Quality & Quantity (2022) 56:2549–2566
https://doi.org/10.1007/s11135-021-01219-2

New insights on respondents' recall ability and memory effects when repeatedly measuring political efficacy

Jan Karem Höhne[1,2]

Accepted: 12 August 2021 / Published online: 14 September 2021

Abstract
Many study designs in social science research rely on repeated measurements implying that the same respondents are asked the same (or nearly the same) questions at least twice. An assumption made by such study designs is that respondents second answer does not depend on their first answer. However, if respondents recall their initial answer and base their second answer on it memory effects may affect the survey outcome. In this study, I investigate respondents' recall ability and memory effects within the same survey and randomly assign respondents to a device type (PC or smartphone) and a response format (response scale or text field) for reporting their previous answer. While the results reveal no differences regarding device types, they reveal differences regarding response formats. Respondents' recall ability is higher when they are provided with the response scale again than when they are only provided with a text field (without displaying the response scale again). The same finding applies to the size of estimated memory effects. This study provides evidence that the size of memory effects may have been overestimated in previous studies.

Keywords Memory effects · Mixed-device survey · Political efficacy · Recall ability · Repeated measurement · Response format

1 Introduction and background

The measurement of political concepts, such as political efficacy, is a key task in social science research and many adjacent research fields, as it allows researchers to draw conclusions about peoples' political attitudes and voting intentions. However, researchers are often not only interested in measuring such political concepts at one single point in time, as is the case in cross-sectional studies, but also over time, as is the case in longitudinal studies with repeated measurements. Panel studies ask the same respondents the same questions at least twice to investigate attitudinal changes over time. For example, this is done in

✉ Jan Karem Höhne
jan.hoehne@uni-due.de

[1] Department of Political Science, University of Duisburg-Essen, Duisburg, Germany

[2] Research and Expertise Centre for Survey Methodology (RECSM), Universitat Pompeu Fabra, Barcelona, Spain

the pre- and post-election survey of the American National Election Study (ANES) using a set of questions on political efficacy.

The observation of change over time is also crucial in experimental research. For example, classical pretest–posttest designs use repeated measurements. The first measurement takes place before a treatment and the second measurement takes place after a treatment. These two measurements are used to infer the impact of the treatment on the respondents under investigation (Dimitrov and Rumrill 2003; Shadish et al. 2002).

Repeated measurements are also important for assessing the measurement quality, such as reliability and validity (Saris and Gallhofer 2014), of data collection methods. In other words, the researcher usually must ask the same respondents the same (or nearly the same) questions at least twice. For example, this is required when using a test–retest design to estimate reliability (Alwin 2010, 2011) and when using a multitrait-multimethod (MTMM) design to estimate reliability and validity (Campbell and Fiske 1959; Saris and Gallhofer 2014; Saris et al. 2004). To avoid attitudinal changes between the repeated measurements that may affect the estimation of measurement quality, the repetitions are commonly conducted within the same survey. For example, the European Social Survey (ESS) regularly conducts MTMM experiments in its rounds by repeating some questions from the beginning of the survey at the very end (see Revilla et al. 2014; Saris et al. 2010). The time interval between a question and its repetition is approx. one hour, depending on the routing and respondents' response speed.

An assumption made by study designs with repeated measurements is that respondents second answer to a question is not affected by their first answer. In general, respondents are assumed to go through each step of the response process again. However, there is a good chance that respondents can draw on their previous answer. This means that they do not complete the response process again but simply repeat their previous answer. If this is the case, then the two measures are not independent of each other. Memory effects have the potential to reduce the precision of parameter estimations (Saris et al. 2010) resulting, for example, in an underestimation of attitudinal change over time or an overestimation of measurement quality.[1]

Although all studies consisting of repeated measurements face the problem of memory effects, there is little research on respondents' recall ability and memory effects. This is particularly true for repeated measurements within the same survey (Revilla and Höhne 2020; Schwarz et al. 2020). For example, van Meurs and Saris (1990) investigated to what extend respondents correctly reproduce their answers to previous questions within the same survey using data from the probability-based NIPO Telepanel. Respondents answered three questions on political parties with a ten-point, end verbalized response scale on a screen without the presence of an interviewer. Van Meurs and Saris (1990) reported that approx. 70% of the respondents who stated that they recall their answer were indeed able to

[1] Even cross-sectional surveys are not free of the threat of memory effects. Since the number of surveys—particularly web surveys—has increased in recent years (Couper 2017), the likelihood that a respondent answers the same (or similar) questions within a short period has dramatically increased. For example, it is possible that a respondent would take part in two independent surveys conducted by the same or different survey companies within a short period (hours or days) that employ similar or even the same questions. This is more likely to occur just before important political events, such as elections, a period during which numerous (scientific) institutions run surveys. The survey that asks a respondent the question the second time may obtain results that are influenced by memory effects.

reproduce it correctly approx. 9 min after the first time the question was asked. In contrast, approx. 36% of the respondents who stated that they cannot recall their answer reproduced it correctly. The authors estimated that approx. 34% (70–36%) of the respondents reproduced their answers correctly because of memory. The other respondents reproduced their answers correctly because of attitude stability or by chance.

