 displays the results for all three models.

The first Model in Table 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.

Table 2

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

	Overall		Beliefs		Attitudes		Behaviors	
	N	%	N	%	N	%	N	%
Observations	3,711	100	1,229	100	1,230	100	1,252	100
Alleged recall: yes	3,124	84	1,024	83	1,058	86	1,042	83
Of these: Correct recall	1,997	64	558	55	720	68	719	69
Mean certainty[a]	7.6		7.1		7.9		7.7	
Alleged recall: no	587	16	205	17	172	14	210	17
Of these: Correct recall	258	44	91	44	71	41	96	46
Mean certainty[a]	5.5		5.3		5.6		5.7	
Overall correct recall	2,255	61	649	53	791	64	815	65
Overall mean certainty[a]	7.3		6.8		7.6		7.4	

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

Table 3

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

	Overall		Beliefs		Attitudes		Behaviors	
	N	%	N	%	N	%	N	%
Observations	614	17	188	15	315	26	111	9
Alleged recall: yes	591	96	176	94	305	97	110	99
Alleged recall: no	23	4	12	6	10	3	1	1
Overall correct recall	523	85	143	76	281	89	99	89
Overall mean certainty[a]	9.2		8.8		9.3		9.4	

[a] On an 11-point scale from 0 “not at all certain” to 10 “absolutely certain”

The second Model in Table 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.

Finally, the third Model in Table 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.[3] 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 of 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

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

Table 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	0.805	0.154	0.258	−0.435	0.306	0.155	0.857	0.209	0.526
Attitudes	0.848	0.167	0.403	−0.124	0.301	0.681	0.598	0.151	0.042
Extreme response	5.622	1.278	0.000	2.065	0.095	0.000	2.861	0.397	0.000
Experienced panelist	0.624	0.121	0.015	0.156	0.164	0.342	0.775	0.105	0.061
Experienced × question type									
Beliefs	1.446	0.395	0.178	−0.171	0.240	0.476	1.203	0.233	0.339
Attitudes	1.546	0.437	0.123	−0.189	0.218	0.386	1.571	0.310	0.022
Age (ref.: <44 years)									
44–58 years	1.774	0.265	0.000	0.045	0.119	0.702	0.782	0.081	0.018
>58 years	1.884	0.303	0.000	−0.252	0.134	0.060	0.751	0.085	0.012
School degree (ref.: basic/none)									
Vocational	1.300	0.240	0.156	0.090	0.165	0.587	1.357	0.172	0.016
High school	0.964	0.165	0.829	0.201	0.155	0.196	1.347	0.162	0.013
Female	0.809	0.093	0.066	−0.018	0.095	0.849	1.322	0.107	0.001
Smartphone respondent	1.124	0.164	0.425	−0.103	0.122	0.396	0.883	0.091	0.227
First question	1.555	0.103	0.000	−0.129	0.046	0.005	0.864	0.059	0.031
Response time	0.998	0.001	0.105	0.000	0.001	0.627	1.001	0.001	0.467
In-between time	1.000	0.000	0.228	0.000	0.000	0.086	1.000	0.000	0.709
Alleged recall	-	-	-	1.821	0.190	0.000	1.693	0.282	0.002
Alleged recall × question type									
Beliefs	-	-	-	−0.191	0.290	0.510	0.578	0.140	0.024
Attitudes	-	-	-	0.023	0.300	0.939	1.086	0.279	0.748
Recall certainty	-	-	-	-	-	-	1.266	0.023	0.000
Constant	3.707	0.913	0.000	5.770	0.279	0.000	0.213	0.054	0.000
Pseudo-$R^2_{\text{McKelvey \& Zavoina}}$		0.161					0.208		
$R^2_{\text{adj.}}$					0.197				
Observations		3,711			3,711			3,711	

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.

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.

6 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 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, 2022, 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, 2022). 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, 2022; 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.

Acknowledgements

This article uses data from the wave 38 of the German Internet Panel (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.

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 measures. In P. V. Marsden & J. Wright (Eds.), *Handbook of survey research* (2nd, 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; 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. (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; 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. https://www.europeansocialsurvey.org/docs/round8/fieldwork/source/

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

Höhne, J. K. (2022). New insights on respondents' recall ability and memory effects when repeatedly measuring political efficacy. *Quality and Quantity*, *56*. 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). Wiley. 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). John Wiley; 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
Tables

Table A1

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

See European Social Survey (2016, pp. 30, 37).

Table A2

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

Questions fielded in German, own translation.

Table A3

Chi-squared tests of differences across experimental groups

	χ^2	df	p
Age	23.50	26	0.605
Gender	2.73	2	0.256
School degree	9.38	12	0.670
Device	2.01	2	0.365
Recruitment sample	0.42	4	0.981

Table A4

Chi-squared tests and one-way ANOVA for differences across question types

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

Table A5

Chi-squared tests and one-way ANOVA for differences between extreme and non-extreme answers

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

Table A6

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

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