



# Differences in measurement quality depending on recall: results for a question about trust in the parliament

Melanie Revilla<sup>1</sup> · Jan Karem Höhne<sup>1,2</sup> · Tobias Rettig<sup>3</sup>

Accepted: 18 May 2022  
© The Author(s) 2022

## Abstract

A large body of studies estimate the measurement quality of survey questions using multitrait-multimethod (MTMM) experiments. In these experiments, respondents are asked (nearly) identical questions at least twice. Most commonly, this is done within the same survey by, for instance, changing the response scale characteristics. However, it has frequently been argued that the estimates from these experiments might be biased when respondents recall their first answer and base their second answer on the initial one. So far, only little is known about the impact of memory effects on the estimates of measurement quality. In this study, we conducted a MTMM experiment in the probability-based German Internet Panel (N=4,401) to investigate whether and to what extent measurement quality estimates differ across respondent groups varying in terms of recall. For this purpose, we use a survey question on trust in the German parliament that was asked with three different response scales. The results reveal that the recall of answers and memory effects vary significantly depending on the scales used for the trust in the parliament question. We also found significant differences in measurement quality across different recall groups. However, more refined research is necessary to better understand the association between answer recall and measurement quality estimates.

**Keywords** Measurement quality · Memory effects · Multitrait-multimethod (MTMM) · Probability-based online panel · Recall of survey answers · Trust in the parliament

---

✉ Melanie Revilla  
melanie.revilla@upf.edu

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

Tobias Rettig  
tobias.rettig@uni-mannheim.de

<sup>1</sup> RECSM-Universitat Pompeu Fabra, Barcelona, Spain

<sup>2</sup> Department of Political Science, University of Duisburg-Essen, Duisburg, Germany

<sup>3</sup> SFB 884 Political Economy of Reforms, University of Mannheim, Mannheim, Germany

## 1 Introduction

Even though it is well-known that survey data suffer from measurement errors, surveys are the predominant method of data collection in social science research. The size of these errors varies depending on many aspects, including the question topic and formulation, country and language, and the characteristics of the response scales (Saris and Gallhofer 2014). Thus, an estimation of the size of measurement errors (under different conditions) before the actual data collection takes place is crucial for improving survey instruments (Revilla et al. 2016). Comparing measurement errors of questions with, for instance, different formulations and/or response scales may help researchers to design questionnaires that minimize measurement errors. In addition, after data collection, the same estimates can be used to correct for remaining measurement errors (Saris and Revilla 2016).

Instead of directly estimating measurement errors, it is common to estimate their complement: measurement quality. Measurement quality is defined as the strength of the relationship between a latent concept of interest and observed survey answers. Measurement quality can be computed as the product of validity and reliability (Saris and Andrews 1991). A common way of estimating measurement quality is through multitrait-multimethod (MTMM) experiments (Althausen et al. 1971; Alwin 1974; Andrews 1984; Campbell and Fiske 1959; Jöreskog 1970, 1971; Saris and Andrews 1991; Werts and Linn 1970). In MTMM experiments, survey questions measuring different latent concepts (called “traits”) are repeated using different methods.

In this study, we focus on MTMM experiments in which the different methods correspond to response scales with different characteristics (e.g., seven vs. eleven response categories). The same respondents are asked to answer the same questions within the same survey several times, systematically varying the response scales. For identification purposes, usually a minimum of three correlated traits are measured; each of them using at least three different methods. This results in a set of at least nine questions that the same respondents are asked to answer.

To reduce the number of repetitions, Saris et al. (2004) proposed to randomly assign respondents to experimental groups that receive a different combination of two instead of three methods. We focus on this split-ballot multitrait-multimethod (SB-MTMM) approach.

A crucial assumption made by MTMM designs is that respondents’ first answer to a particular question does not affect their second answer to it. To put it differently, it is assumed that even if respondents recall the question, they do not recall their initial answer when receiving the question again, but accomplish the information retrieval anew (Tourangeau et al. 2000). However, if respondents use their first answer when processing the second one (e.g., trying to be consistent), then the measures obtained from the repeated questions are not independent of each other, which, in turn, may foster the occurrence of memory effects (van Meurs and Saris 1990). Following Rettig and Blom (2021, p. 3), “the term ‘memory effect’ refers to a specific type of measurement error that occurs when a response is influenced by the respondents’ memory of other responses they have previously given.” Memory effects have the potential to decrease the precision of the parameter estimates that are calculated based on the repeated measures (Saris et al. 2010).

Although the problem of memory effects distorting survey outcomes is frequently mentioned in the literature (see, for instance, Alwin 2007; Moser and Kalton 1972; Rettig and Blom 2021; Saris et al. 2010), only a few studies have empirically investigated memory effects and their consequences for surveys with repeated measurements (Höhne 2021;

Rettig et al. 2019; Revilla and Höhne 2021; Schwarz et al. 2020; van Meurs and Saris 1990). Moreover, to our knowledge, there is no study directly investigating the association between recall of previous answers and measurement quality estimates. In this paper, we start filling this knowledge gap by using data from a SB-MTMM experiment conducted in the probability-based German Internet Panel (GIP) in which we implemented three questions on trust in institutions (i.e., parliament, legal system, and police) that were measured with three different scales: 11-point, item-specific (IS) scale with radio buttons, 7-point, disagree/agree (D/A) scale with radio buttons, and 9-point IS scale that required entering a numeric score in an answer field. The measurement quality estimates are computed for respondents differing in terms of 1) self-reported recall of their first answer to the trust in the parliament question and 2) their ability to correctly reproduce their first answer to this question.

## 2 Literature on recalling answers and memory effects within the same survey

So far, only few studies have empirically investigated the recall of survey answers and the occurrence of memory effects. This particularly applies to repeated measurements within the same survey. One early study was conducted by van Meurs and Saris (1990). The authors used data from the Netherlands Institute for Public Opinion (NIPO) telepanel (Saris 1998). They found that—after a period of about 9 min—between 71 and 85% of the respondents self-reported that they recalled their previous answers (self-reported recall). Moreover, about 70% of the respondents who self-reported that they recalled their previous answers were able to correctly reproduce them when they were asked the same questions again. In contrast, about 36% of the respondents who self-reported that they did not recall their previous answers were able to correctly reproduce them when they were asked the same questions again. Thus, “[it] follows that of the respondents who can reproduce the answers correctly after 9 min approximately (70–36% or) 34% is probably due to memory effects while the others reproduce the result because of the stability of their opinion” (van Meurs and Saris 1990, p. 141). In addition, van Meurs and Saris (1990) found that when the time interval between the two repeated measurements increased and when the questions presented between repetitions dealt with similar topics, the proportion of respondents who correctly reproduced their previous answers decreased. Moreover, respondents giving extreme answers (i.e., selecting the endpoints of the response scales) were more likely to correctly reproduce their previous answers.