Van Meurs and Saris (1990) discovered three variables that affect respondents' recall ability. First, presenting questions with similar topics as the test questions (i.e., the questions that are repeated to determine recall ability and memory effects) and their repetitions in between decreased respondents' likelihood to correctly reproduce their answers. Second, respondents who give extreme answers (i.e., selecting categories at the endpoints of the scale) have a higher likelihood to correctly reproduce their answers. Finally, increasing the in-between time (i.e., the time between the test questions and their repetitions) reduced respondents' likelihood to correctly reproduce their previous answers.

In a subsequent study, Schwarz et al. (2020) conducted a web survey experiment (PC only) in a lab setting using mostly university students in Spain. Respondents answered one question on the difficulty/easiness of dealing with important problems in life using an eleven-point, end verbalized response scale. The authors reported that approx. 60% of the respondents correctly reproduced their previous answer after approx. 20 min. They used the estimation procedure proposed by van Meurs and Saris (1990) and estimated that approx. 17% of the respondents correctly reproduced their previous answer due to memory. This figure is half as large as the one obtained by van Meurs and Saris (1990). Also, Schwarz et al. (2020) did not find evidence that increasing the in-between time decreases respondents' recall ability and memory effects.

In addition, Rettig et al. (2020) and Revilla and Höhne (2020) conducted two independent web survey experiments (no device limitations) in the probability-based German Internet Panel. While Rettig et al. (2020) asked respondents two questions on environmental issues using an eleven-point, end verbalized response scale, Revilla and Höhne (2020) asked respondents one question on political interest using a five-point, completely verbalized response scale. In both studies, the time between the test questions and their repetitions was approx. 20 min. Similar to Schwarz et al. (2020), Rettig et al. (2020) obtained an estimated memory effect of approx. 20%, while the estimated memory effect obtained by Revilla and Höhne (2020) was approx. 7%. This is substantially smaller than the effects obtained in the other studies. The two studies by Rettig et al. (2020) and Revilla and Höhne (2020) used the estimation procedure proposed by van Meurs and Saris (1990). In addition, both studies did not find a significant effect for the time between a question and its repetition on correct reproduction. However, Rettig et al. (2020) found that respondents who give extreme answers have a higher likelihood to correctly reproduce their answers.

Taking a closer look at the studies on respondents' recall ability and memory effects within the same survey, it is observable that there are substantial methodological differences between them. For example, the studies differ regarding survey mode. While van Meurs and Saris (1990) used an early form of a "computer-assisted self-interview" (CASI), the other studies used contemporary self-administered web surveys that differed regarding device type (i.e., PC only vs. no device limitation). The studies also differ regarding survey

Table 1 Research questions

RQ1a	What is the prevalence of respondents stating that they recall their previous answer (stated recall)?
RQ1b	What is the prevalence of respondents correctly reproducing their previous answer (correct reproduction)?
RQ1c	How certain are respondents about recalling their previous answer (recall certainty)?
RQ2	What variables are associated with stated recall (RQ2a), correct reproduction (RQ2b), and recall certainty (RQ2c)?
RQ3	What is the prevalence of correct reproduction due to memory?

setting (i.e., field vs. lab), sample characteristics (i.e., convenience sample vs. probability-based sample), country (i.e., Germany vs. Netherlands vs. Spain), and recall material (i.e., five-point, completely verbalized scales vs. ten- or eleven-point, end verbalized scales).

However, the studies also have something in common. They presented the response scales when they asked respondents whether they can correctly reproduce their previous answers.[2] This circumstance may impede a proper evaluation of respondents' recall ability and memory effects. The reason is that response scales are not simply "measurement devices". The response categories build a "frame of reference" (Schwarz et al. 1985) on which respondents can draw. For example, if respondents cannot correctly reproduce their previous answer (e.g., it depends), but see it among the categories provided, this may help them to come up with the correct answer. Thus, an overestimation of respondents' recall ability and memory effects is likely to occur.

I build on the scarcity of studies and investigate respondents' recall ability and memory effects within the same survey. This is done by having different experimental groups: some are presented with the response scale again for selecting the previous answer and some others are provided with a text field for entering the previous answer (without displaying the response scale again). Respondents are also randomly assigned to a device type (PC or smartphone) to test for device-related differences. In addition, I collect paradata to account for completion conditions, such as on-device media multitasking.

2 Research questions

This experimental study contributes to the small body of research on respondents' recall ability (consisting of the three facets stated recall, correct reproduction, and recall certainty) and memory effects within the same survey. In accordance with the state of research and the existing knowledge gaps outlined above, I address seven research questions. Table 1 displays the research questions.

This study builds on the scarce literature on respondents' recall ability and memory effects within the same survey (see Rettig et al. 2020; Revilla and Höhne 2020; Schwarz et al. 2020; van Meurs and Saris, 1990) by addressing some of their methodological limitations and considering new aspects. First, I investigate the effect of using different response

[2] Van Meurs and Saris (1990) do not clearly state how correct reproduction was measured. I got in touch with the second author who confirmed that the response scale was displayed again when measuring correct reproduction.

formats (response scale or text field) for reporting answers to previous questions. This allows me to draw more accurate and robust conclusions about respondents' recall ability and memory effects. Second, I randomly assign respondents to a device type (PC or smartphone) for web survey completion to account for device-related differences. The previous studies restricted survey completion to a specific device type or had no device type randomization at all. Third, in contrast to the existing studies, I use a quota sample drawn from an opt-in access panel, which is one of the most frequently used data sources in web survey research. Fourth, I account for on-device media multitasking and its effects on respondents' recall ability and memory effects. This is done by collecting paradata about window and browser tab switching during web survey completion (see Höhne and Schlosser 2018; Höhne et al. 2020; Sendelbah et al. 2016). Finally, I use a test question on political efficacy, a topic in social science research and adjacent research fields that has not been investigated in previous studies on respondents' recall ability and memory effects.

