Survey Research Methods (2020)
Vol. 14, No. 3, pp. 325-344
doi:10.18148/srm/2020.v14i3.7579

ISSN 1864-3361
http://www.surveymethods.org

Memory Effects in Repeated Survey Questions Reviving the Empirical Investigation of the Independent Measurements Assumption

Hannah Schwarz
RECSM
Universitat Pompeu Fabra
Barcelona, Spain

Melanie Revilla
RECSM
Universitat Pompeu Fabra
Barcelona, Spain

Wiebke Weber
RECSM
Universitat Pompeu Fabra
Barcelona, Spain

It is common to repeat survey questions in the social sciences, for example to estimate test-retest reliability or in pretest-posttest experimental designs. An underlying assumption is that the repetition of questions leads to independent measurements. Critics point to respondents' memory as a source of bias for the resulting estimates. Yet there is little empirical evidence showing how large memory effects are within the same survey and none showing whether memory effects can be decreased through purposeful intervention during a survey. We aim to address both of these points based on data from a lab-based web survey containing an experiment. We repeated one of the initial questions at the end of the survey (on average 127 questions later) and asked respondents if they recall their previous answer and to reproduce it. Furthermore, we compared respondents' memory of previously given responses between two experimental groups: A control group, where regular survey questions were asked in between repetitions and a treatment group which, additionally, received a memory interference task aimed at decreasing memory. We found that, after an average 20-minute interval, 60% of the respondents were able to correctly reproduce their previous answer, of which we estimated 17% to do so due to memory. We did not observe a decrease in memory as time intervals between repetitions become longer. This indicates a serious challenge to using repeated questions within the same survey. Moreover, the tested memory interference task did not reduce respondents' recall of their previously given answer or the memory effect.

Keywords: Memory effects, survey answer recall, repeated survey questions, web survey experiment, memory interference task.

1 Introduction

Repeated survey questions are employed frequently for different purposes. One important application is the estimation of measurement error. In the process of answering survey questions, both random and systematic errors tend to occur which are commonly referred to as measurement errors (Saris, 1990). Where the size of such measurement errors can be estimated, it is possible to correct for them (Saris & Revilla, 2016). In order to estimate the size of measurement errors, it is necessary to measure the same concept for the same respondent repeatedly. One assumption is that this leads to independent measurements of the same concept. This assumption is crucial for several approaches, in psychological, sociological and political science research, such as test-retest, Quasi-Simplex (Heise, 1969; Wiley & Wiley, 1970), and Multitrait-Multimethod (MTMM) approaches (Andrews, 1984). Repeated survey questions are also employed in experimental research using pretest-posttest designs where the same indicator is measured before and after the treatment and experimental effects are assessed on the basis of comparing these repeated measures (Dimitrov & Rumrill, 2003). Yet another context in which repeated survey questions are frequently used are panel surveys which ask respondents the same questions repeatedly across

Contact information: Hannah Schwarz, Research and Expertise Centre for Survey Methodology (RECSM), Universitat Pompeu Fabra, Edificio Mercè Rodoreda, Despatx 24.406 Ramón Trías Fargas 25–27, 08005 Barcelona (email: hannah.schwarz@upf.edu).

waves in order to assess within person changes over time. However, if asked the same question two or more times, respondents' memory of being asked this question previously may affect their answer in line with Tourangeau and Rasinski's (1988) description of context effects in which prior interview items and respondents' answers to them affect answers to later questions. Similarly, the literature on panel conditioning has stressed how respondents' knowledge of questionnaire structures collected in prior waves can affect their answers in later waves Bach and Eckman (2018). In the present article, we focus on recall of previous answers and memory effects within the course of one interview or one survey (in the case of self-administered questionnaires).

Indeed, the assumption of independent measurements is particularly questionable within the course of one survey as only little time separates the repetitions, making it more likely that respondents remember having been asked the same question previously and remember their previous answer (Alwin, 2007). This issue appears especially in MTMM experiments (Alwin, 2007), or for pretest-posttest designs, for example as used for checking if a treatment in an experiment had an impact (see e.g., Bekker, Fischer, Tobi, & van Trijp, 2017; Hindo & González-Prendes, 2011; Hogg & Grieve, 1999; Momm, Blickle, & Liu, 2013). That respondents are likely to remember their previous answer is all the more worrisome as it is commonly assumed that respondents seek to provide consistent answers (Alwin, 2007), which is backed up by empirical findings in social psychology (Cialdini, Trost, & Newsom, 1995). If respondents remember their previous answer and strive to be consistent with it, repetitions of the questions would not be independent measures anymore. This would result for instance in an overestimation of the test-retest correlation leading to the introduction of bias in the estimates (Heise, 1969; Lord & Novick, 2008) In the literature, this problem has been described as memory effects (Lord & Novick, 2008, Chapter 6; Alwin, 2007; Van Meurs & Saris, 1990). In the psychological literature on cognition, a distinction is made between two linked types of memory. Recognition memory, commonly seen as a combination of two components: a familiarity judgement and a component of conscious recollection, as well as recall memory, commonly seen as depending exclusively on recollective processes (Baddeley, 2013; Kopelman et al., 2007; Manns, Hopkins, Reed, Kitchener, & Squire, 2003). We are interested only in recall memory and we shall, in the following, use the terms "recall', "remember" and "memory" to refer to this type of memory.

There is a lack of empirical evidence assessing whether and how strongly recall of previous answers affects responses to repeated questions (Marx, Menezes, Horovitz, Jones, & Warren, 2003) and whether and how recall of previous answers can be decreased. Van Meurs and Saris (1990) empirically assessed the size of memory effects under different conditions as well as how much time is needed in between question repetitions to avoid memory effects. However, societal and technological changes in the past three decades make a re-assessment of their findings relevant. One factor that could affect changes in the role of memory since the Van Meurs and Saris (1990) study is the switch to web as predominant mode of survey data collection (Callegaro, Manfreda, & Vehovar, 2015). Linked to this is the development that respondents are increasingly trained in taking surveys, especially in web mode, as ever more surveys are being conducted these days (Callegaro et al., 2015). This increased familiarity with the task of answering survey questions could lead to differences in how well respondents recall previously given answers. Beyond this, the recommended survey length for web mode tends to be between 10 (ideal) and 20 (maximum) minutes (Revilla, Ochoa, & Toninelli, 2016). This implies that if repeated measurements are collected within one survey, they can be expected to be collected rather closer together in time in the future. This makes a re-assessment of the strength of memory effects after short periods of time especially urgent. Furthermore, the increasing use of mobile devices to answer web surveys (Revilla, Toninelli, Ochoa, & Loewe, 2016) could also affect memory effects since some studies found that mobile devices are used more often outside of the home (Mavletova & Couper, 2013), in the presence of third parties (Toninelli & Revilla, 2016b), and are associated with different kinds of multitasking (Toninelli & Revilla, 2016a). Thus, we expect mobile respondents to be more distracted, which could reduce their recall of previous answers.

Our overall goal is therefore to revive the empirical investigation of memory effects. More specifically, our aim is twofold. Firstly, we aim to replicate some aspects of Van Meurs and Saris' (1990) study, namely investigating to what extent respondents recall previous answers within the same survey as well as estimating the resulting memory effect. In order to do this, we conducted a lab-based web survey, in which one of the initial questions was presented again at the end of the questionnaire (after answering on average 127 questions). We used a similar design as van Meurs and Saris in which respondents were asked whether they remembered their previous answer and then were requested to reproduce it or provide their best guess about it. We investigated to what extent respondents recall previous answers within the same survey and estimated the resulting memory effect following the strategy of van Meurs and Saris. Secondly, we investigated the possibility of reducing respondents' recall of previously given answers by means of an experiment testing the effectiveness of a memory interference task. If memory effects exist, it is not sufficient to be aware of them but something needs to be done to reduce the bias they can create. Thus, we start investigating a specific way to potentially reduce memory effects: presenting a task to respondents that

aims at interfering with their memory of previous answers.

2 Background

2.1 Why memory is likely to bias repeated measurements

If respondents remember their previous answer, it seems plausible that this may affect the answer they will give to a repeated question. Social psychology research furthermore supports the idea that this influence could be systematic as it shows that respondents tend to strive for providing consistent answers. There is a large body of literature in social psychology about individuals' general aim to be consistent and to appear consistent to others (for reviews see, for example, Gawronski, 2012; Guadagno & Cialdini, 2010). The assertion of people's preference for consistency over inconsistency in order to avoid dissonance has become "almost axiomatic" in the field (Brown, Asher, & Cialdini, 2005, p. 517). Though social psychologists Cialdini et al. (1995) put this overarching aim for consistency into perspective by showing that people differ in their levels of striving for consistency, they still concluded that about half of their studies' subjects showed a strong inherent preference for consistency. Differences in seeking consistency have been measured on a "preference for consistency" scale. The scores on this scale have also shown to be positively associated with individuals' test-retest reliability scores on questionnaires (Guadagno & Cialdini, 2010).