About three decades later, Schwarz et al. (2020) conducted a lab experiment in Spain using an online survey (mainly university students). After an average time interval of 20 min, 66% of the respondents self-reported recall and 60% were able to correctly reproduce their previous answer. Using the estimation procedure proposed by van Meurs and Saris (1990), they found that about 17% of the respondents who correctly reproduced their previous answer did so because of memory and not because of chance or opinion stability. Interestingly, the correct reproduction of previous answers did not decrease over time (up to 30 min). In addition, in an attempt to reduce memory effects, the authors included a “memory interference task” between the initial question and its repetition. This task required respondents to evaluate the grammatical correctness of as many sentences as possible within one minute. In contrast to their expectation, the authors found no significant

reduction of correct reproduction of the previous answer when including the memory interference task.

Using data from a survey experiment conducted in the GIP, Rettig et al. (2019) found that respondents self-reported that they recall their previous answers in about 84% of all cases and correctly reproduced them in about 61%. Applying the van Meurs and Saris' (1990) estimation procedure, about 20% of the respondents correctly reproduced their previous answers because of memory. Compared to attitude and belief questions, behavior questions significantly increase respondents' correct reproduction. Even though the authors did not find a significant effect of panel experience on respondents' correct reproduction, they found a significant increasing effect of extreme answers.

Revilla and Höhne (2021) also used data from a survey experiment conducted in the GIP (about seven months later). In contrast to Rettig et al. (2019), the authors tested a question with five (instead of eleven) response categories. They found that about 90% of the respondents self-reported recall and about 88% correctly reproduced their answer. Applying the estimation procedure by van Meurs and Saris (1990), only about 7% of the respondents reproduced their previous answer because of memory. Following Schwarz et al. (2020), Revilla and Höhne (2021) employed memory interference tasks. More specifically, the authors asked respondents to recall numbers or words. In contrast to their expectations, the tasks did not significantly decrease the correct reproduction due to memory.

Höhne (2021) conducted a survey experiment in a nonprobability access panel (Respondi) in Germany and randomly assigned respondents to a device type (PC or smartphone) and response format (response scale or answer field) for reporting the previous answer. About 89% of the respondents self-reported that they recalled their previous answer and about 48% correctly reproduced it. The author found almost no device differences when it comes to correct reproduction. However, correct reproduction was higher when respondents were provided with the response scale again than when they were provided with an answer field without showing the response categories again. Memory effects varied between 2 and 14%, depending on the response format for reporting the previous answer.

Finally, Rettig and Blom (2021) proposed a conceptual framework of how later given answers can be influenced by respondents' memory of their previously given answers. The authors propose an extension of the cognitive response process model by Tourangeau et al. (2000), introducing the influence of memory effects. Specifically, they argued that memory effects can either lead to a dependent response (consistency model) or to a repeated response (satisficing model).

### 3 Research question and contribution

The overview of previous research on recalling survey answers and memory effects within the same survey points to the occurrence of memory effects when repeating survey questions. However, the level of correct reproduction and memory effects largely vary across studies. The reasons for these variations might be related to differences in question topics (e.g., environmental issues or political efficacy), response scale characteristics (e.g., five or eleven response categories), sample types (e.g., nonprobability or probability-based), or country (e.g., Germany or Spain). Further aspects that may affect respondents' correct reproduction and memory effects are the time and question content between repetitions, the extremeness of the answers, and the question type (e.g., attitude or behavior).

Nevertheless, none of the previously discussed studies investigated empirically how memory affects survey outcomes. In particular, they do not investigate whether and to what extent respondents' self-reported recall and correct reproduction are associated with the estimates of measurement quality. However, this is key to evaluate the accuracy of measurement quality estimates. Moreover, in the case of MTMM experiments, memory effects may differ because of employing different response scales in the repeated questions (instead of employing identical scales, as in the previously discussed studies). Therefore, in this study, we address the following research question:

Does measurement quality differ depending on self-reported recall and correct reproduction of previous answers to a survey question on trust in the parliament?

In addition to addressing this main research question, we also report results on self-reported recall, correct reproduction, recall certainty, and memory effects. We test a different question (trust in the parliament) than previous studies and, most importantly, we use three different response scales (see Table 1). Thus, we add new empirical evidence to the scarce literature on this topic.

## 4 Data

### 4.1 Data collection and sample

Data were collected in a probability-based panel, the GIP. Panelists are invited every two months to participate in an online survey that lasts about 25 min. The collected data are available via the GESIS Data Archive for the Social Sciences after a period of six months. For more details, we refer readers to Blom et al. (2015).

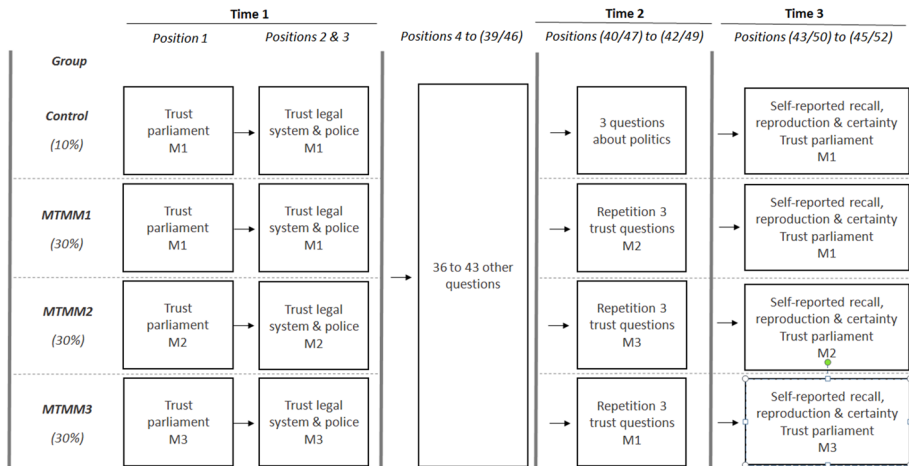
We use data from wave 51 (January 2021; Blom et al., 2021). In total, 4,468 respondents started the wave. Of those, 67 broke-off the survey or did not provide answers to all study-relevant questions. The remaining 4,401 respondents were assigned to one of four experimental groups (see Fig. 1) and are used in the analyses.

The median age of these respondents is 53 years, and 48.4% of them are female. In terms of education, 12.7% graduated from a lower secondary school (low level), 29.4% from an intermediate secondary school (medium level), and 55.6% from a college preparatory secondary school or university (high level). Furthermore, 0.8% were still attending school or left school without a diploma and 1.5% reported having a different degree from those mentioned above. In total, 58.6% completed the online survey with a PC, 6.7% with a tablet, and 34.7% with a smartphone.

To evaluate the effectiveness of the random assignment and the sample composition between the four experimental groups, we conducted chi-square tests. The results showed no significant differences across experimental groups regarding age, gender, education, and device.

### 4.2 Questionnaire

In total, the panel wave included between 58 and 68 questions, depending on the filtering. These questions mainly dealt with respondents' position in the labor market, perceptions



**Fig. 1** Illustration of the course of the online survey including experimental design

of the welfare state, migration, and the corona pandemic. In this study, we focus on 1) the questions of the MTMM experiment and 2) the questions on recall (i.e., self-reported recall, reproduction, and recall certainty) of the answer to the first MTMM question (measuring trait 1; see below).