3 Method

3.1 Questions used in this study

Test question: The question I use to investigate respondents' recall ability and memory effects was adopted from Beierlein et al. (2012). It dealt with internal political efficacy and asked how well respondents understand and assess political issues. The test question was individually presented on the web survey page and had a vertically aligned response scale with the following five categories: agree strongly, agree, it depends, disagree, and disagree strongly.

Follow-up questions: To determine respondents' recall ability and memory effects, I adopted some follow-up questions from the study of van Meurs and Saris (1990). Each question captures one of the three facets that constitute my definition of recall ability: stated recall, correct reproduction, and recall certainty. The first follow-up question, which captures stated recall, asked respondents whether they recall their previous answer using "yes" and "no" response categories. Depending on the answer to the first follow-up question, the second follow-up question, meant to capture correct reproduction, asked respondents either to indicate—if respondents said yes—or to estimate—if respondents said no—their previous answer. In line with the scope of this study, some respondents were asked to report their previous answer by selecting the respective category from the actual response scale and some others were asked to report their previous answer by entering the respective category in a text field (without displaying the response scale again). The third follow-up question, which captures recall certainty, asked respondents how confident they are about their reproduction using an eleven-point, end verbalized response scale running from "not at all certain" to "absolutely certain".

The test question was placed at the beginning of the web survey and the follow-up questions were placed towards the end of the web survey to ensure a sufficient time interval. In total, there were 70 to 74 questions between the test question and the follow-up questions dealing with a variety of topics, such as achievement and job motivation and personality

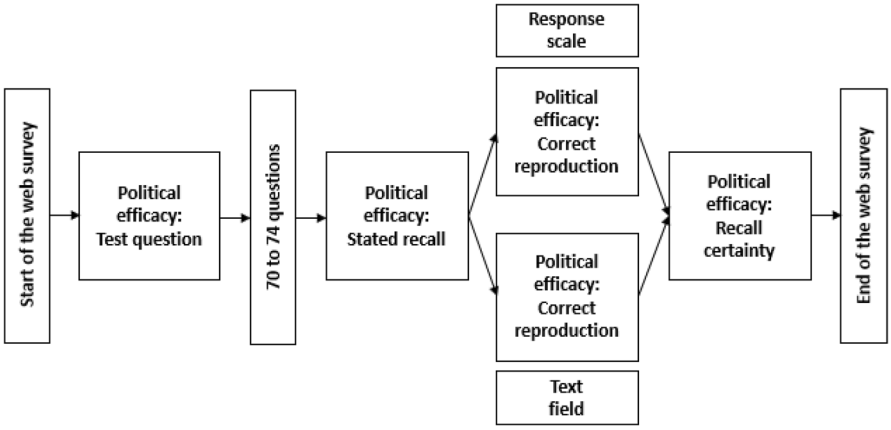

Fig. 1 Course of the web survey and position of the test and follow-up questions. Note. Response scale implies that respondents selected their previous answer from the actual scale. Text field implies that respondents entered their previous answer (the response scale was not displayed again)

Table 2 Experimental design defined by device type and response format

Experimental groups	Device types	Response formats	Group sizes
1	PC	Response scale	823
2	PC	Text field	815
3	Smartphone	Response scale	810
4	Smartphone	Text field	851

traits. Figure 1 illustrates the course of the web survey and Appendix A (see Table 7) shows the wording of the test and follow-up questions.

All questions were designed in German, which was the mother tongue of 96% of the respondents. To improve comparability between PCs and smartphones, I used an optimized survey layout that avoids horizontal scrolling.

3.2 Experimental design

I used a two-step split-ballot design with four experimental groups defined by device type (PC or smartphone) and response format (response scale or text field) for reporting the previous answer to the test question. This resulted in a 2-by-2 factorial design, as displayed in Table 2.

In the first step, respondents were randomly assigned to complete the web survey with a PC or a smartphone. This was done by the survey company before the start of the web survey. Then, in the web survey, respondents were randomly assigned to a response format for reporting their previous answer to the test question within device types. More specifically, respondents were asked to report their previous answer by selecting the respective category from the actual response scale or by entering the respective category in a text field (without displaying the response scale again). Appendix B (See Figs. 2, 3) contains screenshots of the two experimental versions of the follow-up question on correct reproduction for PCs and smartphones.

3.3 Procedure of the study

Data collection was conducted by the survey company Respondi (www.respondi.de) and took place in Germany in July and August 2019. Respondi drew a quota sample from their opt-in access panel based on age and gender, resulting in a 3×2 quota plan designed to represent the German population across these two demographic characteristics. The quotas were calculated based on the German Microcensus, which served as an official population benchmark. In total, Respondi invited 24,246 panelists to take part in the survey. Out of these, 4581 panelists were screened out because the quotas were already achieved or because these respondents tried to access the survey with a different device type than the one they were assigned. In total, 3407 panelists started the web survey. Among these, 108 dropped before being asked the study relevant questions. This leaves 3299 panelists available for statistical analyses.

The email invitation to the web survey included information on the estimated time of the web survey (approx. 20 min), the respective device type to use for survey completion, and a link to the survey. The first page of the survey outlined the general topic and the procedure of the survey and included a statement of confidentiality. Respondents received modest financial compensation from Respondi, which was proportional to the length of the survey.