Such a systematic influence of initial answers on answers to repeated questions, in statistical terms, has been described as correlated errors of the two measures. This has led Alwin (2007, p. 103) to express doubt about repeated measurements being used within the course of one interview. He states that they could only be used "if one can rule out memory as a factor in the covariance of measures over time, and thus, the occasions of measurement must be separated by sufficient periods of time [. . .]".

2.2 Prior research on memory effects

Psychometricians and Sociometricians have alerted to the fact that if measurement is repeated so close in time that respondents may recall their previous answers, this prevents the two measurements from being independent (Heise, 1969; Nunnally, 1978). Various scholars have stated that repeating a measurement in the course of one survey bears a high risk of memory effects and should thus be avoided (Alwin, 2007; Marx et al., 2003). Especially scholars dealing with MTMM experiments have repeatedly pointed to memory as a potential limitation to their approach (Alwin, 2007; Krosnick, 2011; Saris, Satorra, & Coenders, 2004). Besides prolonging the time interval in between repetitions, another approach to deal with memory effects has been to statistically correct for them. Approaches to this have been described for meta-analyses of MTMM experiments (Scherpenzeel & Saris, 1997) as well as for test-retest reliability studies (Laenen et al. n.d.). Furthermore, Saris et al. (2004) proposed to reduce this problem at least somewhat by using a split ballot MTMM design which requires asking the same question to respondents only two instead of three times in the course of one survey. Moreover, Saris (2013) argues that, in the specific case of MTMM experiments, the problem posed by memory is somewhat reduced because respondents are not presented with an exact repetition of a question but a variation of it (different scale).

While there is thus widespread awareness of the problem, scholars have bemoaned the absence of empirical evidence that could guide decisions about how much time in between repeated questions is adequate (Marx et al., 2003). Scholars in the field of psychometrics working with test-retest correlations as reliability indicators have conducted empirical studies with the aim of determining a suitable time interval. On the one hand, time intervals need to be long enough to avoid all kinds of carry-over effects. These can be, for example, caused by memory of the previous answer affecting the answer on the repeated question (memory effects). They can also be due to respondents scoring differently on tasks in the repeated measure because of the practice they gained with the task during the initial measurement (practice effects). On the other hand, time intervals need to be short enough to make actual changes in the measured attributes unlikely.

Few empirical studies have been dedicated to the bias of test-retest reliability caused by memory effects. McKelvie (1992) manipulated retest conditions experimentally by giving respondents different instructions concerning to what extent they should use their memory of their previous answers (three weeks prior) in working out their answers to the repetition of cognitive tasks in the questionnaire. He concluded that recall of previous answers does not seriously inflate test-retest reliability estimates. McConnell, Strand, and Valdés (1998) looked at opinion surveys on leisure time activities and found that carry-over effects are not a problem when asking the same question again after two months. On the contrary, Salinsky, Storzbach, Dodrill, and Binder (2001) found that after a 12 to 16-week interval practice effects were commonly affecting test-retest based reliability scores of neuropsychological questionnaire instruments.

Marx et al. (2003) focused on the time interval between the repetitions and assessed whether it makes a difference to measure test-retest reliability with a time interval of two days or two weeks. They ran an experiment where health status questionnaires were either repeated after two days or after two weeks. Comparing test-retest reliability scores between the two intervals, they did not find significant differences and concluded that it is adequate to use any time interval between two days and two weeks.

However, empirical investigations of how memory effects

play out within the same survey, which is our focus, have been extremely scarce. There is one seminal study by Van Meurs and Saris (1990) investigating respondents' ability to remember previous answers as well as factors affecting their memory. They ran an experiment in the Telepanel of the Netherlands Institute for Public Opinion (NIPO) to compare respondents' memory of answers to different survey questions after an average nine-minute interval and after two weeks. The NIPO telepanel was the first computer-assisted self-interview (CASI) panel based on a nationally representative sample of the Dutch population (Saris, 1998). The total sample size used for this study was 1,537. The authors looked at six items, among them three asking about the attractiveness of three main Dutch political parties (the labour party "PvdA", the Christian democrat party "CDA" and the liberal party 'VVD') and three concerning attitudes towards public services (the public health care service, the public post delivery service and the public transportation service). Ten-point scales were used for all questions. They found that, for those six items, the proportion of respondents who reported to remember their answer after an average nine-minute interval ranged between 71% and 85% and the proportion of respondents who could reproduce their previous answer ranged between 57% and 72%. Furthermore, this number decreased with increasing time intervals between the repetitions (see Figure 1). The authors observed an average decrease of correct reproduction of previous answers across time when looking at all six different items. However, looking at the items separately shows that the decrease was only present for four of six items, namely the first four items concerning political parties and healthcare, but not for the last two items concerning mail delivery and public transport (see Figure 1). The differences in time intervals in their study were a result of respondents going through slightly different versions of the questionnaire due to routing and respondents' different response speeds. Our study design only allows replicating the aspects discussed above and only concerning memory within the same survey. Van Meurs and Saris (1990) also investigated further aspects, which we do not aim to replicate in the present study: They looked at the effects of being asked questions of similar content to the initial question in between repetitions (all questions dealt with politics) versus of different content (questions in between repetitions dealt with issues other than politics). They found that in situations where similar questions were asked in between repetitions, memory effects disappeared after 24 minutes (according to their Table 7). In their NIPO-telepanel study, this corresponded to approximately 70 questions. If questions about other topics were asked in between repetitions, memory effects only disappeared after 80 minutes (Van Meurs & Saris, 1990). Their predictions of these time intervals are based only on respondents who did not give extreme answers to the initial questions. The authors found that respondents who had given extreme answers to the initial question were more likely to remember their answers.

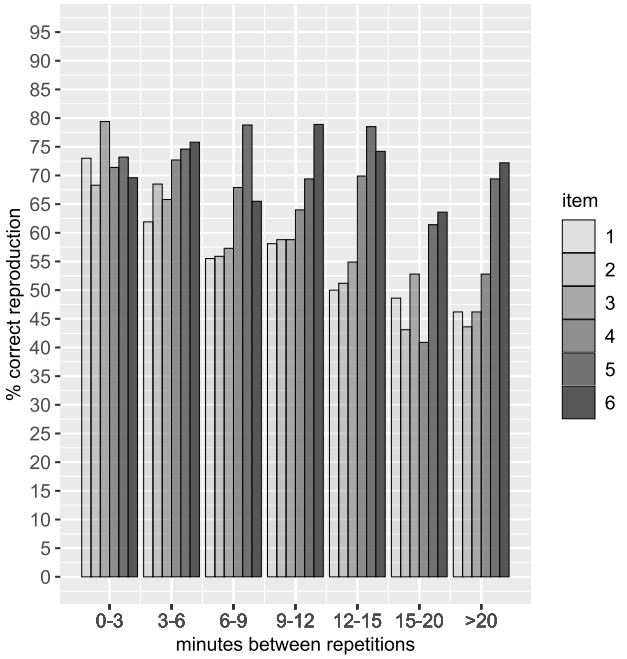

Figure 1. Proportion providing correct reproduction of previous answers across time (in minutes) (Note: Adaptation of Figure 2 from van Meurs and Saris (1990) based on numbers provided by authors in the article in Table 6 (page 142).

Another contribution to the debate is Alwin's (2011) assessment of the potential role of memory effects by looking at results of word recall tests conducted in a nationally representative survey. He observed that after ten minutes of intermediate questions, respondents were rather good at remembering words they had previously learned and recalled. Alwin suggests that this contradicts the assessment of Van Meurs and Saris (1990) that, at the end of an approximately 20-minute questionnaire, a repeated measurement should be free of memory effects (Alwin, 2011). However, while Alwin's contribution gives interesting indications about respondents' ability to recall information they learned earlier in the interview, it remains unclear to what extent this is relevant to the question whether respondents remember their previous survey answers.