#### 4.2.1 Questions of the MTMM experiment

The MTMM experiment includes three traits:

- Trust in the German parliament (trait 1)
- Trust in the German legal system (trait 2)
- Trust in the German police (trait 3)

These traits were selected for the following reasons: they (1) are widely used in political science research, (2) are included in previous MTMM experiments implemented in the European Social Survey (ESS; Rounds 1 to 4), and (3) showed less improper solutions than other ESS concepts when estimating their measurement quality (Revilla et al. 2021).

Each trait was measured using three different methods (see Table 1). Since the association between self-reported recall/correct reproduction and measurement quality may vary depending on response scale characteristics, we decided to include quite different scales. This allows us to investigate the robustness of the results across scale types. For method 1 (M1), we used the same question formulations and scales as in the ESS main questionnaire (i.e., 11-point IS scale) but with a vertical alignment to improve the visibility on smartphone screens. For method 2 (M2), we used a disagree/agree (D/A) format with seven response categories (vertically aligned). Although both scales were end labelled and started with the lowermost category (“no trust at all” or “completely disagree”), they differ regarding other key aspects: (1) number of response categories (eleven vs. seven), (2) evaluative scale dimension (IS vs. D/A), (3) scale polarity (unipolar vs. bipolar), and (4) numeric labels (0 to 10 vs. –3 to 3). Although many studies suggest that D/A questions are

**Table 1** Methods included in the MTMM experiment

| Method | No. of scale points | Evaluative dimension | Scale polarity | Answering method               | Scale   |
|--------|---------------------|----------------------|----------------|--------------------------------|---|
| M1     | 11                  | IS                   | Unipolar       | Clicking on radio button       | 0 (No trust at all)<br>1 2 3 4 5 6 7 8 9<br>10 (Complete trust) |
| M2     | 7                   | D/A                  | Bipolar        | Clicking on radio button       | -3 (Completely disagree)<br>-2 -1 0 1 2<br>3 (Completely agree) |
| M3     | 9                   | IS                   | Unipolar       | Entering score in answer field | 0 (No trust at all)<br>1 2 3 4 5 6 7<br>8 (Complete trust)      |

IS = item-specific; D/A = disagree/agree. Appendix 1 provides the English translations of the main questions used in this study (including response scales) and the Online Supplementary Material 1 provides exemplary screenshots of the MTMM questions

of lower quality than IS questions (see, for instance, Höhne et al. 2018, 2017; Höhne and Krebs 2018; Saris et al. 2010), they are frequently used in applied survey research (Revilla 2017). Furthermore, there are good reasons to assume that answers to D/A questions are not equivalent to answers to IS questions aiming to measure the same concepts. Even if respondents recall their first answer, the effect on measurement quality might be less pronounced when the evaluative scale dimension changes in the repetition. Finally, for method 3 (M3), we used the same request for an answer as for M1, but with a nine-point IS scale requiring respondents to enter the respective score in an answer field instead of selecting a response category. This may lead to different levels of memory effects.

#### 4.2.2 Questions on self-reported recall, correct reproduction, and recall certainty

We adopted the questions proposed by van Meurs and Saris (1990). First, we asked respondents whether they recalled their answer (yes/no) to the first MTMM question (trust in the parliament) using the respective methods received at time 1. We then asked respondents to either indicate (if they self-reported recall) or estimate (if they did not self-report recall) what their answer was. By comparing respondents' answer with their initial answer to the question on trust in the parliament, we determine respondents' correct reproduction. Finally, we asked respondents how certain they were about the recall of their initial answer (0 "not at all certain" to 10 "absolutely certain"). English translations of the original questions are available in Appendix 1.

## 5 Method and analyses

### 5.1 Full experimental design

The SB-MTMM experiment uses a three-group design, as proposed by Saris et al. (2004): respondents are asked to answer three questions using one method at time 1 and then the same questions using a different method at time 2. We used a three-group SB-MTMM

design because it usually performs better than a two-group design (Revilla and Saris 2013). Respondents who were randomly assigned to the first MTMM group (MTMM1) answered M1 at time 1 (the first questions of the survey that were placed after the introduction page) and M2 at time 2 (after 36 to 43 in-between questions and followed by a total of 16 questions on evaluating the survey and last-minute additions due to the corona pandemic). Respondents who were randomly assigned to the MTMM2 group answered M2 at time 1 and M3 at time 2, and respondents in the MTMM3 group answered M3 at time 1 and M1 at time 2.

To investigate whether answering the repeated questions at time 2 affects respondents' answers to the questions on self-reported recall, reproduction, and recall certainty that are asked just after them (time 3), we included also a control group. This control group was of smaller size (10%) than the MTMM1 to MTMM3 groups (30%, respectively).

To have a similar number of in-between questions in the control group as in the other ones, we added three questions on politics instead of three questions with a second method at time 2. We used the following three questions, which are asked just before the questions on trust in institutions in the ESS questionnaire (Round 4): political interest (B1), understanding politics (B2), and forming political opinions (B3). Figure 1 illustrates the course of the survey including the experimental design.

## 5.2 Analyses

### 5.2.1 Preliminary analyses

First, we conducted preliminary analyses on self-reported recall, correct reproduction, recall certainty, and memory effects, following the analytical strategy used in previous studies (Höhne 2021; Rettig et al. 2019; Revilla and Höhne 2021; Schwarz et al. 2020; van Meurs and Saris 1990). Since the trust in the parliament question was asked with different scales across the experimental groups (Control and MTMM1 employed M1, MTMM2 employed M2, and MTMM3 employed M3), we report, for each experimental group, the proportion of respondents self-reporting recall of their previous answer, correctly reproducing their previous answer, and their average recall certainty.

In addition, we compute the proportions of correct reproduction for respondents self-reporting recall and for those self-reporting no recall. The difference between both proportions is used as a memory effect estimate (van Meurs and Saris 1990).

We tested for significant differences between the control and MTMM1 groups to evaluate whether and to what extent receiving the repeated questions at time 2 affects self-reported recall, correct reproduction, recall certainty, and memory effects. We also tested for significant differences between MTMM1 and MTMM2, MTMM2 and MTMM3, and MTMM1 and MTMM3 to investigate differences across scale types. We use two-sample Z-tests when comparing proportions and two-sample T-tests with equal variances when comparing means. The preliminary analyses were conducted in Stata 14 (StataCorp 2015).