Only respondents who owned both a PC and a smartphone were invited to take part in this study. To identify these respondents, I used the profiling information provided by the survey company. If respondents tried to enter the survey with a different device type than the one they were assigned, they were blocked from the survey and asked to switch to the correct device type. I also collected user-agent-strings informing about device properties, such as type and operating system.

In addition to user-agent-strings, I collected response times and window and browser tab switching. This was achieved by using the open source tool "Embedded Client Side Paradata" (ECSP) developed by Schlosser and Höhne (2018). Prior informed consent for the collection of paradata was obtained by Respondi as part of the respondents' registration process.

3.4 Sample

In total, 3299 respondents participated in the study (1638 with a PC and 1661 with a smartphone). This corresponds to a participation rate of 14% among all invitees. The respondents were aged 18 to 70 years, with a mean age of 47 years, and 51% of them were male. In terms of education, 13% had completed lower secondary school (low education level), 35% intermediate secondary school (middle education level), and 53% college preparatory secondary school or university-level education (high education level).

3.5 Analytical strategies

To investigate *RQ1a*, I report the proportions of respondents' stated recall regarding their previous answer to the test question (stated recall: 1 = yes). This is done between PC and smartphone respondents.

To investigate *RQ1b* on the correct reproduction, the answers of the respondents who were randomly assigned to enter the previously selected response category in a text field

had to be coded before statistical analyses. Some respondents provided a correct, but misspelled answer (e.g., "it depent" instead of "it depends"). I decided to count such misspellings as correct reproduction because I am interested in respondents' recall ability rather than their writing or typing skills. However, some other respondents entered a (slightly) different response category than the one they previously selected (e.g., "agree somewhat" instead of "agree" or "neither/nor" instead of "it depends"). I decided to count such discrepancies as incorrect reproduction.

With respect to *RQ1b*, I investigate the proportions of respondents' correct reproduction. This measure is determined by comparing respondents' answer to the test question and their answer to the second follow-up question (correct reproduction: 1 = yes). In line with the scope of the study and the experimental design, I now compare the proportions across device types (PC and smartphone) and response formats (response scale and text field).

Regarding *RQ1c* I investigate the means of respondents' recall certainty, which was measured on an eleven-point, end verbalized response scale (recall certainty: 1 "not at all certain" to 11 "absolutely certain"). This is done between the four experimental groups defined by device type and response format.

In order to investigate *RQ2a* on stated recall and *RQ2b* on correct reproduction, I conduct two logistic regression analyses with stated recall and correct reproduction as binary dependent variables (see coding above), respectively. In contrast, to investigate *RQ2c* on recall certainty, I conduct an ordinary least squares (OLS) regression analysis with recall certainty as the dependent variable (see coding above). In line with the scope of this study and the experimental design, I use response scale (1 = yes) and PC (1 = yes) as the independent variables in all three regression analyses. I also control for several variables that previous research suggests to have an effect on stated recall, correct reproduction, and recall certainty (Rettig et al. 2020; Revilla and Höhne 2020; Schwarz et al. 2020; van Meurs and Saris 1990): extreme answer (1 = yes), response time (in seconds), in-between time (in seconds),[3] on-device media multitasking (1 = yes),[4] age (in years), education with high as reference: low (1 = yes) and middle (1 = yes), male (1 = yes), and survey participation (continuous).[5]

In the logistic regression analysis on correct reproduction, I include stated recall as independent variable. Consequently, in the OLS regression analysis on recall certainty, I include stated recall and correct reproduction as independent variables (see coding above).

With respect to *RQ3*, I use the estimation procedure proposed by van Meurs and Saris (1990) to determine memory effects across device types and response formats. More specifically, I subtract the proportions of respondents stating no recall but correctly

[3] To deal with short and long time measures (response time and in-between time, respectively) I conducted an outlier definition procedure following Lenzner et al. (2010): for both time measures the lower and upper one percentile were defined as outliers and excluded from statistical analyses. Additionally, as a robustness check, I conducted all regression analyses without accounting for outliers. For in-between time, the results remain unchanged but for response time, the results changed.

[4] On-device media multitasking is gathered by so-called JavaScript "OnBlur" functions. The OnBlur property is a JavaScript EventHandler that is triggered when an element, document, window, or browser tab loses focus. It allows to gather how often (called "off-count") and how long (called "off-time") respondents switch away from a web survey page. In total, 2.2% of the respondents engaged in on-device media multitasking when answering the test question on political efficacy.

[5] The variable survey participation was provided by the survey company and informs about the number of survey participations during the last 12 months. On average, respondents participated in 68.3 surveys during the last 12 months.

Table 3 Proportions of stated recall between device types

	PC (Groups 1 and 2)	Smartphone (Groups 3 and 4)
Stated recall (%)	89.5	87.4

N = 3218. Groups 1 and 3 received the response scale again for selecting the previous answer. Groups 2 and 4 received a text field for entering the previous answer

Table 4 Proportions of correct reproduction and means of recall certainty across device types and response formats

	PC with response scale (Group 1)	PC with text field (Group 2)	Smartphone with response scale (Group 3)	Smartphone with text field (Group 4)
Correct reproduction (%)	83.2	12.5	82.0	13.8
Recall certainty (means)	9.04	8.12	8.96	7.81

N = 3185 (correct reproduction) and N = 3184 (recall certainty)

reproducing their previous answer from the proportions of respondents stating recall and correctly reproducing their previous answer. This estimation procedure allows me to distinguish between correct reproduction due to attitude stability and correct reproduction due to memory. In addition, it allows me to compare the size of the memory effects obtained in this study with those obtained in earlier studies (Rettig et al. 2020; Revilla and Höhne 2020; Schwarz et al. 2020; van Meurs and Saris 1990).