2.3 Can recall be decreased? Forgetting through retroactive interference

Next to assessing the presence and strength of memory effects, the second aim of this paper is to investigate whether survey designers could decrease recall of previous answers by amending surveys in a way that would make it harder for respondents to remember their answers. Classic studies in the cognitive psychology field of memory come to the con-

clusion that there are two processes of forgetting – one being a process in which memory fades or decays away with time and the other being a process in which memory gets disrupted or obscured by subsequent learning, or, more broadly, by events occurring between the "learning" of information and the attempt to recall it (Baddeley, 2013; Van Dyke & Johns, 2012). Studies have shown that the amount of forgetting is least when individuals merely rest between the original learning and the recall, that it increases when individuals engage in an unrelated task in between, and that it is highest when individuals engage in a task dealing with similar content in between (Crouse, 1971; Gordon, Hendrick, & Johnson, 2001; Kumar, 2000; McGeoch & McDonald, 1931; Van Dyke & Johns, 2012). The fact that forgetting can be enhanced by increasing cognitive workload as well as by dealing with similar content between learning and recall is described as retroactive interference of memory (Baddeley, 2013; Henderson, 2005). Classical tasks used to induce retroactive memory interference in experiments include asking respondents to memorize certain contents in between the initial learning and the recall request or giving them additional materials to study in between (Baddeley, 2013; Crouse, 1971). However, asking respondents to memorize content or to go through readings allows the researcher little control over whether and to what extent respondents actually engage in these tasks, especially in self-administered survey modes. We thus devised a task aimed at increasing cognitive workload as well as presenting respondents with additional similar content in between repetitions while also allowing us control over the task implementation. Lexical decision tasks which require fast decision making seem to increase cognitive workload as compared to the task of answering standard survey questions. Furthermore, lexical decision tasks offer the advantage that they can be implemented with any type of semantical content.

Therefore, they are also commonly used as a way of presenting semantic content crucial to the experimental treatment to respondents (see, for example, Vitale, Kosson, Resch, & Newman, 2018). We therefore devised a task akin to lexical decision tasks, which required respondents to make quick decisions on the grammatical correctness of sentences they were presented with. They were asked to judge as many sentences as possible within a fixed one-minute time interval. This type of memory interference task offers an additional benefit when implemented in substantive research. While similar content could also be presented by means of additional survey questions about the same topic (see Van Meurs & Saris, 1990), a task asking to judge grammatical correctness of sentences seems less prone to cause conditioning effects which contaminate the repeated measurement. Conditioning effects refer to the effects on responses resulting from previous data collection from the same respondents (United States Federal Committee on Statistical Methodology, 1987). A detailed description of the task can be found in section 3.2. Our expectation is that this task, as it should increase cognitive workload and expose respondents to additional similar content in between repetitions decreases respondents' ability to remember their previous answers. This results in the following hypotheses:

Hypothesis 1. Respondents in the experimental group receiving the memory interference task are less likely to report that they remember their previous answer than those in the control group.

Hypothesis 2. Respondents in the experimental group receiving the memory interference task are less likely to correctly reproduce their previous answer than those in the control group.

Hypothesis 3. Among respondents in the experimental group receiving the memory interference task the memory effect is smaller.

3 Data and methods

3.1 The survey

The aim of our web survey was to assess memory of previously given answers. In order to ensure that we use standard social survey questions, we develop our questionnaire on the basis of the Spanish European Social Survey (ESS) questionnaires (Rounds 6-8, see Table A1 in the appendix). Additionally, following Van Meurs and Saris (1990), one of the initial questions was presented again at the end of the questionnaire followed by two questions assessing respondents' recall of their previous answer to this question. The question was the Spanish translation of 'How difficult or easy do you find it to deal with important problems that come up in your life?'. The answer scale ranged from 0 "Extremely difficult" to 10 "Extremely easy". This question was chosen because of its topic and its answer scale. As we expected our sample to consist to a large part of social science students, we chose a question that did not deal with a political or societal topic as these might have been especially salient for social science students and therefore particularly easy to remember. Furthermore, we consider the 11-point answer scale of this question to be a typical scale used in surveys and therefore relevant to study. We did not use one of the questions used by Van Meurs and Saris (1990), as the exact formulations were not provided in the original study, as the topics of the questions did not seem appropriate for the context of our survey (different time and place) and as we prioritized the above described reasons for choosing a question. At the very end of the survey (on average 127 questions later), respondents were presented the question a second time and were asked whether they could recall the exact answer they had previously given. If they responded yes, they were subsequently

presented the question again with the request to give the same answer as previously. If they responded no, the request was to try to approximate their previous answer (see screenshots in Appendix B). The complete questionnaire contained 134 items, ten of which were only presented to some respondents depending on answers to previous filter questions. The survey was programmed in Qualtrics (2020).

3.2 The survey experiment

The experimental set-up of our lab-based web survey experiment aimed at comparing survey answer recall in a condition where respondents are asked questions about broadly similar topics in broadly similar formats between the two repetitions (approximating the standard situation during a social survey) with a condition where survey designers intentionally include an additional task aimed at decreasing memory between the two repetitions right after the question for which respondents' memory would be measured later. We will refer to this question as test question in the following.

The experiment consists of one treatment and one control group. The test question was first presented as the third question of the survey. The treatment group was presented with the memory interference task directly afterwards. This task was programmed to last exactly one minute and was inspired by lexical decision tasks. Lexical decision tasks ask respondents to decide, as fast as possible, whether a series of letters they are presented with forms a real word or is just a nonsensical string of letters (Meyer & Schvaneveldt, 1971; Ratcliff, Gomez, & McKoon, 2004; Vitale et al., 2018). Even though mostly used for measuring the ease of processing lexical information (Wagenmakers, Ratcliff, Gomez, & McKoon, 2008), such tasks have useful features for our purposes. On the one hand, they allowed us control over the actual task implementation by the respondent. On the other hand, we could choose the content presented to be similar to the repetition question, which is how retroactive interference is commonly induced (Van Dyke & Johns, 2012) and which according to Van Meurs and Saris' (1990) findings reduces memory effects. Moreover, these tasks require quick decision making and should thereby induce higher cognitive workload in respondents than would be induced by just letting them answer standard survey questions. This is another common way of inducing retroactive interference. Hence by both allowing us to present respondents with similar content in between repetitions and to increase cognitive workload, the format of lexical decision tasks seems to suit our purpose of inducing retroactive interference with respondents' memory of the survey answer in question.

However, in order to be able to present content that is maximally similar to the content of the test question, our task differed slightly from lexical decision tasks. Because we would like to interfere with respondents' memory of their survey answers, we created a decision task where respondents are asked to judge sentences rather than single words. The implication was that we asked respondents to judge not the criterion of whether a string of characters formed an actual word but whether a sequence of words formed a grammatically correct sentence. Half of the sentences presented were grammatically correct and half were not. In the introduction to the task, respondents were informed that they would now receive a one-minute task judging sentences in terms of grammatical correctness. They were asked to work fast and judge as many sentences as possible within the one-minute period. By providing a larger number of sentences than could realistically be judged by respondents within the one-minute interval, we ensured that nobody could finish the task before being forwarded to the next questionnaire section after exactly one minute. A timer was displayed to respondents counting down the 60 second they had available for the task. The nature of this interference task allowed us to use sentences containing fragments of the test question and corresponding answer statements as memory interference content. For a detailed depiction of the memory interference task, see Appendix C and tables A2 and A3 in Appendix A. The control group received no such task but instead continued answering the standard series of survey questions after answering the test question for the first time.

3.3 Data collection

Data collection took place on September 25 and October 1, 2018 at University Pompeu Fabra in Spain. Participants took the survey on desktop computers in a computer lab where the experimenter was present at all times. It was thus made sure that participants took the survey without external disruptions and it is unlikely that multitasking or interruption of the survey taking occurred as the experimenter did not notice any such behaviour. Participants were compensated with 5€ upon completing the survey. Initially, 122 observations were collected, but seven of them had to be excluded from the analysis due to technical failure of the programmed survey.