### 5.2.2 Measurement quality analyses

In the analyses on measurement quality, we excluded respondents from the control group, because they only received one method at time 1. Thus, we cannot compute measurement quality for this group. For the other three groups, we report reliability, validity, and



**Table 2** Description of the 12 groups used in the measurement quality analyses including sample sizes (N = 3,945)

| Group | Name    | Self-reported recall | Correct reproduction | Experimental group | Sample size (n) |
|-------|---------|----------------------|----------------------|--------------------|-----------------|
| 1     | R-C-1   | Yes                  | Yes                  | MTMM1              | 645             |
| 2     | R-C-2   | Yes                  | Yes                  | MTMM2              | 735             |
| 3     | R-C-3   | Yes                  | Yes                  | MTMM3              | 346             |
| 4     | R-NC-1  | Yes                  | No                   | MTMM1              | 378             |
| 5     | R-NC-2  | Yes                  | No                   | MTMM2              | 237             |
| 6     | R-NC-3  | Yes                  | No                   | MTMM3              | 484             |
| 7     | NR-C-1  | No                   | Yes                  | MTMM1              | 143             |
| 8     | NR-C-2  | No                   | Yes                  | MTMM2              | 207             |
| 9     | NR-C-3  | No                   | Yes                  | MTMM3              | 179             |
| 10    | NR-NC-1 | No                   | No                   | MTMM1              | 156             |
| 11    | NR-NC-2 | No                   | No                   | MTMM2              | 141             |
| 12    | NR-NC-3 | No                   | No                   | MTMM3              | 294             |

R = self-reported recall; C = correct reproduction; NR = no self-reported recall; NC = no correct reproduction

measurement quality estimates for the trust in the parliament question. In addition, we divide respondents in several groups depending on their recall.

**5.2.2.1 Groups for the measurement quality analyses** To investigate the association between self-reported recall, correct reproduction, and measurement quality, we consider four groups of respondents: (1) those who self-reported that they recalled their first answer and correctly reproduced it, (2) those who self-reported that they recalled their first answer but did not correctly reproduce it, (3) those who self-reported that they did not recall their first answer but correctly reproduced it, and (4) those who self-reported that they did not recall their first answer and did not correctly reproduce it. However, since we use a three-group SB-MTMM design, we have 12 groups when conducting the measurement quality analyses. Table 2 shows these groups including sample sizes. Appendix 2 provides information about the main socio-demographic variables of the respondents in these 12 groups.

**5.2.2.2 Estimation of measurement quality** Different models have been proposed to analyze MTMM experiments. We use the True Score model proposed by Saris and Andrews (1991). This model is used in many recent publications including MTMM analyses and has the advantage of allowing to separately estimate reliability and validity. Thus, we can see whether recall affects them in different ways. The following equations summarize the model:

$$Y_{ij} = r_{ij}T_{ij} + e_{ij} \quad (1)$$

$$T_{ij} = v_{ij}F_i + m_{ij}M_j \quad (2)$$

where  $F_i$  is the  $i$ th trait or factor,  $M_j$  is the  $j$ th method,  $Y_{ij}$  is the observed answer for the  $i$ th trait and the  $j$ th method,  $T_{ij}$  is the true score or systematic component of the answer,  $r_{ij}$  is the reliability coefficient (when completely standardized),  $v_{ij}$  is the validity coefficient

(when completely standardized),  $m_{ij}$  is the effect of the method and  $e_{ij}$  is the random error. The total measurement quality is calculated by multiplying the reliability (square of the reliability coefficient) by the validity (square of the validity coefficient):

$$q_{ij}^2 = r_{ij}^2 * v_{ij}^2$$

In the base model, we make the following assumptions: (1) the random errors are uncorrelated with each other and random errors are uncorrelated with the independent variables in the different equations. (2) The traits are correlated. (3) There is no correlation between method factors and no correlation between method and trait factors. (4) The impact of the method factor (i.e., systematic errors) is the same for the traits measured with a common scale.

The estimates of reliability and validity coefficients are obtained from LISREL 8.72 (Jöreskog and Sörbom 2004) analyzing the covariance matrices using Maximum Likelihood estimation in a multiple group context (the 12 groups presented in Table 2). Initially, we set all parameters invariant across the 12 groups. Then, we test this base model and introduce corrections stepwise.

**5.2.2.3 Testing of the model** To determine which corrections should be applied to the base model, we use a combination of theoretical considerations and indicators of the goodness-of-fit of the model. In a first step, we defined two main types of theoretically justified corrections: (1) due to timing (respondents may get tired or learn; see, for instance, Batista-Foguet et al. 2014) and possible memory effects, we expect that the level of random and systematic errors differ between time 1 and time 2. (2) We also expect differences across the four groups of respondents that were created based on self-reported recall and correct reproduction. More specifically, only respondents with recall should be prone to memory effects. Moreover, these groups are not created randomly so that they may differ regarding several aspects affecting the associations between the variables of interest (e.g., the correlations between traits may differ across the groups).

Next, we test for misspecifications using the JRULE software (van der Veld et al. 2008) that implements the testing procedure developed by Saris et al. (2009). This procedure uses the Expected Parameter Changes (EPC), Modification Indices (MI), and power to determine whether each parameter is or is not misspecified.<sup>1</sup> To decide on the introduction of the first correction, we combine the theoretical considerations listed above with the information on the size of the misspecifications provided by JRULE. The corrections are introduced stepwise until a well-fitting model is obtained (see Online Supplementary Material 2). Then, we use the reliability and validity coefficients of this final model to compute the measurement quality estimates. The syntax of the LISREL input for the final model is available in the Online Supplementary Material 3.

<sup>1</sup> We use the default values of the program (0.10 for causal effects and correlations and 0.40 for loadings).

**Table 3** Self-reported recall, correct reproduction, and recall certainty of the trust in parliament question (N=4,401). \* $p < .05$ , \*\* $p < .01$ , Exp. Group=experimental group. Correct rep.=correct reproduction. SE=standard errors. NS=non-significant ( $p \geq .05$ ). M1 is a 11-point IS radio button scale, M2 is a 7-point D/A radio-button scale, and M3 is a 9-point IS scale where respondents have to type in a score

| Exp. group                  | Control    | MTMM1      | MTMM2      | MTMM3      | Significance differences |     |     |     |
|-----------------------------|------------|------------|------------|------------|--------------------------|-----|-----|-----|
|                             |            |            |            |            | Control-1                | 1-2 | 1-3 | 2-3 |
| Method                      | M1         | M1         | M2         | M3         |                          |     |     |     |
| % Self-reported recall (SE) | 81.5 (.02) | 77.4 (.01) | 73.5 (.01) | 63.7 (.01) | NS                       | *   | **  | **  |
| % Correct rep. (SE)         | 63.7 (.02) | 59.6 (.01) | 71.3 (.01) | 40.3 (.01) | NS                       | **  | **  | **  |
| Mean recall certainty (SE)  | 7.4 (.11)  | 7.4 (.06)  | 7.2 (.06)  | 7.2 (.07)  | NS                       | *   | *   | NS  |
| <i>n</i>                    | 444        | 1,323      | 1,325      | 1,309      |                          |     |     |     |