4 Results

4.1 Stated recall, correct reproduction, and recall certainty

Overall, 88.4% of the respondents stated that they recall their previous answer, which indicates a comparatively high stated recall. As Table 3 reveals, the proportion of PC respondents stating recall is slightly higher than that of smartphone respondents stating recall.

In the next step, I look at respondents' correct reproduction across device types (PC and smartphone) and response formats (response scale and text field). The overall correct reproduction is 48.0%. As Table 4 reveals, correct reproduction is substantially higher for the groups receiving the response scale again than for the groups receiving a text field (without displaying the response scale again). This finding applies to both PC and smartphone respondents. In contrast, the differences between device types within response formats are comparatively small.

Finally, I examine respondents' recall certainty. This is also done across device types and response formats. The overall mean of recall certainty is 8.48. As Table 4 reveals,

Table 5 Logistic regressions of stated recall and correct reproduction and OLS regression of recall certainty (standard errors in parentheses)

Independent variables	Stated recall Coefficients (SE)	Correct reproduction Coefficients (SE)	Recall certainty Coefficients (SE)
Response scale	0.04 (0.12)	3.67*** (0.11)	0.30** (0.11)
PC	0.05 (0.13)	0.05 (0.11)	–0.06 (0.08)
Extreme answer	0.68*** (0.16)	–1.09*** (0.12)	1.25*** (0.09)
Response time	–0.00* (0.00)	–0.00* (0.00)	–0.00 (0.00)
In-between time	0.00 (0.00)	0.00 (0.00)	–0.00 (0.00)
On-device media multitasking	–0.41 (0.45)	1.00* (0.48)	–0.08 (0.35)
Age	0.03*** (0.00)	–0.01* (0.00)	0.01*** (0.00)
Education with high as reference			
Low	–0.71*** (0.17)	–0.27 (0.17)	–0.55*** (0.12)
Middle	–0.14 (0.14)	–0.18 (0.12)	–0.17 (0.09)
Male	0.06 (0.13)	0.03 (0.11)	0.27** (0.08)
Survey participation	0.01*** (0.00)	0.00 (0.00)	0.00** (0.00)
Stated recall	–	0.77*** (0.17)	1.71*** (0.13)
Correct reproduction	–	–	1.14*** (0.11)
Pseudo-R^2	0.05	0.41	–
Adjusted-R^2	–	–	0.22
N	3118	3088	3085

Intercepts are statistically significant. See variable coding in chapter "3.5. Analytical strategies"

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

the results on recall certainty draw a similar picture as the results on correct reproduction. Recall certainty is about one scale point higher for respondents that received the response scale again than for respondents that received a text field for entering the previously selected response category. Again, the differences between PC and smartphone respondents within response formats are very small.

Table 6 Proportions of correct reproduction (depending on stated recall) and estimated memory effects across device types and response formats

	PC with response scale (Group 1)	PC with text field (Group 2)	Smartphone with response scale (Group 3)	Smartphone with text field (Group 4)
Correct reproduction if stated recall: yes (%)	84.6	12.9	83.6	14.0
Correct reproduction if stated recall: no (%)	72.2	8.3	69.9	12.4
Estimated memory effects (%)	12.4	4.6	13.7	1.6

N = 3,185. Estimated memory effects: First row proportions minus second row proportions

4.2 Variables associated with stated recall, correct reproduction, and recall certainty

To answer *RQ2a*, *RQ2b*, and *RQ2c*, I investigated several variables that were claimed by previous research to have an impact on respondents' recall ability (Rettig et al. 2020; Revilla and Höhne 2020; Schwarz et al. 2020; van Meurs and Saris 1990). Table 5 displays the results of the two logistic regressions of stated recall [LR $\chi^2(11) = 108.80$, $p < 0.001$] and correct reproduction [LR $\chi^2(12) = 1767.30$, $p < 0.001$] and the OLS regression of recall certainty [$F(13,3071) = 68.84$, $p < 0.001$].

As Table 5 reveals, response scale significantly increases correct reproduction and recall certainty, which is indicated by the positive coefficients. Respondents reporting their previous answer by selecting the respective response category show a higher correct reproduction and recall certainty than respondents reporting their previous answer by entering the respective response category in a text field. This finding is in line with the previously reported descriptive results. There is no significant effect of response scale on stated recall. PC does not have a significant effect on stated recall, correct reproduction, and recall certainty. Again, this corroborates the previous descriptive results and indicates that the device type does not affect respondents' recall ability in terms of stated recall, correct reproduction, and recall certainty.

Extreme answer has a significant effect in all three regressions. Interestingly, this effect does not consistently tend towards the same direction. While extreme answer has an increasing effect on stated recall and recall certainty, it has a decreasing effect on correct reproduction. This implies that respondents with extreme answers have lower levels of correct reproduction even though they show higher stated recall and recall certainty. I address this point further in the discussion and conclusion section.