3.4 Sample

The final sample was a convenience sample of 115 individuals recruited via email from a database of students and former students at University Pompeu Fabra who had indicated they were willing to participate in experiments. It consisted of $n = 56$ observations in the treatment group and $n = 59$ observations in the control group. The sample showed an average age of 22 years (min = 19, max = 29) and consisted of 65% women. 69% of the sample indicated a high school degree to be their highest achieved level of education while 29% had already obtained a university degree. 84% of the sample had participated in a survey of any kind within the past 12 months and 43% had participated

in four or more surveys in this period. Significant differences between control and treatment group were found in terms of age, with respondents in the treatment group being on average 0.92 years older than those in the control group ($t = -2.41$, $p = 0.018$, Cohen's $D = 0.45$), and proportion of respondents with a university degree, with 18% more of respondents in the treatment than in the control group having a university degree already ($z = -2.03$, $p = 0.042$, Cohen's $H = 0.38$). No significant differences were found in terms of gender ($z = 0.60$, $p = 0.551$, Cohen's $H = 0.11$), in terms of whether respondents had taken any surveys in the past 12 months ($z = 0.68$, $p = 0.499$, Cohen's $H = 0.13$), or whether they had taken at least four surveys in the past 12 months ($z = -0.14$,$p = 0.892$, Cohen's $H = 0.03$). We check whether the results of our experimental analysis hold also when controlling for the two variables that differed significantly between the treatment and control group: age and proportion holding a university degree (see section 4.2). In Table 1, we summarize the characteristics of the complete sample as well as of the two groups separately.

3.5 Analyses

Observational analysis—recall of previous answer and estimation of memory effect. The observational part of the analysis contains aspects that were also investigated by Van Meurs and Saris (1990) except for section 4.1.4 (response speed and memory). Firstly, to assess to what extent people recall their survey answers, we examined the proportion of the sample which indicated to recall their previous answer and those which, in addition, managed to correctly reproduce their previous answer. Furthermore, we looked at the proportion of respondents who correctly reproduced their previous answer even though they indicated not to remember it. Following Van Meurs and Saris (1990), combining these two pieces of information allows us to estimate the proportion of respondents who give consistent answers due to memory by subtracting the proportion of consistent respondents without memory from the proportion of consistent respondents with memory. The underlying idea by Van Meurs and Saris (1990) is that when respondents manage to reproduce their answer even without remembering it, this must be due to the stability of their opinion (or other underlying concept assessed by the question) or due to chance (Saris, 2013). The same share of consistent answers is likely to be attributable to stability of the assessed concept or chance among respondents who remember their previous answer than among those who do not remember. Hence, subtracting the proportion of consistent respondents indicating that they do not remember from the proportion of consistent respondents indicating that they do remember will give us an estimate of the proportion of respondents whose consistent answers are due to memory.

Furthermore, we assessed the time passed in between repetitions, which was automatically collected by the survey software as server-side paradata[1], as well as its association with respondents being able to reproduce their previous answer correctly. We looked at the proportion of correctly reproduced answers in all decile groups of time intervals between repetitions. Moreover, to allow for a more direct comparison with Van Meurs and Saris' (1990) findings, we compared findings pertaining to the most similar time interval between repetitions covered by their study and ours.

Another point we aimed to examine was whether extreme responses increased recall capability, as concluded by Van Meurs and Saris (1990). However, as no extreme answers were provided to the test question in our study, we were unable to assess this.

Experimental analysis—can recall of previous answers and the memory effect be decreased? The experimental part is a new contribution, not considered by Van Meurs and Saris (1990). To investigate whether recall of survey answers can be decreased by our memory interference task, we compared a) the proportions of respondents who indicated to remember their previous answer and b) the proportions of respondents who correctly reproduced their previous answer between the control and the treatment group. We conducted logistic regressions in order to assess whether receiving the memory interference task affected a) respondents' likelihood of indicating to remember their previous answer and b) respondents' likelihood of correctly reproducing their previous answer. This enabled us to control for the differences in time between repetitions when looking at the effect of the memory interference task as well as for age and proportion of respondents with a university degree, two variables which differ significantly between treatment and control group (see section 3.4). In a next step, we compared the estimated memory effects between the two groups.

4 Results

4.1 Observational results—recall of previous answer and estimation of memory effect

How many respondents remember their previous answer? Firstly, to assess respondents' capability to recall their previous answer, we looked at our observational results. We then compared them with Van Meurs and Saris' (1990) results. Table 2 contains a comparison of both studies concerning self-reported memory as well as correct reproduction of previous survey answers.

We found that after an average 20-minute interval between repetitions (sd = 3.59, min = 11, max = 29), 66% of all re-

[1]It should be noted that we could not completely control for respondents visiting other websites during their completion of the web survey (see, for example, Höhne & Schlosser, 2018). However, in the controlled lab setting with the experimenter being present in the computer lab, it is unlikely that such behavior would have gone unnoticed.

Table 1
Sample characteristics

	Complete sample	Treatment group	Control group	p-value (two tailed)	Effect size[a]
Women (%)	65	63	68	0.551	0.11
Age (mean)	22	22	21	0.018	0.45
Obtained university degree (%)	29	38	20	0.042	0.38
Any survey experience in past 12 months (%)	84	82	86	0.499	0.13
Participated in at least four surveys in past 12 months (%)	43	44	42	0.892	0.03
N	115	56	59		

[a] Cohen's *D* for means, Cohen's *H* for proportions

Table 2
Comparison proportions self-reported memory and correct reproduction of previous answer(s) between current study and van Meurs and Saris (1990)

	Current study (average 20-minute interval)	Van Meurs & Saris (1990) (average nine-minute interval)
Self-reported memory (%)	66	Ranging between 71 and 85
Correct reproduction (%)	60	Ranging between 57 and 72

spondents indicated that they did remember the answer they gave previously. In comparison, Van Meurs and Saris (1990) found that after an average nine-minute interval between repetitions, among the six items they looked at, the proportion of respondents who indicated they did remember their answer ranged between 71% and 85%. Furthermore, in terms of correct reproductions of previous answers, 60% of our respondents managed to do this after an average 20-minute interval. Van Meurs and Saris (1990) found that after an average nine-minute interval between repetitions, the proportion of respondents who reproduced their previous answer correctly ranged between 57% and 72%.

Development of memory over time. We made a more direct comparison, comparing results for the most similar time interval from our study with Van Meurs and Saris (1990) study. The time interval between 12 and 20 minutes allows for a more direct comparison as it is covered by both studies.

We found that of the respondents who had this time interval in between repetitions, between 42% and 73% could correctly reproduce their answer (see Figure 2). This is in line with van Meurs and Saris' earlier findings, showing that among respondents with intervals between repetitions ranging from 12 to 20 minutes, between 41% and 79% could correctly reproduce their answers (see Figure 1).

In our study, a decrease of memory across the captured time period does not show up. For our test question, an evaluation of how easy or difficult respondents find it to deal with problems in their life (11-point scale), it seems to make no difference whether 11 or 30 minutes passed between repetitions. Indeed, the correlation between the time passed between repetitions and whether respondents gave consistent answers is not significant ($r = 0.07$, $p = 0.436$). Instead, the highest rate of correct answers is observed among those respondents who were neither particularly fast nor particularly slow, i.e., those taking between 21 and 23 minutes (see Figure 2). For comparison, Van Meurs and Saris (1990) observed an average decrease of correct reproduction of previous answers across time for four of the six items they studied.

Estimating the size of the memory effect. Next, we compare the estimated memory effect in our study with that estimated by Van Meurs and Saris (1990). Table 3 shows the correct reproduction of answers among respondents who indicated they remembered and among those who indicated they did not remember as well as the resulting estimation of the size of the memory effect for both studies.

The proportion of correct responses among those who indicated to remember their previous answer is similar in the two studies: 66% as compared to an average of 70% with min = 66% and max = 79%. However, the proportion of those who give correct responses among the ones who indicate not to remember their previous answer is higher in the current study: 49% as compared to their average of 36% with

Table 3
Comparison proportions correct reproduction of previous answer(s) among respondents who remember and who do not remember between current study and van Meurs and Saris (1990)

	Current study (average 20-minute interval)	van Meurs and Saris (1990) (average nine-minute interval)
Correct reproduction among self-reported memory (%)	66	Average 70 (range 66 to 79)
Correct reproduction among self-reported no memory (%)	49	Average 36 (range 34 to 47)
Estimated memory effect (%)	17	34