**Table 4** Correct reproduction of the first answer to the trust in parliament question of respondents self-reporting recall, self-reporting no recall, and estimated memory effects (N=4,401). \* $p < .05$ , \*\* $p < .01$ , Exp. Group=experimental group. Correct rep.=correct reproduction. SE=standard errors. NS=non-significant ( $p \geq .05$ ). M1 is a 11-point IS radio button scale, M2 is a 7-point D/A radio-button scale, and M3 is a 9-point IS scale where respondents have to type in a score

| Exp. group                                    | Control    | MTMM1      | MTMM2      | MTMM3      | Significance differences |     |     |     |
|---|------------|------------|------------|------------|--------------------------|-----|-----|-----|
|   |            |            |            |            | Control-1                | 1-2 | 1-3 | 2-3 |
| Method  | M1         | M1         | M2         | M3         |                          |     |     |     |
| % Correct rep. self-reported recall: yes (SE) | 66.6 (.02) | 63.1 (.01) | 75.6 (.01) | 41.6 (.02) | NS                       | **  | **  | **  |
| % Correct rep. self-reported recall: no (SE)  | 51.2 (.06) | 47.8 (.03) | 59.5 (.03) | 37.9 (.02) | NS                       | **  | **  | **  |
| % Estimated memory effects                    | 15.4       | 15.3       | 16.1       | 3.7        | NS                       | NS  | **  | **  |
| <i>n</i>                                      | 444        | 1,323      | 1,325      | 1,309      |                          |     |     |     |

## 6 Results

### 6.1 Results of the preliminary analyses

First, we report the proportions of self-reported recall and correct reproduction and the means of recall certainty. Table 3 provides the results for the four experimental groups. The proportions of self-reported recall vary between 63.7% (MTMM3) and 81.5% (Control). This is in line with the self-reported recall levels observed in previous studies. Correct reproduction varies between 40.3% (MTMM3) and 71.3% (MTMM2). The average recall certainty varies between 7.2 (MTMM2 and MTMM3) and 7.4 (Control and MTMM1). Recall certainty was measured on a scale from 0 to 10.

There is no significant difference between the Control and MTMM1 groups. Thus, the fact that respondents get a repetition of the trust in parliament question (with a different scale) in the MTMM1 group does not seem to affect recall. However, we find significant differences when comparing the three MTMM groups (except for recall certainty between MTMM2 and MTMM3), suggesting that the scales used affect self-reported recall, correct reproduction, and recall certainty. Asking respondents to answer by typing in a score (M3) results in the lowest self-reported recall, correct reproduction, and recall certainty.

To estimate the proportion of correct reproduction due to memory, we follow the procedure proposed by van Meurs and Saris (1990). Table 4 reports the proportions of respondents who correctly reproduced their answer (when self-reporting recall and when self-reporting no recall) and the estimated memory effects (difference between the first and second rows).

We estimate that between 3.7% (MTMM3) and 16.1% (MTMM2) of the respondents correctly reproduced their previous answer due to memory. This is lower than what was found by van Meurs and Saris (1990), but it is in line with what other studies with similar in-between times found (see, for instance, Höhne 2021; Revilla and Höhne 2021). Differences for the three variables are significant across all MTMM groups, except for one comparison (memory effects between MTMM1 and MTMM2). This suggests that the response scale used makes a difference. The lowest memory effect is found for M3 (IS scale with entering a score), while the highest one is found for M2 (D/A scale with radio buttons). Again, no significant differences are found between the control and MTMM1 groups.

## 6.2 Measurement quality estimates

First, the average measurement quality over all estimates is 0.82. This implies that 82% of the variance in the observed survey answers comes from the latent concept of interest. The remaining 18% are due to measurement errors. Even though this finding indicates that measurement errors exist, this can be considered as “good” quality.<sup>2</sup> For instance, Poses et al. (2021) report an average measurement quality of 0.65 for 67 ESS questions.

Second, the average measurement quality across all groups, methods, and time points is very similar for the three traits included in the MTMM experiment. It varies between 0.81 for the trust in the parliament question and 0.84 for the trust in the police question.

Since we only used the first trait (trust in the parliament) to investigate recall, Table 5 reports the reliability, validity, and measurement quality for this trait at different time points (time 1 and 2), for different methods (M1, M2, and M3), and for four recall groups (R-C, R-NC, NR-C, and NR-NC).

### 6.2.1 Measurement quality

We find significant differences in measurement quality across the four groups. This similarly applies to time 1 and time 2. For instance, at time 1 for M2, the difference in measurement quality is 0.28 between the R-C (0.85) and R-NC (0.57) groups. There are two main possible reasons for this finding.

First, the recall groups are not created randomly and thus respondents in each group may have different characteristics (e.g., the level of attention that they have paid to the survey question). This circumstance may have led to different levels of measurement quality. Indeed, respondents of the different groups analyzed in the MTMM experiment differ in terms of socio-demographic characteristics (see Appendix 2) and correlations between the

<sup>2</sup> Following DeCastellarnau and Revilla (2017), we use similar thresholds as the ones proposed for Cronbach’s alpha (Bland and Altman 1997) to interpret the estimates of measurement quality.

**Table 5** Estimates of reliability, validity, and measurement quality of the trust in the parliament question for different time points, methods, and recall groups (N=3,945). R = self-reported recall; C=correct reproduction; NR=no self-reported recall; NC=no correct reproduction. M1 is a 11-point IS radio button scale, M2 is a 7-point D/A radio-button scale, and M3 is a 9-point IS scale where respondents have to type in a score

| Time | Methods | Reliability ( $r^2$ ) |      |      |      |       |      | Validity ( $v^2$ ) |      |      |      |       |      | Quality ( $q^2$ ) |      |      |      |       |      |
|------|---------|-----------------------|------|------|------|-------|------|--------------------|------|------|------|-------|------|-------------------|------|------|------|-------|------|
|      |         | R-C                   |      | R-NC |      | NR-NC |      | R-C                |      | R-NC |      | NR-NC |      | R-C               |      | R-NC |      | NR-NC |      |
|      |         | R-C                   | R-NC | R-C  | R-NC | R-C   | R-NC | R-C                | R-NC | R-C  | R-NC | R-C   | R-NC | R-C               | R-NC | R-C  | R-NC | R-C   | R-NC |
| 1    | M1      | .94                   | .92  | .94  | .94  | .86   | .96  | .92                | .92  | 1    | .88  | .90   | .85  | .94               | .76  |      |      |       |      |
| 1    | M2      | .92                   | .64  | .90  | .83  | .83   | .92  | .88                | .92  | .92  | .92  | .85   | .57  | .83               | .76  |      |      |       |      |
| 1    | M3      | .92                   | .94  | .83  | .94  | .94   | .96  | 1                  | .96  | .96  | .96  | .89   | .94  | .80               | .90  |      |      |       |      |
| 2    | M1      | .94                   | .94  | .94  | .86  | .86   | .96  | .81                | .96  | .96  | .88  | .90   | .76  | .90               | .76  |      |      |       |      |
| 2    | M2      | .92                   | .64  | .90  | .83  | .83   | .92  | .88                | .92  | .92  | .92  | .85   | .57  | .83               | .76  |      |      |       |      |
| 2    | M3      | .94                   | .94  | .94  | .77  | .77   | .96  | .96                | .86  | .88  | .90  | .90   | .90  | .81               | .68  |      |      |       |      |

trust in the parliament question and the other two trust questions at time 1 (see Appendix 3).<sup>3</sup>

Second, the presence of memory effects can also play a role in the observed differences in measurement quality. Indeed, if respondents in the recall groups use their first answer to decide about their second one, an overestimation of measurement quality in these groups is to be expected. Based on our preliminary results on recall, we would expect a higher overestimation for M2 (highest estimated memory effects; see Table 4) and M1 (second highest estimated memory effects) than for M3 (lowest estimated memory effects).