Response time has a significant negative effect on stated recall and correct reproduction. However, this effect is very small in both cases and it disappears when not accounting for outliers (using the lower and upper one percentile as thresholds). Thus, this effect is not robust and negligible. Surprisingly, on-device media multitasking, as measured by means of paradata, has an increasing effect on correct reproduction. This implies that respondents engaging in on-device media multitasking (e.g., checking emails or social

media notifications) while answering the test question seem to be more likely to correctly reproduce their answer. Again, I address this point further in the discussion and conclusion section.

I also controlled for several respondent characteristics in all three regressions. Even though the results reveal some significant effects of age, education, and male on stated recall, correct reproduction, and recall certainty, no systematic pattern can be observed. Survey participation has a positive effect on stated recall and recall certainty. However, similar to response time, these effects are negligibly small.

4.3 Estimating memory effects

In order to investigate the extent of memory effects I followed the estimation procedure proposed by van Meurs and Saris (1990): subtracting the proportions of respondents stating no recall but correctly reproducing their previous answer from the proportions of respondents stating recall and correctly reproducing their previous answer. Overall, memory effects are relatively small (7.5%), a finding that is similar to the findings reported by Revilla and Höhne (2020). As Table 6 reveals, the results look somewhat different when comparing estimated memory effects between response formats (response scale and text field). Respondents reporting their previous answer by selecting the respective category from the actual response scale show substantially higher memory effects than respondents reporting their previous answer by entering the respective category in a text field. Memory effects are between 3.7 times (PC respondents) and 8.5 times (smartphone respondents) higher for the group that received the response scale than for the group that received a text field. These results reveal that the response format matters when it comes to estimating memory effects in web surveys with repeated measurements.

5 Discussion and conclusion

The goal of this study was to investigate respondents' recall ability and memory effects within the same survey. For this purpose, I conducted a web survey experiment in an opt-in access panel in Germany with experimental groups defined by device type (PC or smartphone) and response format (response scale or text field) for reporting the previous answer to the test question. In addition, I collected paradata, such as window and browser tab switching, to account for respondents' completion conditions. The results revealed substantial differences with respect to respondents' recall ability and memory effects between response formats. However, they revealed almost no differences between device types used for survey completion.

The results on stated recall (*RQ1a*) showed that the majority of the respondents stated that they recall their previous answer when asked if they do so. This is irrespective of the device type used. However, the results on correct reproduction (*RQ1b*) showed that less respondents correctly reproduced their previous answer. This particularly applies when respondents received a text field instead of the actual response scale. One explanation is that response scales provide a frame of reference (Schwarz et al. 1985), which may help

respondents to come up with the same answer as to the previous question. Text fields do not provide such a frame, impeding correct reproduction. As such, text fields require a greater effort in terms of concentration, consideration, and patience (Bradburn et al. 2004), which, in turn, minimizes the likelihood of correctly reproducing the previous answer simply by guessing. Thus, an overestimation of respondents' recall ability and memory effects is less likely.

The findings on recall certainty (*RQ1c*) corresponded to the findings on correct reproduction. This implies that providing a text field instead of the actual response scale reduces not only correct reproduction but also recall certainty. One explanation might be that, again, compared to response scales, text fields do not provide a frame of reference that respondents can use to infer their previous answer. Thus, these findings indicate that the response format matters when it comes to respondents' recall ability.

The results on the variables associated with stated recall (*RQ2a*), correct reproduction (*RQ2b*), and recall certainty (*RQ2c*) drew a similar picture as the descriptive results. Response scale significantly increased correct reproduction and recall certainty. In contrast, PC neither had an effect on stated recall nor on correct reproduction and recall certainty. As previously reported, extreme answer had a significant effect in all three regressions. In line with the findings reported by Rettig et al. (2020), extreme answer increases stated recall and recall certainty. However, for correct reproduction, the effect of extreme answer flipped and indicated a decreasing effect. Looking at the test question it is to see that it uses "it depends" as middle category, which might not be the most commonly used middle category in agree/disagree response scales. This comparatively unique response scale characteristic may have positively affected correct reproduction of respondents selecting the middle category. It would be worthwhile if future research investigates the impact of the wording of (agree/disagree) response categories on respondents' recall ability and memory effects.

In addition, I found an increasing effect of on-device media multitasking on correct reproduction. This implies that respondents who engage in on-device media multitasking (e.g., checking emails or social media notifications) when answering the test question on political efficacy are more likely to correctly reproduce their previous answer. However, the overall proportion of on-device media multitasking in this study was very low (about 2%) and, thus, this finding should be interpreted with caution. Further research on the relationship between correct reproduction and on-device media multitasking is necessary.

The results on memory effects (*RQ3*) revealed systematic differences between response formats. More specifically, the estimated memory effects were substantially higher for respondents receiving the actual response scale again than for respondents receiving a text field (without displaying the response scale again). As "cooperative communicators" (Schwarz 1996), respondents use the information provided by the response scale to correctly reproduce their previous answer, resulting in an overestimation of memory effects. In particular, this behavior seems to apply to repeated measurements within the same survey in which the in-between time is comparatively short. The results for respondents receiving a text field instead of the actual response scale seem to draw a more precise picture of the size of memory effects. These effects are small.