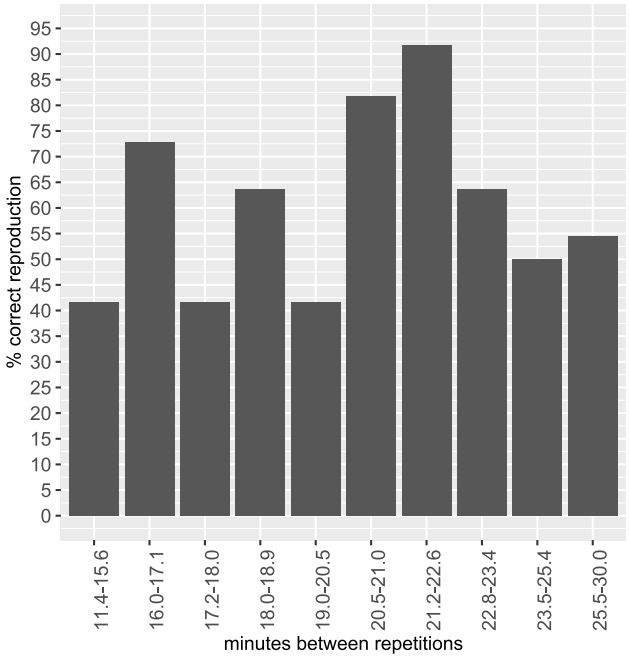

Figure 2. Proportion providing correct reproduction of previous answers in present study across time intervals (based on deciles with $n = 11$ or $n = 12$) (Note: We display the proportions within decile groups based on the minutes between repetitions variable. One decile group might have as maximum value 18.0 minutes while the next decile might start with that value as minimum value. In order to display no overlap, we would have to display more decimal digits here. The fact that some values do not appear is due to the fact that not all possible values of "minutes between repetitions" did occur among respondents.)

min = 34% and max = 47%. Therefore, we estimate the resulting memory effect to be lower (66% − 49% = 17%) than the original study (70% − 36% = 34%).

Response speed and memory. In the following, no comparison will be made with the results of Van Meurs and Saris (1990), as the discussed aspects were not covered by their study. We found that the response time for the test question when first presented at the beginning of the survey is not significantly correlated with correctly reproducing this answer later. However, we found a significant negative correlation between response time for the question asking to reproduce the previous answer and whether respondents reproduced it correctly ($r = -0.28$, $p = 0.002$). As respondents received slightly different requests asking them to restate or approximate their previous answer, depending on whether they had reported to remember it or not, we checked whether the correlation was present in each of these two groups. We found that it only holds for the group of respondents who indicated they did remember their previous answer. Hence, among respondents reporting to remember their previous answer, the less time they spent on reproducing their previous answer, the more likely they were to reproduce it correctly ($r = -0.40$, $p = 0.000$). This suggests that it required those who remembered correctly less cognitive effort to reproduce their previous answer than those who did not remember their exact previous answer.

4.2 Experimental results—can recall of previous answers and the memory effect be reduced?

Can recall of previous answers be reduced? Our second aim was to investigate whether a memory interference task could decrease recall of previously given answers and the estimated memory effect. Table 4 presents the differences in proportions between memory interference group and control group on self-reported memory as well as correct reproduction of the previous answer.

Contrary to our expectations formulated in hypothesis 1, the proportion of respondents who reported they remembered their previous answer was higher in the treatment than in the control group (75% versus 58%). This difference is statistically significant ($z = -1.97$, $p = 0.049$, Cohen's $H = 0.37$). Furthermore, hypothesis 2 receives no support as the proportion of respondents who correctly reproduced their answer was slightly higher in the treatment group (63% versus 58%) but this difference is not significant ($z = -0.53$, $p = 0.594$, Cohen's $H = 0.10$).

Moreover, we looked at the difference in time interval between repetitions across groups. This was on average 17 seconds and thus less than expected as the fixed time respondents had to spend on the interference task was one minute. This indicates that respondents in the treatment group filled

Table 4
Comparison proportions self-reported memory and correct reproduction of previous answer between treatment and control group

	Control group %	Treatment group %	p-value (two tailed)	Effect size (Cohen's H)
Self-reported memory	58	75	0.049	0.37
Correct reproduction	58	63	0.594	0.10

in the rest of the survey somewhat faster than those in the control group. However, the difference is not significant ($t = -0.43$, $p = 0.669$, Cohen's D = 0.08).

Table 5 shows the results of the logistic regression models employed to determine whether the memory interference task had an effect on self-reported memory as well as on correct reproduction of the previous answer.

The results confirm those of the bivariate analyses: There is no support for hypothesis 1 as respondents in the treatment group were significantly more likely to indicate that they did remember their previous answer than those in the control group (OR = 2.84, $p = 0.017$). More specifically, for a respondent who received the memory interference task, the odds of stating to remember their previous answer were almost three times as large as for respondents in the control group. However, the treatment did not have a significant effect on whether or not respondents correctly reproduced their answer (OR = 1.28, $p = 0.527$), thus there is also no support for hypothesis 2.

Can the memory effect be reduced? Table 6 contains the differences in proportions of correct reproduction of the previous answer among those who stated they remembered it and among those who stated they did not remember it.

The proportion of respondents who correctly reproduced their answer among those who stated to remember it is higher in the treatment group, while the proportion of respondents who correctly reproduced their answer among those who stated not to remember it is lower in the treatment group. However, neither difference is statistically significant ($z = -0.64$, $p = 0.522$, Cohen's $H = 0.15$ and $z = 0.54$, $p = 0.590$, Cohen's $H = 0.18$, respectively). As a consequence, the memory effect is estimated to be higher in the treatment than in the control group but this difference is also not statistically significant ($z = -1.77$, $p = 0.076$, Cohen's $H = 0.43$). Hypothesis 3 thus does not receive support.

5 Conclusion and discussion

Our goals in this article were 1) to investigate to what extent respondents recall previous answers within the same survey as well as to estimate the memory effect, both inspired by the study of Van Meurs and Saris (1990) and 2) to investigate if the memory effect could be reduced by implementing an interference task, going beyond the study of Van Meurs and Saris (1990). We asked a student sample to take a web survey with a vast majority of questions being adopted from the ESS. The third question of the survey (asking how difficult or easy respondents find it to deal with problems in their daily life on an 11-point scale) was displayed again at the very end of the questionnaire asking respondents whether they remembered their previous answer to it. Depending on their answer (yes or no), the next question asked them to reproduce their previous answer or approximate it as best as they could, respectively. Furthermore, half of the respondents were randomly assigned to completing a memory interference task after answering the third question of the survey.

Regarding our first goal, we found that, after an average 20-minute time interval (after answering on average 127 questions), 66% of the respondents in our sample stated to remember their previous response. Furthermore, overall, 60% of the respondents did indeed reproduce their previous answer correctly and no decreasing pattern could be observed for respondents with longer time intervals between repetitions. Thus, respondents with longer time intervals between repetitions did not do worse in our study. Among the respondents who stated they could remember their previous answer, 66% indeed correctly reproduced their answer, versus 49% of the respondents who stated they would not remember their previous answer. Following Van Meurs and Saris (1990), we therefore assume that 49% correctly reproduce their answer due to consistency or chance and thus that only 66% – 49% = 17% correctly reproduce their answer due to memory. This estimate is lower than what was found in the original study by Van Meurs and Saris (1990).

Furthermore, we could not find any clear pattern in the development of respondents' memory in the assessed time range between 11 minutes and 30 minutes in between repetitions. We speculate that this could be due to two distinct processes going on simultaneously which cancel each other out. While memory has clearly been found to decay over time (Baddeley, 2013), filling in a survey more slowly is likely associated with more cognitive effort spent on the task (Revilla & Ochoa, 2015) which, in turn, can be expected to improve recall (Baddeley, 2013). It should be noted that the lack of a decreasing pattern is not directly contradicting Van Meurs and Saris (1990) findings, as they similarly found no decreas-

Table 5
Logistic regression models assessing effect of treatment on self-reported memory and correct reproduction of previous answer

	Self-reported memory		Correct reproduction	
	OR	SE	OR	SE
Treatment	2.84*	1.24	1.28	0.51
Minutes between repetitions	0.96	0.06	1.04	0.06
Age	0.85	0.11	0.90	0.11
University	0.76	0.43	1.27	0.70
Intercept	94.25	274.27	5.35	14.64
Pseudo R^2	0.06		0.01	

* $p \leq 0.05$

Table 6
Comparison proportions correct reproduction of previous answer among respondents who remember and who do not remember and estimated memory effect between treatment and control group

	Control group %	Treatment group %	p-value (two tailed)	Effect size (Cohen's H)
Correct reproduction among self-reported memory (%)	62	69	0.522	0.15
Correct reproduction among self-reported no memory (%)	52	43	0.590	0.18
Estimated memory effect (%)	10	26	0.076	0.43

Number of observations self-reported memory: control group $n = 34$, treatment group $n = 42$; Number of observations self-reported no memory: control group $n = 25$, treatment group $n = 14$

ing pattern of memory over time in two out of the six items on different topics they assessed. Hence, the question topic might play a role here.