However, there is no clear pattern indicating that the recall groups show a higher overall measurement quality than the others. For instance, at time 1 for M1, the highest measurement quality is found for one of the groups with no self-reported recall (NR-C). Moreover, for M2, one of the groups with self-reported recall (R-NC) shows the lowest measurement quality. This similarly applies to time 1 and time 2. Overall, there is no systematic trend and no strong evidence of an overestimation of measurement quality due to memory effects.

In addition, even if it can be assumed that respondents get tired and/or bored during the course of the survey, which, in turn, may increase measurement errors and decrease measurement quality at time 2 (see, for instance, Batista-Foguet et al. 2014), most of the measurement quality estimates do not substantially change between time 1 and time 2. One explanation might be that the survey was only about 25 min and/or that we use data from a well-established panel (with experienced respondents). In the R-C group (the group in which we mainly expect memory effects), all three quality estimates are very similar. There are only two substantial quality estimate changes between time 1 and time 2. One for the R-NC group (M1) and one for the NR-NC group (M3). However, these changes do not seem to be related to recall since one occurs in a group self-reporting recall and the other one in a group self-reporting no recall.

## 6.2.2 Reliability and validity

Considering reliability and validity estimates separately, it is to observe that there is more variation across reliabilities (varying between 0.64 and 0.94) than across validities (varying between 0.81 and 1). In addition, the validity estimates are higher than the reliability estimates (an average of 0.93 vs. 0.88). This suggests that random errors might be more problematic than systematic errors. When comparing the groups based on recall, at different time points and for different methods, similar conclusions can be obtained as for measurement quality. In particular, there is no clear pattern across the recall groups in terms of reliability and validity. Depending on the methods and time points, sometimes the groups with recall and sometimes the groups without recall show the highest estimates of reliability and/or validity.

<sup>3</sup> These correlations cannot differ because of memory effects since they are observed at time 1. Thus, correlation differences support the idea that the groups differ in key aspects for our analyses.

## 7 Discussion

The main goal of this study was to investigate whether and to what extent measurement quality estimates differ depending on respondents' self-reported recall and correct reproduction of their answer to a question on trust in the parliament. For this purpose, we conducted a three-group SB-MTMM experiment including three traits on trust in institutions. Each trait was asked using three different scales. Moreover, we used similar questions as van Meurs and Saris (1990) to determine respondents' recall ability. Data collection was conducted in the German Internet Panel (GIP).

The preliminary analyses reveal that memory effects occur. This applies to all three methods (or response scales). The level of self-reported recall, correct reproduction, and memory effects varies across scales. Using an IS question with a nine-point scale requiring respondents to enter their score in an answer field (M3) leads to the lowest level of self-reported recall, correct reproduction, and memory effects. In contrast, using a D/A question with seven response categories (M2) leads to the highest level of correct reproduction and memory effects. Self-reported recall and recall certainty are highest for an IS question with eleven response categories (M1). Finally, comparing the results of the control and MTMM1 groups also suggests that answering at time 2 (using a different scale than at time 1) does not affect self-reported recall, correct reproduction, recall certainty, and memory effects. One limitation of these analyses is the way in which memory effects were estimated. Since some respondents self-report recalling their answers but then do not reproduce it, it is difficult to determine which answers are actually influenced by respondents' memory of previously given answers. We tried to tackle this problem by following the estimation procedure proposed by van Meurs and Saris (1990), which is also used by other recent publications (see, for instance, Höhne 2021; Rettig et al. 2019; Revilla and Höhne 2021; Schwarz et al. 2020). Nevertheless, errors cannot be precluded.

With respect to our main research question (*Does measurement quality differ depending on self-reported recall and correct reproduction of previous answers to a survey question on trust in the parliament?*), our results show significant differences in measurement quality depending on self-reported recall and correct reproduction (both at time 1 and time 2). Reliability and validity estimates vary as well. However, this finding cannot only be attributed to the recall of answers to previous questions. It might be associated with the non-random nature of our groups of comparison (i.e., R-C, R-NC, NR-C, and NR-NC). The groups are based on respondents' self-reported recall and correct reproduction and thus they may differ with respect to key characteristics (e.g., attention that is paid to the survey question). We have supporting evidence that they differ in terms of socio-demographic characteristics and correlations between the trust questions at time 1. This circumstance may have affected measurement quality estimates. Thus, the observed differences may be a combination of the non-random nature and memory effects.

Further research is necessary to better disentangle the effect of memory from other differences between the four groups of interest on measurement quality. In addition, the differences in measurement quality do not follow a systematic trend. In contrast to the frequently mentioned apprehension (see, for instance, Saris et al. 2010), we did not find strong evidence for an overestimation of measurement quality estimates at time 2. One possible reason for this finding is that we used different response scales at time 2. Thus, when answering at time 2, respondents cannot simply repeat their answer from time 1 but need to use their answer from time 1 and transform it to decide which response category to select on the new scale. However, this might not be a simple task for respondents. Particularly, if

the response scales are very different (e.g., 11-point IS scale instead of 7-point D/A scale). Thus, the memory effects are expected to be lower when using different scales at time 2 than when asking to reproduce the answer using the same scale.

Another limitation is that we only used one question dealing with trust in the parliament. To test the robustness of our results, similar analyses should be conducted using different questions. For instance, future studies could use question topics that vary in terms of saliency. Robustness checks for other methodological aspects, such as the response scale characteristics, the type of sample or panel, the data collection mode, the country, or the time interval and content of the questions, would be worthwhile.

In addition, even though we defined a clear testing procedure, there is some unavoidable subjectivity with respect to the decisions on the model specifications and the introduced corrections, as in all Structural Equation Model (SEM) analyses. These decisions may have affected the results reported in this study. In our opinion, further research on how to make the testing procedure in the field of SEM more objective would be desirable.

Overall, our results suggest that researchers should carefully select the response scales for studies with measurement repetitions because they matter when it comes to recalling answers to previously asked survey questions. However, we did not find clear evidence for an overestimation of measurement quality estimates due to memory effects. One reason might be that the response scales used at time 1 and time 2 were sufficiently different. Further research with more similar scales would be useful. So far, it seems that there is no strong reason to be concerned about a bias in measurement quality estimates due to memory; at least if similar conditions as in this study are applied (e.g., similar in-between time and similar sample type).

## Appendix 1

See Table 6.