Similar to previous studies on respondents' recall ability and memory effects (see Rettig et al. 2020; Revilla and Höhne 2020; Schwarz et al. 2020; van Meurs and Saris 1990), in this study, the estimation of memory effects is based on correct reproduction (i.e., selecting

or reporting the identical response category again). However, the requirement of correct reproduction may appear somewhat strict. Another approach would be to go for a kind of "vague reproduction" instead. Following this notion, it would be possible to collapse, for instance, agreeing, middle, and disagreeing response categories. This would be less strict and account for the fact that respondents correctly remember their response tendency.

This study offers some avenues for future research. First, I only used one test question to determine respondents' recall ability and memory effects. In order to provide more robust evidence and to increase the external validity of the results obtained in this study, future research can employ multiple questions that are of different types and cover different topics. Second, I used a test question with a five-point, completely verbalized response scale that was vertically aligned. However, there are numerous scale design aspects that can be varied when investigating respondents' recall ability and memory effects. In relation to this point, it would be interesting to systematically vary the length of the response scales (e.g., three-, four-, five-, six-, and seven-points). Third, in contrast to van Meurs and Saris (1990), I did not systematically vary the questions (e.g., questions with similar and dissimilar topics as the test question) or time (e.g., question number or fixed time periods) between the test question and its repetition; this similarly applies to the other studies on respondents' recall ability and memory effects (Rettig et al. 2020; Revilla and Höhne 2020; Schwarz et al. 2020). This is a crucial aspect that should be addressed in future studies. In relation to this last point, it would also be worthwhile to take the context in which the questions of interest were asked into account. More specifically, future research could take the survey setting (e.g., at home or a coffee house) into account when investigating respondents' recall ability and memory effects.

This study provided new evidence on respondents' recall ability and showed that previous studies seem to overestimate the size of memory effects. Importantly, this similarly applies to web surveys completed on PCs and smartphones. When thinking of web surveys, we need to think of mixed-device surveys because the share of smartphones in web surveys increases continuously (see, for instance, Gummer et al. 2019; Peterson et al. 2017; Revilla et al. 2016). Thus, it is good news that no device effect is at work when it comes to respondents' recall ability and memory effects.

Appendix A

English translations of the test question on political efficacy and the three follow-up questions. The original German wordings of the questions are available from the author on request.

Table 7 Test and follow-up questions for determining respondents' recall ability and memory effects

Test question	Question stems	Response formats and scales
Political efficacy	To what extent do you agree or disagree with the following statement? I am good at understanding and assessing important political issues.	Agree strongly, agree, it depends, disagree, disagree strongly
Follow-up questions		
Stated recall	Earlier we asked you the following question: [TEST QUESTION] Can you recall your exact answer to it?	Yes or no
Correct reproduction if stated recall yes	Please indicate, what your answer to the previous question was: [TEST QUESTION]	Same scale as test question or text field
Correct reproduction if stated recall no	Even if you do not exactly recall. Please try to indicate, what your answer to the previous question was: [TEST QUESTION]	Same scale as test question or text field
Recall certainty	How certain are you about the recall of your answer to the previous question? [TEST QUESTION]	1 not at all certain to 11 absolutely certain

The follow-up question on correct reproduction with a text field included an additional instruction explicitly stating that respondents should enter the previously selected response category

Appendix B

Fig. 2 Screenshots of the follow-up question on correct reproduction for the test question on political efficacy (PC only). Note. Second follow-up question on correct reproduction if respondents stated that they recall their previous answer. The upper screenshot shows the follow-up question providing the response scale again and the lower screenshot shows the follow-up question providing a text field

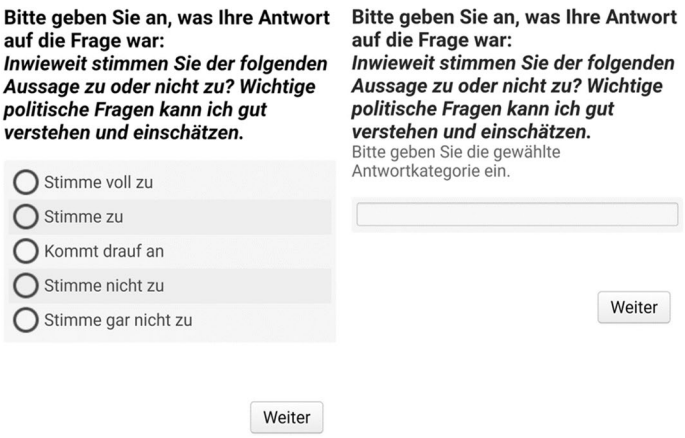

Fig. 3 Screenshots of the follow-up question on correct reproduction for the test question on political efficacy (smartphone only). Note. Second follow-up question on correct reproduction if respondents stated that they recall their previous answer. The upper screenshot shows the follow-up question providing the response scale again and the lower screenshot shows the follow-up question providing a text field

Acknowledgements The author is grateful to Sebastian Lundmark (University of Gothenburg) for his support in conducting this research.

Funding Open Access funding enabled and organized by Projekt DEAL. I acknowledge financial support by the German Research Foundation (Grant Number: 139943784) through the Collaborative Research Center 884 "Political Economy of Reforms" at the University of Mannheim (Germany).

Availability of data and materials Data may be accessed on-site at the Collaborative Research Center 884 "Political Economy of Reforms" at the University of Mannheim (Germany).

Code availability Codes may be accessed by contacting the author (jan.hoehne@uni-due.de).

Declarations

Conflict of interest There are no conflicts of interest or competing interests.

Consent to participate Consent for participation was obtained through the survey company.

Consent for publication I have consent to publish this study.

Ethics approval The study was conducted in line with existing ethical research standards.