The most important implication of our findings is that even though the estimated memory effect of 17% is much lower than that found by Van Meurs and Saris (1990), it can still be expected that this can create bias in estimates derived from research based on repeated measures within one survey interview, like the estimation of measurement error or pretest-posttest experimental designs.

Our first goal was to determine the extent to which respondents could recall previous answers within the same survey and to estimate the memory effect, following the Van Meurs and Saris' (1990) study and we thus focused on comparing our results to theirs. However, there are many differences between the two studies which might explain the different findings, such as: 1) the higher average time in between repetitions in our study and the fact that our variance in time intervals was smaller than theirs, 2) the different samples used, 3) the lab versus the field setting, and 4) the fact that we only looked at one test question while van Meurs and Saris looked at six different items. These differences imply limitations to the comparisons we can draw. However, they allow insights about how robust van Meurs and Saris' findings are to different contexts.

Another advantage of our study is that differences in time passed between repetitions are mostly due to respondent's response speed while in Van Meurs and Saris (1990) study, this is to a larger degree confounded with the amount of questions presented in between repetitions. Separating the effects of these two variables should be an aim of future studies. This could be done by fixing the time respondents have available for each question. Web questionnaires including the feature of fixed response times have been tested and could be used in future surveys (Revilla, Ochoa, & Turbina, 2017).

The second goal of this study was to investigate the possibility of reducing respondents' memory of their previous survey answer by asking them to complete a memory interference task. Using a survey task aimed at interfering with memory of the previously given answer by increasing the

amount of similar content presented between repetitions for half of the respondents, we did not find the expected effects. The memory interference task did not decrease respondents self-reported memory (hypothesis 1), nor their correct reproduction of respondents' previous answers (hypothesis 2) nor the memory effect (hypothesis 3). This might be because presenting respondents with additional similar content could have also caused additional retrieval which, in turn, could have increased respondents' ability to remember (Baddeley, Eysenck, & Anderson, 2009). The finding that respondents in the treatment group were significantly more confident to remember their answers is in line with this idea. This finding seems to clash with the finding of Van Meurs and Saris (1990) that similar content between repetitions decreases memory. In our study, additional similar content might have rather caused additional retrieval of the initial question and answer. Further research is needed to investigate whether slightly different tasks of a similarly compact format as the interference task used in the current study could be used to the desired end of decreasing respondents' memory of their previous survey answers.

As in most studies, the design of the current study is subject to limitations. Firstly, we used a small convenience sample (n = 115) of students between 19 and 29 years, of which 65% were women, which limits the generalizability of our results. Especially the young age of our respondents can be expected to make a difference in terms of memory performance, given that memory has commonly been found to decrease with age in the cognitive aging literature (Luo & Craik, 2008). At the same time, respondents in this age range are especially interesting to investigate because they can be expected to be particularly familiar with completing tasks in online environments (Hartman & McCambridge, 2011). Differences in age and familiarity with online tasks might be part of the explanation why we could not observe a decay of memory in our sample, contrary to the findings of Van Meurs and Saris (1990). Yet, this would need to be further tested with different, preferably larger and more representative samples. A critical view at Van Meurs and Saris' (1990) way of estimating the memory effect is, furthermore, in order. Particularly, the question arises whether some respondents indicating that they do not remember their previous answer were not simply trying to avoid the burden of answering a follow-up question. This would result in an underestimation of the memory effect. Future research should address such possibilities and work on developing alternative ways of estimating the memory effect.

In conclusion, our findings are relevant for different kinds of research employing repeated measurement of survey items, such as for the estimation of test-retest reliability or for experimental designs based on collecting the same measure pre- and post-treatment. Such research needs to find ways to account for the fact that a substantive part of consistency in answers can be attributed to respondents' memory. Until now, the two main approaches to dealing with the problems of memory effects have been correcting for them statistically (Laenen, Alonso, Molenberghs, & Vangeneugden, 2006; Scherpenzeel & Saris, 1997) or letting adequate time periods pass in between repetitions. However, very little is known yet about the size of memory effects within a single survey, depending on different aspects (number of questions, topics, scale formats, etc.). Our study provides some results to start filling this gap. Moreover, it allows us to identify a research agenda for future research on the aspects that could influence the size of memory effect. These aspects, which can be studied separately or combined, are the following: (1) the number of items between repetitions, to see how memory develops when these are increased or decreased (2) the time interval between repetitions, for the same reason, (3) the type of items (different topics, likely to either cause extreme answers or not, different scales), (4) response speed, as discussed above (5) different samples, preferably large and representative, that allow for inferences to populations, (7) conducting a field study as opposed to a lab study to see if results change when memory effects are investigated outside of the artificial lab environment, in the field setting where surveys normally take place, and (6) investigating interference tasks. This last point is a suggestion of a novel way of dealing with the problem, namely to purposefully interfere with respondents memory during the course of the survey.

Acknowledgement

We thank Mark Asensio and Oriol Bosch for their support during the survey development as well as Willem Saris, Daniel Oberski, Jan Karem Höhne and the anonymous reviewers for their helpful comments. We also thank the Faculty of Political and Social Sciences at Universitat Pompeu Fabra for supporting this research with a grant.

References

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. Hoboken: John Wiley & Sons.

Alwin, D. F. (2011). Evaluating the reliability and validity of survey interview data using the MTMM approach. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), *Question evaluation methods. Contributing to the science of data quality* (pp. 263–293). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, *48*(2), 409–442. doi:10.1086/268840

Bach, R. L., & Eckman, S. (2018). Motivated misreporting in web panels. *Journal of Survey Statistics and Methodology*, *6*(3), 418–430. doi:10.1093/jssam/smx030

Baddeley, A. (2013). *Essentials of human memory (classic edition)*. Hove: Psychology Press.

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory*. Hove: Psychology Press.

Bekker, G. A., Fischer, A. R., Tobi, H., & van Trijp, H. C. (2017). Explicit and implicit attitude toward an emerging food technology: The case of cultured meat. *Appetite*, *108*, 245–254. doi:10.1016/j.appet.2016.10.002

Brown, S. L., Asher, T., & Cialdini, R. B. (2005). Evidence of a positive relationship between age and preference for consistency. *Journal of Research in Personality*, *39*(5), 517–533. doi:10.1016/j.jrp.2004.07.001

Callegaro, M., Manfreda, K. L., & Vehovar, V. (2015). *Web survey methodology*. London: Sage.

Cialdini, R. B., Trost, M. R., & Newsom, J. T. (1995). Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of Personality and Social Psychology*, *69*(2), 318. doi:10.1037/0022-3514.69.2.318

Crouse, J. H. (1971). Retroactive interference in reading prose materials. *Journal of Educational Psychology*, *62*(1), 39–44. doi:10.1037/h0030652

Dimitrov, D. M., & Rumrill, P. D. (2003). Pretest-posttest designs and measurement of change. *Work*, *20*(2), 159–165.

Gawronski, B. (2012). Back to the future of dissonance theory: Cognitive consistency as a core motive. *Social Cognition*, *30*(6), 652–668. doi:10.1521/soco.2012.30.6.652

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1411. doi:10.1037/0278-7393.27.6.1411

Guadagno, R. E., & Cialdini, R. B. (2010). Preference for consistency and social influence: A review of current research findings. *Social Influence*, *5*(3), 152–163. doi:10.1080/15534510903332378

Hartman, J. L., & McCambridge, J. (2011). Optimizing millennials' communication styles. *Business Communication Quarterly*, *74*(1), 22–44. doi:10.1177/1080569910395564

Heise, D. R. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review*, *34*(1), 93–101. doi:10.2307/2092790

Henderson, J. (2005). *Memory and forgetting*. London: Routledge.

Hindo, C. S., & González-Prendes, A. A. (2011). One-session exposure treatment for social anxiety with specific fear of public speaking. *Research on Social Work Practice*, *21*(5), 528–538.