**Table 6** English translations of the main questions used in this study (including response scales)

|   | Question stems  | Response scales and formats  |
|---|---|--|
| Trust in parliament question (M1)           | How much trust do you have in the German parliament?  | 0 "no trust at all"<br>1 2 3 4 5 6 7 8 9<br>10 "complete trust"        |
| Trust in parliament question (M2)           | How much do you agree or disagree with the following statement?<br>I have trust in the German parliament  | -3 "completely disagree"<br>-2 -1 0 1 2<br>3 "completely agree"        |
| Trust in parliament question (M3)           | How much trust do you have in the German parliament?<br>Please rate your trust on a scale from 0 to 8, where 0 means "no trust at all" and 8 means "complete trust". You can grade your response with the numbers in-between (1 to 7) | [field for numeric input]  |
| Self-reported recall                        | Earlier we asked you the following question: [Trust in parliament question]<br>Can you recall your exact answer to it?  | Yes/no   |
| Reproduction (if self-reported recall: yes) | Please indicate, what your answer was   | [same scale as trust in parliament question]                           |
| Reproduction (if self-reported recall: no)  | Even if you do not exactly recall:<br>Please estimate, what your answer was   | [same scale as trust in parliament question]                           |
| Recall certainty                            | How certain are you about your answer?  | 0 "Not at all certain"<br>1 2 3 4 5 6 7 8 9<br>10 "Absolutely certain" |

## Appendix 2

See Table 7.

**Table 7** Information about the main socio-demographic characteristics of respondents in the 12 recall groups

|                   | R-C-1 | R-C-2 | R-C-3 | R-NC-1 | R-NC-2 | R-NC-3 | NR-C-1 | NR-C-2 | NR-C-3 | NR-NC-1 | NR-NC-2 | NR-NC-3 |
|-------------------|-------|-------|-------|--------|--------|--------|--------|--------|--------|---------|---------|---------|
| Female (%)        | 46.1  | 42.9  | 45.8  | 48.1   | 49.8   | 62.2   | 55.0   | 43.9   | 60.3   | 53.5    | 53.4    |         |
| Time of birth (%) |       |       |       |        |        |        |        |        |        |         |         |         |
| < 1955            | 18.7  | 19.9  | 19.3  | 22.2   | 25.7   | 18.1   | 12.6   | 17.7   | 21.3   | 18.1    | 17.6    | 23.5    |
| 1955–1964         | 25.7  | 22.6  | 20.7  | 21.7   | 17.7   | 22.0   | 21.0   | 21.5   | 30.9   | 20.0    | 22.5    | 21.4    |
| 1965–1974         | 19.7  | 22.3  | 26.8  | 19.0   | 24.0   | 20.8   | 18.2   | 19.1   | 18.5   | 21.3    | 22.5    | 19.4    |
| 1975–1984         | 15.9  | 17.3  | 16.7  | 16.4   | 14.8   | 16.3   | 21.0   | 12.4   | 14.6   | 17.4    | 12.7    | 15.3    |
| 1985+             | 20.0  | 18.0  | 16.4  | 20.6   | 17.7   | 22.8   | 27.3   | 29.2   | 14.6   | 23.2    | 24.6    | 20.4    |
| Education (%)     |       |       |       |        |        |        |        |        |        |         |         |         |
| Low               | 11.8  | 13.0  | 12.7  | 13.0   | 16.8   | 12.5   | 14.0   | 14.3   | 18.9   | 17.3    | 12.7    | 11.2    |
| Medium            | 26.5  | 28.3  | 28.8  | 32.0   | 30.2   | 28.5   | 30.1   | 30.1   | 26.1   | 32.0    | 22.5    | 39.3    |
| High              | 60.2  | 57.1  | 57.3  | 52.6   | 50.4   | 57.5   | 53.8   | 53.1   | 53.9   | 49.4    | 64.8    | 49.1    |

### Appendix 3

See Table 8.

**Table 8** Correlations between the trust in parliament question and the other two trust questions (at time 1) across the 12 groups used in the MTMM analyses

|                             | R-C-1 | R-C-2 | R-C-3 | R-NC-1 | R-NC-2 | R-NC-3 | NR-C-1 | NR-C-2 | NR-C-3 | NR-NC-1 | NR-NC-2 | NR-NC-3 |
|-----------------------------|-------|-------|-------|--------|--------|--------|--------|--------|--------|---------|---------|---------|
| Parliament and legal system | .80   | .77   | .91   | .66    | .58    | .52    | .65    | .56    | .75    | .55     | .47     | .43     |
| Parliament and police       | .64   | .63   | .69   | .49    | .40    | .47    | .50    | .46    | .32    | .45     | .30     | .28     |

Groups finishing by "1" (respectively, "2" or "3") answered the trust questions at time 1 using M1 (respectively, M2 or M3)

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11135-022-01441-6>.

**Acknowledgements** The authors are grateful to Willem Saris for his support while doing this research.

**Funding** Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This article uses data from wave 51 of the German Internet Panel (Blom et al. 2021). A study description can be found in Blom et al. (2015). The German Internet Panel is part of the Collaborative Research Center 884 (SFB 884) funded by the German Research Foundation (DFG)—Project Number 139943784—SFB 884.

## Declarations

**Conflict of interest** The authors have not disclosed any competing interests.

**Ethical approval** The results are not included in any other paper and the paper is not under submission in any other journal. The results have only been presented by the first author in a webinar organized by the Department of Statistics of the University of Stockholm (in April 2021).

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Althausser, R.P., Heberlein, T.A., Scott, R.A.: A causal assessment of validity: the augmented multitrait-multimethod matrix. In: Blalock, H.M. (ed.) *Causal Models in the Social Sciences*, pp. 151–169. Aldine-Atherton, Chicago, New York (1971)
- Alwin, D.F.: An analytic comparison of four approaches to the interpretation of relationships in the multitrait-multimethod matrix. In: Costner, H.L. (ed.) *Sociological Methodology*, pp. 79–105. Jossey-Bass, San Francisco (1974)
- Alwin, D.F.: *Margins of Error: A Study of Reliability in Survey Measurement*. Wiley, Hoboken (2007)
- Andrews, F.M.: Construct validity and error components of survey measures: a structural modelling approach. *Public Opin. q.* **48**(2), 409–442 (1984). <https://doi.org/10.1086/268840>
- Batista-Foguet, J.M., Revilla, M., Saris, W.E., Boyatzis, R., Serlavós, R.: Reassessing the effect of survey characteristics on common method bias in emotional and social intelligence competencies assessment. *Struct. Equ. Model.* **21**(4), 596–607 (2014). <https://doi.org/10.1080/10705511.2014.934767>
- Bland, J.M., Altman, D.G.: Statistical notes: Cronbach's Alpha. *BMJ* **314**, 572 (1997). <https://doi.org/10.1136/bmj.314.7080.572>
- Blom, A.G., Gathmann, C., Krieger, U.: Setting up an online panel representative of the general population: the German internet panel. *Field Methods* **27**(4), 391–408 (2015). <https://doi.org/10.1177/1525822X15574494>
- Blom, A.G., Gonzalez Ocano, M., Fikel, M., Krieger, U., Rettig, T., SFB 884 'Political Economy of Reforms', Universität Mannheim: German Internet Panel, Wave 51.: GESIS Data Archive, Cologne. ZA7759 Data file Version 2.0.0 (2021). <https://doi.org/10.4232/1.13833>
- Campbell, D.T., Fiske, D.W.: Convergent and discriminant validation by the multitrait-multimethod matrices. *Psychol. Bull.* **56**(2), 81–105 (1959)
- DeCastellarnau, A., Revilla, M.: Two approaches to evaluate measurement quality in online surveys: an application using the Norwegian Citizen Panel. *Surv. Res. Methods* **11**(4), 415–433 (2017). <https://doi.org/10.18148/srm/2017.v11i4.7226>
- Höhne, J.K.: New insights on respondents' recall ability and memory effects when repeatedly measuring political efficacy. *Qual. Quant.* (2021). <https://doi.org/10.1007/s11135-021-01219-2>