References

Alwin, D.F.: How good is survey measurement? Assessing the reliability and validity of survey measures. In: Marsden, P.V., Wright, J. (eds.) Handbook of Survey Research, pp. 405–434. Emerald Group Publishing, London (2010)

Alwin, D.F.: Evaluating the reliability and validity of survey interview data using the MTMM approach. In: Madans, J., Miller, K., Maitland, A., Willis, G. (eds.) Question Evaluation Methods, pp. 265–295. Wiley, Hoboken (2011)

Beierlein, C., Kemper, C. J., Kovaleva, A., Rammstedt, B.: PEKS—Political Efficacy Kurzskala. In: Leibniz-Zentrum für Psychologische Information und Dokumentation (ed.) Elektronisches Testarchiv (PSYNDEX Tests-Nr. 9006492). ZPID, Trier (2012)

Bradburn, N., Sudman, S., Wansink, B.: Asking Questions: The Definitive Guide to Questionnaire Design – For Market Research, Political Polls, and Social and Health Questionnaires. Wiley, San Francisco (2004)

Campbell, D.T., Fiske, D.W.: Convergent and discriminant validation by the multi-trait-multimethod matrix. Psychol. Bull. **56**, 81–105 (1959)

Cohen, J.: Statistical Power Analysis for the Behavioral Science. Academic Press, New York (1969)

Couper, M.P.: New developments in survey data collection. Ann. Rev. Sociol. **43**, 121–145 (2017)

Dimitrov, D.M., Rumrill, P.D.: Pretest-posttest designs and measurement of change. Work **20**, 159–165 (2003)

Gummer, T., Quoß, F., Roßmann, J.: Does increasing mobile device coverage reduce heterogeneity in completing web surveys on smartphones? Soc. Sci. Comput. Rev. **37**, 371–384 (2019)

Höhne, J.K., Schlosser, S.: Investigating the adequacy of response time outlier definitions in computer-based web surveys using paradata SurveyFocus. Soc. Sci. Comput. Rev. **36**, 369–378 (2018)

Höhne, J.K., Schlosser, S., Couper, M.P., Blom, A.G.: Switching away: exploring on-device media multitasking in web surveys. Comput. Hum. Behav. (2020). https://doi.org/10.1016/j.chb.2020.106417

Lenzner, T., Kaczmirek, L., Lenzner, A.: Cognitive burden of survey questions and response times: a psycholinguistic experiment. Appl. Cognit. Psychol. **24**, 1003–1020 (2010)
Peterson, G., Griffin, J., LaFrance, J., Li, J.: Smartphone participation in web surveys. In: Biemer, P.P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L.E., Tucker, N.C., West, B.T. (eds.) Total Survey Error in Practice, pp. 203–233. Wiley, Hoboken (2017)
Revilla, M.A., Höhne, J.K.: Repeatedly measuring political interest: can we reduce respondents' recall ability and memory effects in surveys using memory interference tasks? Int. J. Public Opinion Res. (2020). https://doi.org/10.1093/ijpor/edaa035
Revilla, M.A., Saris, W.E., Krosnick, J.A.: Choosing the number of categories in agree-disagree scales. Sociol. Methods Res. **43**, 73–97 (2014)
Revilla, M., Toninelli, D., Ochoa, C., Loewe, G.: Do online access panels really need to allow and adapt surveys to mobile devices? Internet Res. **26**, 1209–1227 (2016)
Rettig, T., Höhne, J.K., Blom, A.G.: Recalling survey answers: a comparison across question types and different levels of online panel experience. Survey Res. Methods (under review) (2020)
Saris, W.E., Gallhofer, I.N.: Design, Evaluation, and Analysis of Questionnaires for Survey Research. Wiley, Hoboken (2014)
Saris, W.E., Revilla, M.A., Krosnick, J.A., Shaeffer, E.M.: Comparing questions with agree/disagree response options to questions with item-specific response options. Survey Res. Methods **4**, 61–79 (2010)
Saris, W.E., Satorra, A., Coenders, G.: A new approach to evaluating the quality of measurement instruments: the split-ballot MTMM design. Sociol. Methodol. **34**, 311–347 (2004)
Schlosser, S., Höhne, J.K.: ECSP – Embedded Client Side Paradata. Zenodo (2018). https://doi.org/10.5281/zenodo.1218941
Schwarz, N.: Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation. Lawrence Erlbaum Associates, Mahwah (1996)
Schwarz, H., Revilla, M., Weber, W.: Memory effects in repeated survey questions: reviving the empirical investigation of the independent measurement assumption. Survey Res. Methods **14**, 325–344 (2020)
Schwarz, N., Hippler, H.-J., Deutsch, B., Strack, F.: Response scales: effects of category range on reported behavior and comparative judgments. Public Opin. Q. **49**, 388–395 (1985)
Sendelbah, A., Vehovar, V., Slavec, A., Petrovčič, A.: Investigating respondent multitasking in web surveys using paradata. Comput. Hum. Behav. **55**, 777–787 (2016)
Shadish, W.R., Cook, T.D., Campbell, D.T.: Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin Company, Boston (2002)
van Meurs, A., Saris, W. E.: Memory effects in MTMM studies. In: Saris, W. E., van Meurs, A. (eds.) Evaluation of Measurement Instruments by Meta-Analysis of Multitrait Multimethod Studies, pp. 134–146. North Holland, Amsterdam (1990)