Hogg, M. A., & Grieve, P. (1999). Social identity theory and the crisis of confidence in social psychology: A commentary, and some research on uncertainty reduction. *Asian Journal of Social Psychology*, *2*(1), 79–93. doi:10.1111/1467-839X.00027

Höhne, J. K., & Schlosser, S. (2018). Investigating the adequacy of response time outlier definitions in computer-based web surveys using paradata surveyfocus. *Social Science Computer Review*, *36*(3), 369–378. doi:10.1177/0894439317710450

Kopelman, M. D., Bright, P., Buckman, J., Fradera, A., Yoshimasu, H., Jacobson, C., & Colchester, A. C. (2007). Recall and recognition memory in amnesia: Patients with hippocampal, medial temporal, temporal lobe or frontal pathology. *Neuropsychologia*, *45*(6), 1232–1246. doi:10.1016/j.neuropsychologia.2006.10.005

Krosnick, J. A. (2011). Experiments for evaluating survey questions. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), *Question evaluation methods. contributing to the science of data quality* (pp. 215–238). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Kumar, A. (2000). Interference effects of contextual cues in advertisements on memory for ad content. *Journal of Consumer Psychology*, *9*(3), 155–166. doi:10.1207/S15327663JCP0903_3

Laenen, A., Alonso, A., Molenberghs, G., & Vangeneugden, T. (2006). Accounting for the rater's memory effect in reliability estimation. Technical Report 0656 IAP Statistics Network Interuniversity Attraction Pole. Retrieved from https://sites.uclouvain.be/IAP-Stat-Phase-V-VI/PhaseV/publications_2006/TR/TR0656.pdf

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. Information Age Publishing Inc.

Luo, L., & Craik, F. I. (2008). Aging and memory: A cognitive approach. *The Canadian Journal of Psychiatry*, *53*(6), 346–353. doi:10.1177/070674370805300603

Manns, J. R., Hopkins, R. O., Reed, J. M., Kitchener, E. G., & Squire, L. R. (2003). Recognition memory and the human hippocampus. *Neuron*, *37*(1), 171–180. doi:10.1016/S0896-6273(02)01147-9

Marx, R. G., Menezes, A., Horovitz, L., Jones, E. C., & Warren, R. F. (2003). A comparison of two time intervals for test-retest reliability of health status instruments. *Journal of clinical epidemiology*, *56*(8), 730–735. doi:10.1016/S0895-4356(03)00084-2

Mavletova, A., & Couper, M. P. (2013). Sensitive topics in PC web and mobile web surveys: is there a difference? *Survey Research Methods*, *7*(3), 191–205. doi:10.18148/srm/2013.v7i3.5458

McConnell, K. E., Strand, I. E., & Valdés, S. (1998). Testing temporal reliability and carry-over effect: The role of correlated responses in test-retest reliability studies.

Environmental and Resource Economics, *12*(3), 357–374. doi:10.1023/A:1008264922331

McGeoch, J. A., & McDonald, W. T. (1931). Meaningful relation and retroactive inhibition. *The American Journal of Psychology*, *43*(4), 579–588. doi:10.2307/1415159

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

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. doi:10.1037/h0031564

Momm, T., Blickle, G., & Liu, Y. (2013). Political skill and emotional cue learning via voices: A training study. *Journal of Applied Social Psychology*, *43*(11), 2307–2317.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Qualtrics. (2020). Provo, Utah, USA: Qualtrics. Retrieved from https://www.qualtrics%20.com

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*(1), 159–182. doi:10.1037/0033-295X.111.1.159

Revilla, M., & Ochoa, C. (2015). What are the links in a Web survey among response time, quality, and auto-evaluation of the efforts done? *Social Science Computer Review*, *33*(1), 97–114. doi:10.1177/0894439314531214

Revilla, M., Ochoa, C., & Toninelli, D. (2016). PCs versus Smartphones in answering web surveys: does the device make a difference? *Survey Practice*, *9*(4), 2804. doi:10.29115/SP-2016-0021

Revilla, M., Ochoa, C., & Turbina, A. (2017). Making use of internet interactivity to propose a dynamic presentation of web questionnaires. *Quality & Quantity*, *51*(3), 1321–1336. doi:10.1007/s11135-016-0333-2

Revilla, M., Toninelli, D., Ochoa, C., & Loewe, G. (2016). Do online access panels need to adapt surveys for mobile devices? *Internet Research*, *26*(5), 1209–1227. doi:10.1108/IntR-02-2015-0032

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

Saris, W. E. (1990). Models for evaluation of measurement instruments. In A. Van Meurs & W. E. Saris (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait multimethod studies* (pp. 52–80). Amsterdam: North Holland.

Saris, W. E. (1998). Ten years of interviewing without interviewers: The telepanel. In M. P. Couper, R. P. Baker, J. Bethlehem, C. Clark, J. Martin, W. L. Nicholls, & J. M. O'Reilly (Eds.), *Computer assisted survey information collection* (pp. 409–429). New York: John Wiley & Sons.

Saris, W. E. (2013). Is there anything wrong with the MTMM approach to question evaluation? *Pan-Pacific Management Review*, *16*(1), 49–77.

Saris, W. E., & Revilla, M. (2016). Correction for measurement errors in survey research: Necessary and possible. *Social Indicators Research*, *127*(3), 1005–1020. doi:10.1007/s11205-015-1002-x

Saris, W. E., Satorra, A., & Coenders, G. (2004). A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design. *Sociological Methodology*, *34*(1), 311–347.

Scherpenzeel, A. C., & Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods & Research*, *25*(3), 341–383.

Toninelli, D., & Revilla, M. (2016a). Is the smartphone participation affecting the web survey experience? 48th Scientific Meeting of the Italian Statistical Society. IT.

Toninelli, D., & Revilla, M. (2016b). Smartphones vs PCs: Does the device affect the web survey experience and the measurement error for sensitive topics?-A replication of the Mavletova & Couper's 2013 experiment. *Survey Research Methods*, *10*(2), 153–169. doi:10.18148/srm/2016.v10i2.6274

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*(3), 299. doi:10.1037/0033-2909.103.3.299

United States Federal Committee on Statistical Methodology. (1987). Statistical Policy Working Paper 4 - Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics. Retrieved from https://stats.oecd.org/glossary/detail.asp?ID=4534

Van Dyke, J. A., & Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Language and Linguistics Compass*, *6*(4), 193–211. doi:10.1002/lnc3.330

Van Meurs, A., & Saris, W. E. (1990). Memory effects in MTMM studies. In A. Van Meurs & W. E. Saris (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait multimethod studies* (pp. 134–146). Amsterdam: North Holland.

Vitale, J., Kosson, D. S., Resch, Z., & Newman, J. P. (2018). Speed-accuracy tradeoffs on an affective lexical decision task: Implications for the affect regulation theory of psychopathy. *Journal of Psychopathology and Be-*

havioral Assessment, *40*(3), 412–418. doi:10 . 1007 / s10862-018-9652-z

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140–159. doi:10.1016/j.jml.2007.04.006

Wiley, D. E., & Wiley, J. A. (1970). The estimation of measurement error in panel data. *American Sociological Review*, *35*(1), 112–117. doi:10.2307/2093858

Appendix A
Tables

Table A1
Overview of questionnaire for control and treatment group

Questions Control Group	Questions Treatment Group
ESS R6 D28 – D30 (D30= Repeated question) – 3 question	**ESS R6** D28 – D30 (D30= Repeated question) – 3 questions
–	**One-minute Memory Interference task**
ESS R6 34 questions from D Block (Personal and social wellbeing, helping others, feelings in the last week, life satisfaction, physical activity)	**ESS R6** 34 questions from D Block (Personal and social wellbeing, helping others, feelings in the last week, life satisfaction, physical activity)
ESS R8 A1 – A6 (Media use; Internet use; social trust) – 6 questions	**ESS R8** A1 – A6 (Media use; Internet use; social trust) – 6 questions
ESS R8 31 questions from D Block (climate change)	**ESS R8** 31 questions from D Block (climate change)
ESS R7 19 questions from E Block (health)	**ESS R7** 19 questions from E Block (health)
10 Demographics questions **2 Survey experience questions**	**10 Demographics questions** **2 Survey experience questions**
ESS R6/8 HF1/HF2 (A-U) (Human values scale) – 21 questions	**ESS R6/8** HF1/HF2 (A-U) (Human values scale) – 21 questions
Two questions to measure recall (see Appendix B)	**Two questions to measure recall** (see Appendix B)

Table A2
Examples of sentences used in memory interference task

Incorrect	Considero ocuparme imposible de los problemas importantes de mi vida. ¿Hasta qué punto crees que es difícil o fácil tratar con vecinos ruidosas? Los problemas de la vida me dificulta el día a día. Considero más bien fácil lidiar con los problemas que me da la vida de cuanto en cuanto.
Correct	¿Cuán difícil o fácil piensa que es ocuparse de los problemas importantes de la vida? ¿Hasta qué punto hallas fácil tratar con los problemas importantes con los que te encuentras en la vida? Tratar con una pareja celosa es un serio problema. Es fácil adaptarse a una situación adversa.