- Höhne, J.K., Krebs, D.: Scale direction effects in agree/disagree and item-specific questions: a comparison of question formats. *Int. J. Soc. Res. Methodol.* **21**, 91–103 (2018). <https://doi.org/10.1080/13645579.2017.1325566>
- Höhne, J.K., Schlosser, S., Krebs, D.: Investigating cognitive effort and response quality of question formats in web surveys using paradata. *Field Methods* **29**(4), 365–382 (2017)
- Höhne, J.K., Revilla, M., Lenzner, T.: Comparing the performance of agree/disagree and item-specific questions across PCs and smartphones. *Methodology* **14**(3), 109–118 (2018). <https://doi.org/10.1027/1614-2241/a000151>
- Jöreskog, K.G.: A general method for analysis of covariance structures. *Biometrika* **57**(2), 239–251 (1970). <https://doi.org/10.1093/biomet/57.2.239>
- Jöreskog, K.G.: Statistical analysis of sets of congeneric tests. *Psychometrika* **36**(2), 109–133 (1971). <https://doi.org/10.1007/BF02291393>
- Jöreskog, K.G., Sörbom, D.: LISREL for Windows (Version 8.72). Computer software. Scientific Software International, Skokie (2004)
- Moser, C.A., Kalton, G.: *Survey Methods in Social Investigation*. Basic Books, New York (1972)
- Poses, C., Revilla, M., Asensio, M., Schwarz, H., Weber, W.: Measurement quality of 67 common social sciences questions across countries/languages: results of 28 multitrait-multimethod experiments implemented in the European Social Survey. *Surv. Res. Methods* **15**(3), 235–256 (2021). <https://doi.org/10.18148/srm/2021.v15i3.7816>
- Rettig, T., Blom, A.G.: Memory effects as a source of bias in repeated survey measurement. In: Cernat, A., Sakshaug, J.W. (eds.) *Measurement Error in Longitudinal Data*, pp. 3–18. Oxford University Press, Oxford (2021)
- Rettig, T., Höhne, J.K., Blom, A.G.: Recalling Survey Answers: A Comparison Across Question Types and Different Levels of Online Panel Experience. ESRA Conference, July 15–19, Zagreb, Croatia (2019)
- Revilla, M.: Analyzing the survey characteristics, participation, and evaluation across 186 surveys in an online opt-in panel in Spain. *Methods Data Anal.* **11**(2), 135–162 (2017). <https://doi.org/10.12758/mda.2017.02>
- Revilla, M., Höhne, J.K.: Repeatedly measuring political interest: can we reduce respondent' recall ability and memory effects in surveys using memory interference tasks? *Int. J. Public Opin. Res.* **33**(3), 678–689 (2021). <https://doi.org/10.1093/ijpor/edaa035>
- Revilla, M., Saris, W.E.: The split-ballot multitrait-multimethod approach: implementation and problems. *Struct. Equ. Model.* **20**(1), 27–46 (2013). <https://doi.org/10.1080/10705511.2013.742379>
- Revilla, M., Zavala-Rojas, D., Saris, W.E.: Creating a good question: how to use cumulative experience. In: Wolf, C., Joye, D., Smith, T.W., Fu, Y. (eds.) *The SAGE Handbook of Survey Methodology*, pp. 236–254. SAGE, London (2016)
- Revilla, M., Poses, C., Serra, O., Asensio, M., Schwarz, H., Weber, W.: Applying the estimation using pooled data approach to the multitrait-multimethod experiments of the European Social Survey (rounds 1 to 7). *Struct. Equ. Model.* **28**(3), 463–474 (2021). <https://doi.org/10.1080/10705511.2020.1807988>
- Saris, W.E.: Ten years of interviewing without interviewers: the Telepanel. In: Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C., Martin, J., Nicholls, W.L., O'Reilly, J.M. (eds.) *Computer-Assisted Survey Information Collection*, pp. 409–431. Wiley, New York (1998)
- Saris, W.E., Andrews, F.M.: Evaluation of measurement instruments using a structural modeling approach. In: Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., Sudman, S. (eds.) *Measurement Errors in Surveys*, pp. 575–598. Wiley, Hoboken (1991)
- Saris, W.E., Gallhofer, I.N.: *Design, Evaluation, and Analysis of Questionnaires for Survey Research*, 2nd edn. Wiley, Hoboken (2014)
- Saris, W.E., Revilla, M.: Correction for measurement errors in survey research: necessary and possible. *Soc. Indic. Res.* **127**(3), 1005–1020 (2016). <https://doi.org/10.1007/s11205-015-1002-x>
- 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**(1), 311–347 (2004). <https://doi.org/10.1111/j.0081-1750.2004.00155.x>
- Saris, W.E., Satorra, A., Van der Veld, W.M.: Testing structural equation models or detection of misspecifications? *Struct. Equ. Model.* **16**(4), 561–582 (2009). <https://doi.org/10.1080/10705510903203433>
- Saris, W.E., Revilla, M., Krosnick, J.A., Shaeffer, E.M.: Comparing questions with agree/disagree response options to questions with construct-specific response options. *Surv. Res. Methods* **4**(1), 61–79 (2010). <https://doi.org/10.18148/srm/2010.v4i1.2682>
- Schwarz, H., Revilla, M., Weber, W.: Memory effects in repeated survey questions—reviving the empirical investigation of the independent measurements assumption. *Surv Res Methods* **14**(3), 325–344 (2020). <https://doi.org/10.18148/srm/2020.v14i3.7579>

- StataCorp.: Stata Statistical Software: Release 14. Computer Software. StataCorp LP, College Station (2015)
- Tourangeau, R., Rips, L.J., Rasinski, K.: *The Psychology of Survey Response*. Cambridge University Press, Cambridge (2000)
- 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)
- Van der Veld, W.M., Saris, W.E., Satorra, A.: *Judgement Rule Aid for Structural Equation Models (Version 3.0.4 beta)*. Computer software. Radboud University, Nijmegen (2008)
- Werts, C.E., Linn, R.L.: Path analysis: psychological examples. *Psychol. Bull.* **74**(3), 193–212 (1970). <https://doi.org/10.1037/h0029778>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.