Table A3
Translation of examples of sentences used in memory interference task

Incorrect	I consider it impossible to take care of the important problems of my life. [grammatical mistake in word order] To what extent do you think it is difficult or easy to deal with noisy neighbours? [grammatical mistake in adjective] Life's problems make daily life difficult for me. [grammatical mistake in verb] I consider it rather easy to deal with the problems that life gives me occasionally. [grammatical mistake in adverb at the end of sentence]
Correct	How difficult or easy do you think it is to take care of the important problems of life? To what extent do you find it easy to deal with the important problems that you encounter in life? Dealing with a jealous partner is a serious problem. It is easy to adapt to an adverse situation.

Appendix B
Screenshots and translations of questions to measure recall

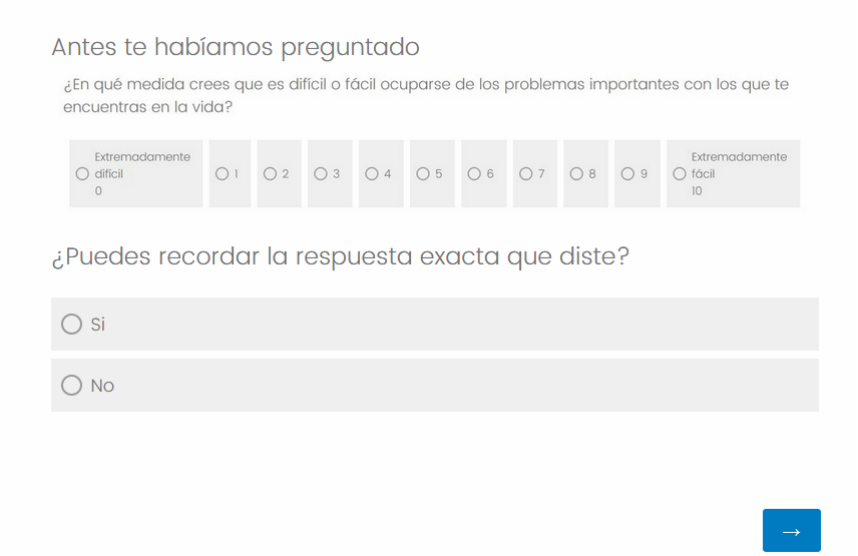

Figure B1. Second presentation of test question
Translation: "We asked you the following question previously: How difficult or easy do you find it to deal with important problems that come up in your life?" Label scale point 0: "Extremely difficult", Label scale point 10: "Extremely easy". "Can you remember your exact answer to this question?" "Yes", "No"

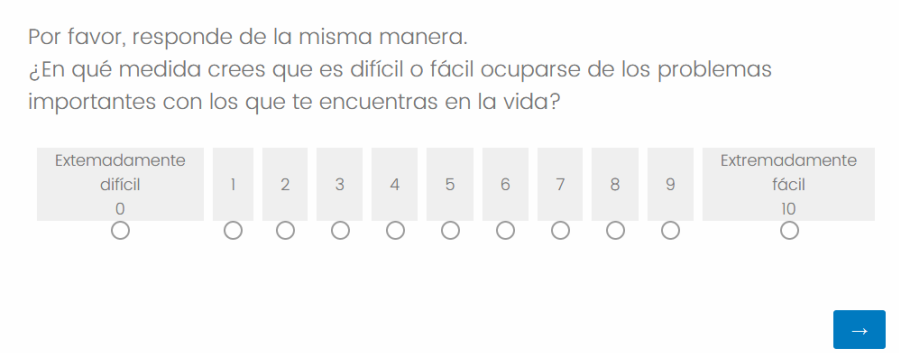

Figure B2. Follow-up question asking to restate previous answer
Translation IF YES: "Please provide the answer you gave before. How difficult or easy do you find it to deal with important problems that come up in your life?" Label scale point 0: "Extremely difficult" Label scale point 10: "Extremely easy".

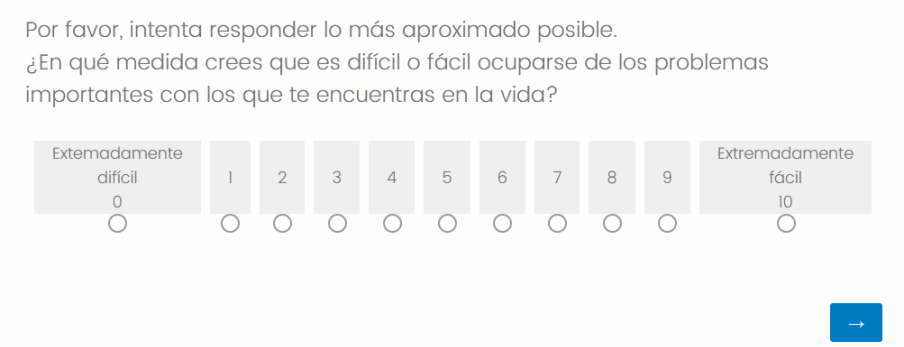

Figure B3. Follow-up question asking to approximate previous answer
Translation IF NO: "Please provide your best guess of the answer you gave before. How difficult or easy do you find it to deal with important problems that come up in your life?" Label scale point 0: "Extremely difficult" Label scale point 10: "Extremely easy".

Appendix C
Exemplary screenshots of treatment questionnaire with translations

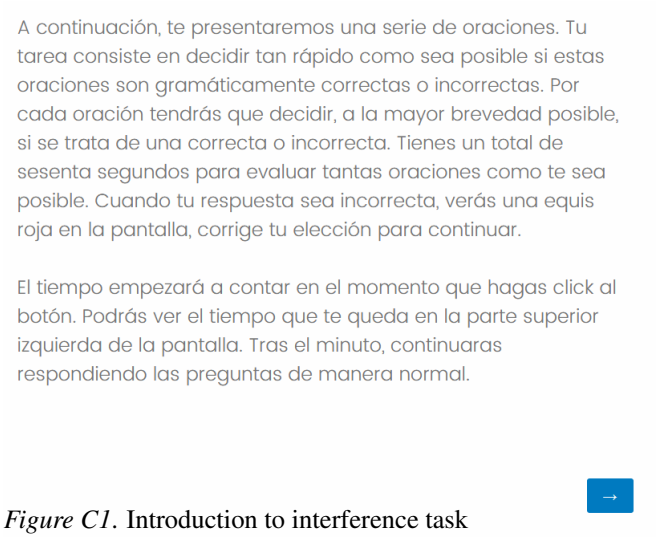

Figure C1. Introduction to interference task
Translation "In the following, we present you a task in order to find out how fast you can judge whether a sentence is non-sensical, i.e. grammatically incorrect or whether it is a real, correct sentence. Therefore, we will show you a series of sentences, some correct and others incorrect. We will ask you for each to decide as quickly as possible whether or not the sentence is correct. You have one minute time and your task is to evaluate as many sentences as possible during this time. The timer starts when you hit the next button. Your remaining time will be displayed in the upper left hand corner. After 1 minute, you will continue with answering normal survey questions."

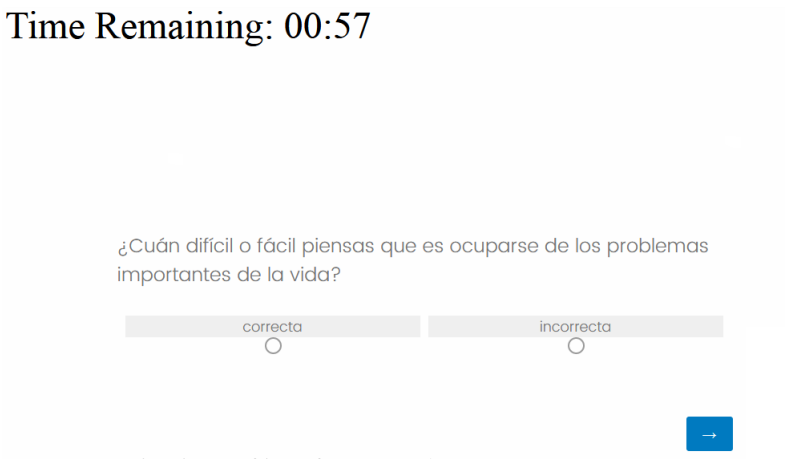

Figure C2. First item of interference task
Translation "How difficult or easy do you think it is to deal with important problems in life?" "Correct", "incorrect".